Printed: ISBN: 978-0-9564944-0-5

European Council for Modelling and Simulation

CD: ISBN: 978-0-9564944-1-2

Logo Design "Simulation meets global challenges" by www.thisisoctarine.com
Copyright 2010

final layout by **Digitaldruck Pirrot GmbH**
66125 Sbr.-Dudweiler, Germany

printed by **Maha Cetak Trading SDN BHD**
43300 Seri Kembangan, Malaysia

PROCEEDINGS

24th European Conference on Modelling and Simulation ECMS 2010

June 1st – 4th, 2010
Kuala Lumpur, Malaysia

Edited by:

Andrzej Bargiela
Sayed Azam Ali
David Crowley
Eugène J.H. Kerckhoffs

Organized by:

ECMS - European Council for Modelling and Simulation

Hosted by:

The University of Nottingham Malaysia Campus

Sponsored by:

The University of Nottingham Malaysia Campus

InfoValley Bhd

International Co-Sponsors:

IEEE - Institute of Electrical and Electronics Engineers
ASIM - German Speaking Simulation Society
EUROSIM - Federation of European Simulation Societies
PTSK - Polish Society of Computer Simulation
LSS - Latvian Simulation Society

ECMS 2010 ORGANIZATION

General Conference and Programme Chair

Andrzej Bargiela

The University of Nottingham Malaysia Campus
Malaysia

Conference Co-Chair

Sayed Azam Ali

The University of Nottingham Malaysia Campus
Malaysia

Programme Co-Chair

David Crowley

The University of Nottingham Malaysia Campus
Malaysia

Honorary Conference Chair

Eugène J.H. Kerckhoffs

TU Delft
The Netherlands

President of European Council for Modelling and Simulation

Khalid Al-Begain

University of Glamorgan
United Kingdom

Managing Editor

Martina-Maria Seidel

St. Ingbert
Germany

INTERNATIONAL PROGRAMME COMMITTEE

Agent-Based Simulation

Track Chair: **Eugène J.H. Kerckhoffs**
TU Delft, The Netherlands

Track Co-Chair: **Pavel Nahodil**
Czech Technical University of Prague, Czech Republic

Simulation of Complex Systems & Methodologies

Track Chair: **Krzysztof Amborski**
Warsaw University of Technology, Poland

Programme Chair: **Jaroslav Sklenar**
University of Malta, Malta

Simulation in Industry, Business and Services

Track Chair: **Alessandra Orsoni**
University of Kingston, United Kingdom

Programme Chair: **Serhiy Kovela**
University of Kingston, United Kingdom

Simulation Applications in Industry

Track Chair: **Agostino Bruzzone**
MISS DIP University of Genoa, Italy

Track Co- Chair: **Francesco Longo**
University of Calabria, Italy

Programme Chair: **Alberto Tremori**
Simulation Team, Italy

Finance and Economics

Track Chair: **Javier Otamendi**
University of Rey Juan Carlos Madrid, Spain

Track Co-Chair: **José Luis Montes**
University of Rey Juan Carlos Madrid, Spain

Simulation of Intelligent Systems

Track Chair: **Lars Nolle**
Nottingham Trent University, United Kingdom

Track Co-Chair: **Ivan Zelinka**
Tomas Bata University of Zlín, Czech Republic

Programme Chair: **Zuzana Oplatková**
Tomas Bata University of Zlín, Czech Republic

Programme Co-Chair: **Shane Lee**
University of Wales, United Kingdom

Simulation, Experimental Science and Engineering

Track Chair: **Jan Amborski**
Institute of Aviation, Poland

Track Co-Chair: **Rafal Kajka**
Institute of Aviation, Poland

Electrical and Electromechanical Engineering

Track Chair: **Sergiu Ivanov**
University of Craiova, Romania

Track Co-Chair: **Francis Labrique**
Catholic University of Louvain, Belgium

Programme Chair: **Maria José Resende**
Technical University of Lisbon, Portugal

Modelling, Simulation and Control of Technological Processes

Track Chair: **Jiři Vojtěšek**
Tomas Bata University in Zlín, Czech Republic

Track Co-Chair: **Petr Dostál**
Tomas Bata University in Zlín, Czech Republic

Programme Chair: **František Gazdoš**
Tomas Bata University in Zlín, Czech Republic

Multi-Resolution and Granular Modelling

Track Chair: **Siang Yew Chong**
University of Nottingham Malaysia Campus, Malaysia

IPC Members in Alphabetical Order

Tony Allen, Nottingham Trent University, United Kingdom
Monika Bakošová, Slovak University of Technology in Bratislava, Slovakia
John Bland, Nottingham Trent University, United Kingdom
Juan C. Burguillo-Rial, University of Vigo, Spain
Ester Camiña Centeno, Complutense University Madrid, Spain
Petr Chalupa, Tomas Bata University in Zlín, Czech Republic
Huanhuan Chen, University of Birmingham, United Kingdom
Antonio Cimino, MSC-LES, Italy
Duilio Curcio, University of Calabria, Italy
Donald Davendra, Tomas Bata University in Zlín, Czech Republic
Bruno Dehez, Catholic University of Louvain, Belgium
Andrea Del Pizzo, University of Naples Federico II, Italy
Ángel Diaz Chao, Rey Juan Carlos University, Spain
Luis Miguel Doncel Pedrera, University Rey Juan Carlos Madrid, Spain
František Dušek, University of Pardubice, Czech Republic
Andrzej Dzielinski, Warsaw University of Technology, Poland
Miroslav Fikar, Slovak University of Technology in Bratislava, Slovakia
Charlotte Gerritsen, VU University of Amsterdam, The Netherlands
Maria Grazia Gnoni, University of Salento, Italy
Chi Keong Goh, Rolls Royce, Singapore
Pilar Grau, University Rey Juan Carlos Madrid, Spain
Jana Hájková, University of West Bohemia, Czech Republic
Mark Hoogendorn, VU University of Amsterdam, The Netherlands
Daniel Honc, University of Pardubice, Czech Republic
Martin Ihrig, University of Pennsylvania, USA
Teruaki Ito, University of Tokushima, Japan
Patrick Kirchhof, University of Osnabrück, Germany
Joanna Kolodziej, University of Bielsko-Biala, Poland
Petia Koprinkova-Hristova, Bulgarian Academy of Sciences, Bulgaria
Mladen Kos, University of Zagreb, Croatia
Igor Kotenko, St. Petersburg Institute for Informatics, Russia
Marek Kubalcik, Tomas Bata University in Zlín, Czech Republic

Bin Li, University of Science and Technology, China
Ahmad Lotfi, Nottingham Trent University, United Kingdom
Ulf Lotzmann, University of Koblenz, Germany
Dorin Lucache, Technical University of Lasi, Romania
Susan Lysecky, University of Arizona, USA
Radek Matoušek, Brno University of Technology, Czech Republic
Radek Matušů, Tomas Bata University in Zlín, Czech Republic
Tomás Maul, Nottingham University Malaysia Campus, Malaysia
Nicolas Meseth, University of Osnabrück, Germany
Dan Mihai, University of Craiova, Romania
Marek Miller, Institute of Aviation, Poland
Michael Möhring, University of Koblenz, Germany
Tomoharu Nakashima, Osaka Prefecture University, Japan
Libero Nigro, University of Calabria, Italy
Jakub Novák, Tomas Bata University in Zlín, Czech Republic
Ilie Nuca, Technical University of Moldova, Moldova
Yew Soon Ong, Nanyang Technical University, Singapore
Teodor Pana, Technical University Cluj-Napoca, Romania
Evtim Peytchev, University of Nottingham, United Kingdom
Ioan Popa, University of Craiova, Romania
Matthijs Pontier, VU University of Amsterdam, The Netherlands
Napoleon H. Reyes, Massey University, New Zealand
Boris Rohal-Ilkiv, Slovak University of Technology in Bratislava, Slovakia
Sabrina Rondinelli, University of Calabria, Italy
Leon Rothkrantz, Delft University of Technology, The Netherlands
Jorge Sáinz, University Rey Juan Carlos, Spain
Sancho Salcedo-Sanz, University of Alcala, Spain
Ismael Sanz Labrador, University Rey Juan Carlos, Spain
Yvonne Seow, University of Nottingham Malaysia Campus, Malaysia
Rober Signorile, Boston College, USA
Andrzej Sluzek, Nanyang Technological University, Singapore
Mojca Stemberger, University of Ljubliana, Slovenia

Loránd Szabó, Technical University Cluj-Napoca, Romania
Roger Tait, Nottingham Trent University, United Kingdom
Kay Chen Tan, National University of Singapore, Singapore
Peter Tiňo, University of Birmingham, United Kingdom
Klaus Troitzsch, University of Koblenz-Landau, Germany
Christopher Tubb, University of Wales, Newport, United Kingdom
Domenico Umbrello, University of Calabria, Italy
Roland Wertz, Fraunhofer Institute IPA, Stuttgart, Germany
Edward Williams, University of Dearborn, USA
Ivan Yatchev, Technical University Sofia, Bulgaria

PREFACE

The European Conference on Modelling and Simulation has provided, over the years, a forum for scientific discussions for simulationists from Europe, Asia, Australia, North and South America and Africa. It is this broad international participation at the ECMS conferences that has prompted a bold change of venue for the ECMS2010. This year's conference, the 24th ECMS is held in Kuala Lumpur, Malaysia, and is hosted by the University of Nottingham.

Organised under the headline "Simulation Meets Global Challenges" the conference emphasises the fact that in today's economy and society the most challenging problems have a global nature. The availability of energy sources and water supply, the challenges of logistics, food production and environmental sustainability are all global concerns. This is why the conference emphasises the importance of networking of individuals and research organisations addressing the same issues from different geographical and disciplinary perspectives. With the outsourcing of much of industrial production from Europe and the USA to developing countries in Asia there is much to be gained from exchanging experiences relating to social and environmental impact of industrialization; so as to build simulation models that can serve as a reliable source of foresight in decision-making processes.

The University of Nottingham, Malaysia Campus is pleased to host the conference as its focus relates closely to the University's commitment to providing our students with a global perspective on real-world problems.

The ECMS 2010 is organised around 10 thematic tracks, which highlight the breadth and relevance of simulation modelling methodology to problem solving. The thematic tracks such as Agent-Based Simulation, Simulation of Complex Systems, Multi-Resolution and Granular Modelling and Intelligent Systems Simulation provide theoretical underpinnings for a broad range of applications while the tracks such as Simulation in Industry, Business and Services, Simulation in Electrical and Electromechanical Engineering and Simulation and Control of Technological Processes provide a forum for the discussion of domain-specific simulation results. However, the list of application areas represented by the conference tracks is not exhaustive and the conference programme includes some papers dealing with computer simulations in agriculture and crop science which are certain to become much more prominent in future.

The track record of 24 years of ECMS conferences is rooted in the vibrancy of the international community of simulationists who play a key role in helping to understand complex real-life problems through simulation modelling studies. We, the organisers of the ECMS2010, are confident that the cross-fertilisation of ideas afforded through formal presentations and informal discussions at the conference will provide a step forward in tackling global challenges.

We wish all participants a fruitful conference and an enjoyable stay in Kuala Lumpur.

Andrzej Bargiela
General Conference and Programme Chair

Sayed Azam Ali
General Conference Co-Chair

David Crowley
General Programme Co-Chair

Eugène J.H. Kerckhoffs
Honorary Conference Chair

Logo Design by www.thisisoctarine.com Copyright 2010

TABLE OF CONTENTS

Plenary Papers

Papers submitted to the following topics

Agent-Based Simulation
Simulation of Complex Systems and Methodologies
Electrical and Electromechanical Engineering
Simulation, Experimental Science and Engineering
Finance and Economics
Simulation in Industry, Business and Services
Simulation Applications in Industry
Simulation of Intelligent Systems
Modelling, Simulation and Control of Technological Processes
Multi-Resolution and Granular Modelling
Late papers

ECMS 2010
SCIENTIFIC PROGRAM

Plenary Papers

COLLABORATIVE GRANULAR MODELING AND SIMULATION

Witold Pedrycz
Department of Electrical & Computer Engineering
University of Alberta, Edmonton Canada
and
School of Computer Science, University of Nottingham, Nottingham,
NG7 2RD, UK.
and
Systems Research Institute, Polish Academy of Sciences
Warsaw, Poland
e-mail: pedrycz@ee.ualberta.ca

KEYWORDS fuzzy model, family of fuzzy models, principle of justifiable granularity, information granularity, collaboration

ABSTRACT
With the remarkably diversified plethora of design methodologies and algorithmic pursuits present today in system modeling including fuzzy modeling, we also witness a surprisingly high level of homogeneity in the sense that the resulting models are predominantly concerned with and built by using a data set coming from a single data source.
In this study, we introduce a concept of collaborative granular modeling. In a nutshell, we are faced with a number of separate sources of data and the resulting individual models formed on their basis. An ultimate objective is to realize modeling at the global basis by invoking effective mechanisms of knowledge sharing and collaboration. In this way, each model is formed not only by relying on a data set that becomes locally available but also is exposed to some general modeling perspective by effectively communicating with other models and sharing and reconciling revealed local sources of knowledge.
Several fundamental modes of collaboration (by varying with respect to the levels of interaction) are investigated along with the concepts of collaboration mechanisms leading to the effective way of knowledge sharing and reconciling or calibrating the individual modeling points of view. The predominant role of information granules with this regard is stressed.
For illustrative purposes, the underlying architecture of granular models investigated in this talk is concerned with rule-based topologies and rules of the form "if R_i then f_i" with R_i being a certain information granule (typically set, fuzzy set or rough set) formed in the input space and f_i denoting any local model realizing a certain mapping confined to the local region of the input space and specified by R_i.

1. INTRODUCTORY COMMENTS

Fuzzy modeling (Angelov et al., 2008; Crtespo and Weber, 2005; Kacprzyk and Zadrozny, 2005; Kilic et al, 2007; Molina et al., 2006; Pedrycz and Gomide, 1998; Pham and Castellani, 2006) exhibits a surprisingly diversity of design methodologies. The concepts and architectures of neurofuzzy systems, evolutionary fuzzy systems are becoming more present in the literature. In spite of this variety, there is one very visible development aspect that cuts across the entire field of fuzzy modeling, that is fuzzy models are built on around a single data set. What becomes more apparent nowadays is a tendency of modeling a variety of distributed systems or phenomena, in which there are separate data sets, quite often quite remote in terms of location or distant in time. The same complex phenomenon could be perceived and modeled using different data sets collected individually and usually not shared. The data might be expressed in different feature spaces as the view at the process could be secured from different perspectives. The models developed individually could be treated as a multitude of sources of knowledge. Along with the individual design of fuzzy models, it could be beneficial to share sources of knowledge (models), reconcile findings, collaborate with intent of forming a model, which might offer a global, unified, comprehensive and holistic view at the underlying phenomenon. Under these

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

circumstances an effective way of knowledge sharing and reconciliation through a sound communication platform becomes of paramount relevance, see Figure 1.

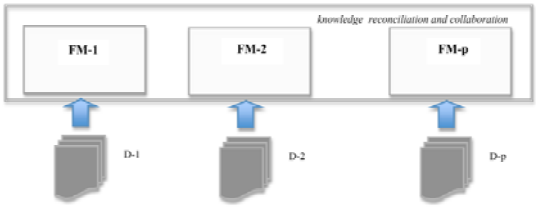


Figure 1: A General Platform of Knowledge Reconciliation and Collaboration in Fuzzy Modeling

A situation portrayed in Figure 1 is shown in a somewhat general way not moving into the details. It is essential to note that the mechanisms of collaboration and reconciliation are realized through passing information granules rather than detailed numeric entities.

The general category of fuzzy models under investigation embrace models described as a family of pairs $<R_i, f_i>$, i=1, 2, …,c. In essence, these pairs can be sought as concise representations of rules with R_i forming the condition part of the i-th rule and f_i standing in the corresponding conclusion part. It is beneficial to emphasize that in such rules, we admit a genuine diversity of the local models formalized by f_i. From the modeling perspective the expression $f_i(\mathbf{x},\mathbf{a}_i)$ could be literally *any* modeling construct, namely

- fuzzy set,
- linear or nonlinear regression function,
- difference or differential equation,
- finite state machine,
- neural network

One can cast the fuzzy models in a certain perspective by noting that by determining a collection of information granules (fuzzy sets) R_i, one establishes a certain view at the system/phenomenon. Subsequently, the conclusion parts (f_i) are implied by the information granules and their detailed determination is realized once R_i have been fixed or further adjusted (refined).

In light of the discussion on knowledge reconciliation and mechanisms of collaboration, it becomes apparent that the interaction can focus on information granules R_i and communication schemes that invoke exchange of granules whereas conclusion parts can be adjusted accordingly once the collaborative development of information granules has been completed.

The main objectives of the study, that is reflected by the organization of the material, is to formulate and discuss a variety of collaborative models of fuzzy models as well as highlight the design principles. The constructs resulting through such collaboration give rise in one way or another to granular constructs of higher order, where the elevated level of granularity is a consequence of reconciliation of knowledge coming from the individual models. The principle of justifiable granularity is presented and shows how granularity emerges as a result of summarization of numeric information (and numeric membership values, in particular). A number of collaborative schemes are discussed where we identify main concepts and present some general ways in which such schemes can be realized. We also show how type-2 fuzzy sets (including interval-valued fuzzy sets) are formed as an immediate result of collaboration.

Throughout this study, we adhere to the standard notation. In particular information granules – fuzzy sets are denoted by capital letters. The notation and terminology is the one being in common usage in the area of fuzzy sets.

2. THE PRINCIPLE OF JUSTIFIABLE GRANULARITY

The essence of the principle of justifiable granularity (Pedrycz, 2005) is that a meaningful representation of a collection of numeric values (real numbers), say $\{x_1, x_2, \ldots, x_N\}$ can be realized as a certain information granule (Bargiela and Pedrycz, 2003, 2008; Zadeh, 1997, 2005) rather than a single numeric entity, no matter how such single individual has been selected. What is being done in statistics is an example of this principle that is realized in the language of *probabilistic* information granules. A sample of numeric data is represented not only by its mean or median (which is a very rough description) but also by the standard deviation. Both the mean and the standard deviation imply a realization of a certain probabilistic information granule, such as e.g., a Gaussian one. The probabilistic information granules are just one of the possibilities to construct an information granule to represent a collection of numeric data.

In case of other formalisms of information granulation, the development of the corresponding granules is guided by a certain optimization criterion. In general, in such criteria, we manage two conflicting requirements. The one is about forming an

information granule of sufficiently high level of experimental evidence accumulated behind it and in this way supporting its existence. The second one is about maintaining high specificity of the resulting information granule.

We discuss several general cases to venture into more algorithmic details of the realization of information granules. We show a construction of interval-based information granules as well as information granules represented as fuzzy sets.

(a) the design of interval-based information granule of numeric data. The data are illustrated in Figure 2.

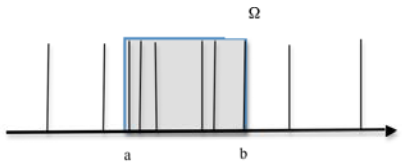


Figure 2: Realization of the Principle of Justifiable Granularity For Numeric Data and Interval Form of Information Granules

We span the numeric interval Ω (=[a, b]) in such a way that (i) the numeric evidence accumulated within the bounds of Ω is as high as possible. We quantify this requirement by counting the number of data falling within the bounds of Ω, that is card{ $x_k \in \Omega$}, which has to be maximized. At the same time, we require that (ii) the support of Ω is as low as possible, which makes Ω specific (detailed) enough. These two requirements are in conflict. A possible way to combine them into a single criterion is to consider the ratio

$$Q = \frac{\text{card}(x_k \in \Omega)}{\text{supp}(\Omega)} = \frac{\text{card}(x_k \in \Omega)}{|b - a|} \qquad (1)$$

which is maximized with regard to the end-points of the interval, namely $\max_{a,b} Q$. The modified version of (1), which offers more flexibility in the development of the information granule involves a decreasing function of the length of the interval |b-a, say| f(|b-a|) which along with some parameters helps control an impact of granularity of the interval on the maximization of Q. For instance, we consider $f(|b-a|) = \exp(-\alpha|b-a|)$ with a being a certain parameter assuming positive values. The Q reads as

$$Q = \text{card}(x_k \in \Omega) \exp(-\alpha |b - a|) \qquad (2)$$

(b) here we design the interval information granule considering that the numeric data come with membership values, that is we are concerned with the pairs (x_k, μ_k) where μ_k stands for the k-th membership value. This specific design scenario is included in Figure 3.

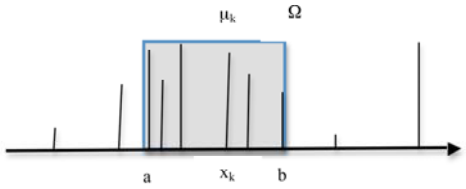


Figure 3: Realization of the Principle of Justifiable Granularity for Numeric Data with Membership -graded (weighted numeric data) and Interval Form of Information Granules

The same design development as discussed in (a) applies here. As each data point comes with the associated membership value, the numeric evidence accumulated within Ω has to be computed in such a way that they are present in the calculations and contribute to the accumulated experimental evidence behind Ω. We determine the σ-sum of the evidence, that is

$\sum_{x_k \in \Omega} \mu_k$ This leads to the maximization of the following performance index

$$Q = \frac{\sum_{x_k \in \Omega} \mu_k}{\text{supp}(\Omega)} \qquad (3)$$

Alternatively, we can focus on the formation of the information granule Ω, which leads to the minimum of changes of the membership grades of the corresponding data. To admit x_k with membership μ_k to Ω, we need to change (elevate) the membership grade and this change is equal to $1-\mu_k$. Similarly, if we exclude x_k from Ω, the corresponding change (suppress) in membership value is μ_k. Refer to Figure 4. The criterion of interest is that of the sum of all possible changes made to the membership grades. We construct Ω in such a way that the changes in membership values are as low as possible. Formally, the performance index is expressed as

$$Q = \sum_{x_k \in \Omega} (1 - \mu_k) + \sum_{x_k \notin \Omega} \mu_k \qquad (4)$$

and its minimization leads to the interval-type of information granule,

$$\text{Min}_{a,b:\, a<b}\, Q \qquad (5)$$

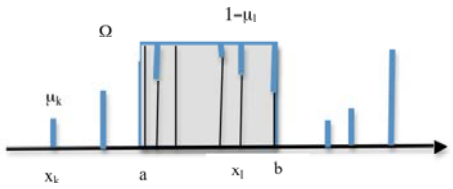


Figure 4: The Design of Interval Information Granule Realized Through the Minimization of the Criterion of Modification of Membership Grades

If the constructed information granule of interest is a fuzzy set rather than the interval, the above considerations are slightly revisited to account for membership degrees of the information granule A. An example of this type of optimization is illustrated in Figure 5.

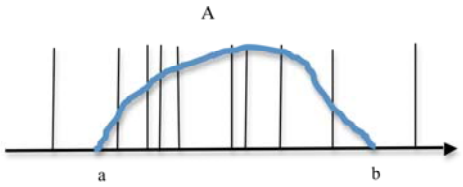


Figure 5: Realization of the Principle of Justifiable Granularity for Numeric Data and Information Granule Represented as Fuzzy Set A

The position of the modal value of A is determined by taking the numeric representative of the data (say, mean or median). Typically, to arrive at semantically meaningful A, we require that the membership function A is unimodal. Given some type of the fuzzy set (say, triangular, parabolic, etc), the optimization of the spreads of the fuzzy set is realized independently for the left- and right-hand spread. The performance index considered here is a slightly modified version of (1), that is

$$Q = \frac{\sum_{k=1}^{N} A(x_k)}{\text{supp}(A)} \tag{6}$$

Considering the fixed form of the membership functions, here are two optimization problems of parametric character: Min_a Q and Min_b Q.
Some further flexibility can be added to the problem by introducing a parameter-enhanced version of Q, which reads as follows

$$Q = \frac{\sum_{k=1}^{N} A^{\gamma}(x_k)}{\text{supp}(A)} \tag{7}$$

where $\gamma > 0$. For $\gamma \in [0,1]$ there is less emphasis is placed on the membership values in the sense these values are "inflated". Note that if $\gamma \rightarrow 0$ then (6) reduces to the previous interval type of information granule. In a more general setting one can consider any continuous and increasing function g_1 of $\sum_{k=1}^{N} A(x_k)$, that is $g_1(\sum_{k=1}^{N} A(x_k))$ and a decreasing function g_2 of supp(A) that is $g_2(\text{supp}(A))$.
In case, the numeric data are associated with some membership values (μ_k), those are taken into account in the modified version of the performance index, which includes these values

$$Q = \frac{\sum_{k=1}^{N} \mu_k A(x_k)}{\text{supp}(A)} \tag{8}$$

All the algorithms realizing the principle of justifiable granularity produce an information granule (either an interval or a fuzzy set) based on a collection of numeric data. The nature of the numeric data themselves can be quite different. Two situations are worth highlighting here:

(a) The numeric data could result from measuring some variables present in the system. In this case, information granules are treated as non-numeric data, which can be then used in the design of the model and highlight the structure of a large number of numeric data.

(b) The numeric data are membership values of some fuzzy sets reported for a certain point of the universe of discourse. The granular representation resulting from the discussed construct gives rise to the information granule of higher type, fuzzy set of type-2, to be more specific. Depending on the nature of the information granule formed here, we arrive at interval-valued type-2 fuzzy sets or just type-2 fuzzy sets.

The principle of justifiable granularity can be used in case of functions, which are then made granular. Given the pairs of input-output data ($\mathbf{x}_k$, y_k), k=1, 2,…, N and having the best numeric mapping "f" we realize its granular mapping, say interval-like format (f_-, f_+). The level of granularity expressed here as the integral of

difference between the bounds. The objective is to make the value of the integral as low as possible while "covering" as many output data as possible, see Figure 6.

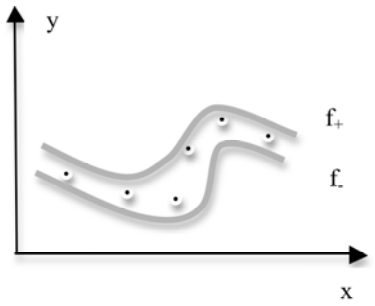


Figure 6: Realization of the Principle of Justifiable Granularity: From Numeric Mapping to its Granular Realization

3. KNOWLEDGE RECONCILIATION: MECHANISMS OF COLLABORATION

The collaboration in the formation of fuzzy models is mostly focused on the collaborative formation of information granules as they form a backbone of fuzzy models. The conclusion parts are mostly realized based on the locally available data and they are constructed once the information granules have been established.
Here there are a number of possible mechanisms of interaction between the individual models when exchanging the findings about the structure of information granules. In contrast to the hierarchical mode of collaboration (to be discussed in Section 4), the mechanisms presented here can be referred to as a one-level collaboration. The form of interaction depends on the level of compatibility considering available spaces of data and spaces of features (inputs) and commonalities among them. Refer to Figure 7.

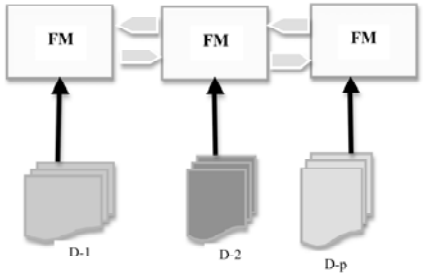
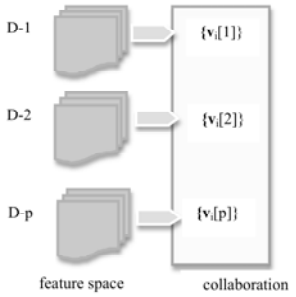
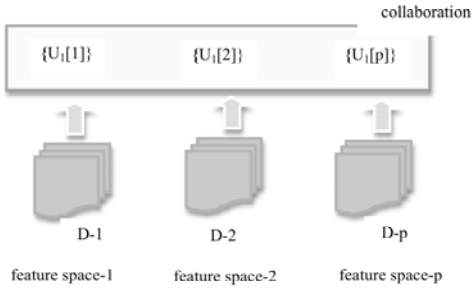


Figure 7: Collaboration Among Fuzzy Models Realized Through Communication At the Level of Information Granules

The findings of the corresponding character are exchanged (communicated among the models) and actively used when carrying out information granulation at the level of the individually available data sets. In what follows, we elaborate on the main aspects of the collaboration modes referring the reader to the literature on their algorithmic details (Pedrycz, 2005; Pedrycz and Rai, 2008). The taxonomy provided here is based on the commonalities encountered throughout the individual data sources. Those could be present in terms of the same feature space or the same data being expressed in different feature spaces.
Collaboration through exchange of prototypes Here, as shown in Figure 8, the data are described in the same feature space and an interaction is realized through prototypes produced locally.



(a)



(b)

Figure 8: A Schematic View at Collaboration Through Exchange of Prototypes (a) and (b) Partition Matrices

Collaboration through exchange of partition matrices Here the data are described in different feature spaces (they might overlap but are not the same). The data in each data set are the same but described in different feature spaces. The exchange of findings and collaboration is realized through interaction at the level of partition matrices. Note that these matrices

abstract from the feature spaces (the spaces do not appear there in an explicit way) but the corresponding rows of the partition matrices have to coincide meaning that we are concerned with the same data).

Formally the underlying optimization problem can be expressed by an augmented objective function, which is composed of two components

$$Q = Q(D - ii) + \alpha \sum_{\substack{jj=1,\\ jj \neq ii}}^{p} \| \mathbf{G}(ii) - \mathbf{G}(jj) \|^2 \qquad (9)$$

The first one, Q(D-ii) is focused on the optimization of the structure based on the locally available data (so the structure one is looking based on D-ii). The second one is concerned with achieving consistency between the granular structure **G**(ii) and the structure revealed based on other data. The positive weight (α) is used to set up a certain balance between these two components of the augmented objective function (local structure and consistency among the local structures). The notation **G**(ii) is used to concisely denote a collection of information granule obtained there, say **G**(ii) ={ G_1[ii], G_2[ii],…, G_c[ii]}. As mentioned, such granules could be represented (described) by their prototypes or partition matrices.

If we consider the FCM-like optimization (Bezdek, 1981), the objective function can be written down in a more explicit fashion as follows

$$Q = \sum_{i=1}^{c} \sum_{\substack{k=1\\ \mathbf{x}_k \in \mathbf{D} - ii}}^{N} u_{ik}^{m} \| \mathbf{x}_k - \mathbf{v}_i[ii] \|^2 + \alpha \sum_{\substack{jj=1,\\ jj \neq ii}}^{p} \| \mathbf{G}(ii) - \mathbf{G}(jj) \|^2 \qquad (10)$$

In case of communication at the level of the prototypes, Figure 8(a), the objective function becomes refined and its term guiding the collaboration effect arises in the form

$$\sum_{\substack{jj=1,\\ jj \neq ii}}^{p} \| \mathbf{G}(ii) - \mathbf{G}(jj) \|^2 = \sum_{i=1}^{c} \sum_{\substack{jj=1,\\ jj \neq ii}}^{p} \| \mathbf{v}_i[ii] - \mathbf{v}_i[jj] \|^2 \qquad (11)$$

For the communication with the aid of partition matrices, Figure 8(b) the detailed expression for the objective function reads as follows

$$\sum_{\substack{jj=1,\\ jj \neq ii}}^{p} \| \mathbf{G}(ii) - \mathbf{G}(jj) \|^2 = \sum_{i=1}^{c} \sum_{k=1}^{N} \sum_{\substack{jj=1,\\ jj \neq ii}}^{p} (u_{ik}[ii] - u_{ik}[jj])^2 \| \mathbf{v}_i[ii] - \mathbf{v}_j[jj] \|^2 \qquad (12)$$

It could be noted that there is a certain direct correspondence between the prototypes and the partition matrix in the sense that each of one could be inferred given that the other one has been provided. More specifically, we envision the following pair of mappings supplying equivalence transformations,

$$\{U, D\} \rightarrow \mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_c\}$$
$$\{\mathbf{V}, D\} \rightarrow U \qquad (13)$$

This transformation can bring a certain unified perspective at the mechanisms of exchange of information granules. For instance, one can convey a collection of the prototypes and they can induce a partition matrix over any data set.

4. KNOWLEDGE RECONCILIATION: A HIERARCHY OF FUZZY MODELS

The overall schematic view of the hierarchical knowledge reconciliation is presented in Figure 9. The knowledge acquired at the level of the models FM-1, FM-2, …, FM-p is concisely arranged in a certain knowledge signature that is a collection of information granules and the associated local models. To emphasize their origin, let us an extra index in squared brackets. For instance, the knowledge signature coming from the ii-th location is denoted as { <R_i[ii], f_i[ii]>}.

In the hierarchical reconciliation of knowledge, we distinguish two general approaches, which depend on a way in which the knowledge is being utilized. The corresponding visualization of the essence of this mechanism is presented in Figure 9.

Passive approach In this approach, we are provided with the knowledge signature, mainly the information granules Ri[ii], which are reconciled at the higher level of the hierarchy.

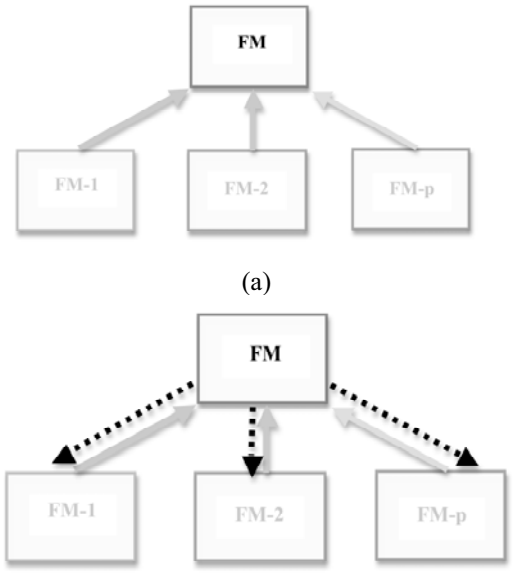


Figure 9: Passive and Active Hierarchical Knowledge Reconciliation Through Interaction Among Fuzzy Models

The prototypes obtained at the lower level are available at the higher level of hierarchy. Here two main directions are sought:

(a) we choose a subset of prototypes, which are the most representative (in terms of a certain criterion) of all $R_i[ii]$s, i=1, 2, …, c_{ii}, ii=1, 2, …, P. Some reconstruction criterion can be involved here using which we express to which extent Ri[ii] are "reconstructed" by the most representative subset of the prototypes. The problem formulated in this way is of combinatorial nature and may invoke the use of methods of evolutionary optimization through which an optimal subset of the prototypes can be established.

(b) the reconciliation of knowledge is realized through clustering of these prototypes, which results in a family of high-level prototypes (and information granules) over which a fuzzy model is constructed. The associated local models are formed on a basis of some data available at the lower level. The approach is called *passive* as the formation of the model at the higher level does not impact (adjust) in any way the local knowledge (models) available at the lower level.

Active approach In contrast to the passive approach, here we allow for some mechanisms of interaction. The results of clustering of the prototypes realized at the higher level of the hierarchy are used to assess the quality of the information granules of the models present at the lower level. For instance, one determines how well a certain information granule is expressed (reconstructed) by the information granules developed at the higher level of the hierarchy. This feedback signal, shown in Fig 9 (b) by a dotted line, is sent back to the lower level where the formation of information granules is guided by the quality of the clusters quantified by the feedback message. The clustering mechanism applied at the lower level is updated and the clusters, which were deemed the least fit are adjusted to become more in line with the findings produced at the higher level of the hierarchy. To accommodate this feedback, the modification of the clustering method can be realized by incorporating the changes to the objective function. For instance, in the FCM, the objective function is adjusted as follows where γ_i is a certain function quantifying the quality of the i-th cluster and supplied from the higher level of the hierarchy. For instance, if this cluster is evaluated there as being irrelevant from the global perspective, then the low values of γ_i used in the objective function discounts the relevance of the i-th cluster at the lower level.

Irrespective of the passive or active approach, in both cases the hierarchical structure gives rise to the hierarchy of information granules. As visualized in Figure 9, there are two stages of information granulation: first information granules are built on basis of data (and those are used in forming fuzzy models at the lower level) and then they are reconciled.

The term information granules of type-2 has to be referred to the original data, so type-1 granulation pertains to the process involving original data whereas the type-2 constructs pertain to information granules built on a basis of prototypes of the information granules at the lower level; in this sense these could be sought of type-2 vis-à-vis original data. Again, here we can distinguish between two ways of forming information granules of higher type, that is (a) selection, and (b) successive granulation. In the selection mode, we choose the most representative subset of information granules from each of the models. As such, this is a combinatorial optimization problem. In the method of successive granulation, the prototypes of information granules produced at the lower level are the objects to be clustered.

It is worth noting that one could have a combination of the hierarchical as well as one-level collaboration mechanisms.

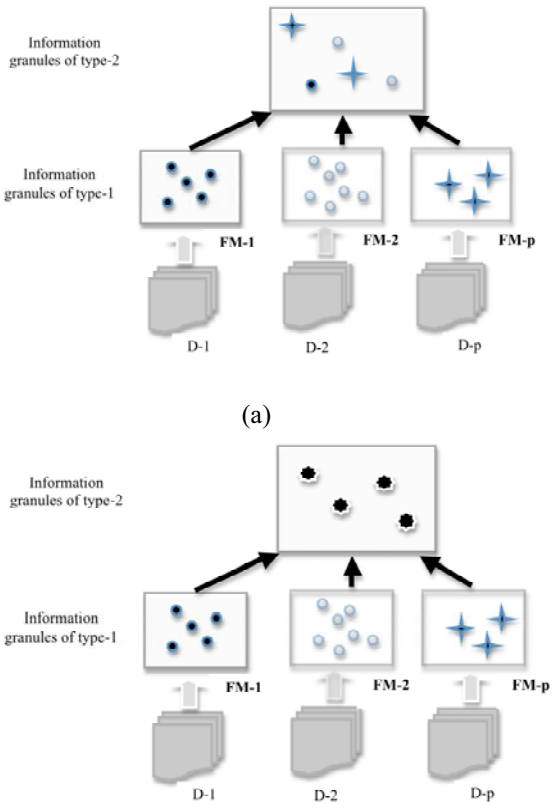


Figure 10: From Data to Information Granules of Type-1 and Type-2 - Two Ways of Design: (a) by Selection, and (b) Successive Granulation Using Prototypes Present at the Lower Level of Hierarchy

5. THE DEVELOPMENT OF FUZZY MODELS OF TYPE-2

The hierarchical way of knowledge reconciliation outlined in Section 4, not matter whether being realized in a passive or active way, leads to fuzzy models, which are inherently associated with information granules and directly engage the principle of justifiable granularity. Let us elaborate on two general cases, which illustrate a way in which granulation of information comes to the picture.

Fuzzy models with granular outputs. The prototypes (or *metaprototypes*, being more descriptive) developed at the higher level are associated with the corresponding outputs. Consider a certain prototype, $\mathbf{v}_i$. We determine the outputs of the fuzzy models present at the lower level, that is FM-1($\mathbf{v}_i$), FM-2($\mathbf{v}_i$), and FM-p($\mathbf{v}_i$). We repeat the same calculations for all other prototypes. Apparently, the mapping from the set of prototypes to the output space is one-to-many as for each input $\mathbf{v}_i$ we typically encounter several different numeric outputs. Through the use of any of the technique of granulation (see Section 2), we arrive at a collection of numeric inputs-granular output pairs of the form

$$\{ (\mathbf{v}_1 , \mathcal{G}(\text{FM-1}(\mathbf{v}_1), \text{FM-2}(\mathbf{v}_1), .., \text{FM-p}(\mathbf{v}_1))), (\mathbf{v}_2 , \mathcal{G}(\text{FM-1}(\mathbf{v}_2), \text{FM-2}(\mathbf{v}_2), \ldots, \text{FM-p}(\mathbf{v}_2))), \ldots (\mathbf{v}_i , \mathcal{G}(\text{FM-1}(\mathbf{v}_i), \text{FM-2}(\mathbf{v}_i), \ldots, \text{FM-p}(\mathbf{vi}_1))), \ldots (\mathbf{v}_c , \mathcal{G}(\text{FM-1}(\mathbf{v}_c), \text{FM-2}(\mathbf{v}_c), \ldots, \text{FM-p}(\mathbf{v}_c))) \} \quad (14)$$

where $\mathcal{G}$(FM-1($\mathbf{v}_i$), FM-2($\mathbf{v}_i$), and FM-p($\mathbf{v}_i$)) is a result of granulation of the corresponding set of numeric outputs of the fuzzy models. Considering the data set (17) visualized schematically in Figure 11, a fuzzy model can be built in different ways, say a collection of rules, neural network with granular outputs, schemes of case-based reasoning (CBR) or fuzzy regression.

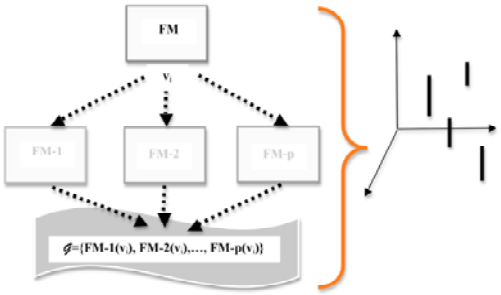


Figure 11: The Formation of Granular Outputs and a Realization of Granular Model of the CBR Architecture

6. CONCLUSIONS

The study has focused on the new category of collaborative fuzzy modeling, which has emerged when dealing with numerous sources of knowledge (local fuzzy models build on a basis of locally accessible data).

Granularity of information plays a pivotal role in fuzzy modeling: while type-1 information granules are the building blocks of fuzzy models, information granules of higher type are essential to formalize and quantify the effect of collaboration and reconciliation of knowledge, which inherently has to quantify a variety of

sources of knowledge coming from individual fuzzy models. The passive and active modes of collaboration help reach a significant level of consensus and, what is equally essential, quantify the level through higher type of information granules. The detailed algorithmic aspects can be realized in different ways and those topics could be a subject of further studies.

The principle of justifiable granularity in application to granular mappings can be exploited in the design of granular models such as granular neural networks, type-2 rule-based systems, or granular regression. In particular, one can consider an admissible level of granularity (treated as a knowledge representation resource), which has to be distributed in an optimal fashion so that the highest level of coverage of numeric data can be achieved. In the formulation of the problem done in this way, one can envision the use of methods of evolutionary optimization as a vehicle to allocate optimal levels of granularity to the individual parameters of the model (say, connections of the neural network).

REFERENCES

Angelov P.; Lughofer E.; Zhou X. 2008, "Evolving fuzzy classifiers using different model architectures" *Fuzzy Sets and Systems*, 159, 23, 3160-3182.

Bargiela, A.; Pedrycz, W. 2003. *Granular Computing: An Introduction*, Kluwer Academic Publishers, Dordrecht

Bargiela, A.;Pedrycz, W. 2008. Toward a theory of Granular Computing for human-centered information processing, *IEEE Transactions on Fuzzy Systems*, 16, 2, 320 – 330.

Bezdek, J.C. 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, N. York, N.Y.

Crespo, F.; Weber, R. 2005. A methodology for dynamic data mining based on fuzzy clustering, *Fuzzy Sets & Systems*, 150, 267-284

Kacprzyk, J.; Zadrozny, S. 2005. Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools, *Information Sciences*, 173, 4, 281-304.

Kiliç, K.; Uncu, O.; Türksen, I.B. 2007. Comparison of different strategies of utilizing fuzzy clustering in structure identification, *Information Sciences*, 177, 23, 5153-5162.

Molina, C.; Rodríguez-Ariza, L.; Sánchez, D.; Amparo Vila, M. 2006. A new fuzzy multidimensional model, *IEEE Trans. on Fuzzy Systems*, 14, 6, 897-912.

Pedrycz, W.; Gomide, F.1998. *An Introduction to Fuzzy Sets: Analysis and Design*, MIT Press, Cambridge, MA

Pedrycz, W.; Rai, P. 2008. Collaborative clustering with the use of Fuzzy C-Means and its quantification, *Fuzzy Sets and Systems*, 159, 18, 2399-2427.

Pedrycz,W. 2005. *Knowledge-Based Clustering: From Data to Information Granules*, J. Wiley, Hoboken, NJ

Pham, D.T.; Castellani, M. 2006. Evolutionary learning of fuzzy models, *Engineering Applications of Artificial Intelligence*, 19, 6, 583-592.

Zadeh, L.A. 1997. Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, *Fuzzy Sets and Systems*, 90, 111-117.

Zadeh, L.A. 2005. Toward a generalized theory of uncertainty (GTU) — an outline, *Information Sciences,* 172, 1- 40.

CONSENSUS CLUSTERING AND FUZZY CLASSIFICATION FOR BREAST CANCER PROGNOSIS

Jonathan M. Garibaldi and Daniele Soria
School of Computer Science
University of Nottingham, Jubilee Campus
Wollaton Road, Nottingham, NG14 5AR, UK
Email: jmg@cs.nott.ac.uk

Khairul A. Rasmani
Faculty of Information Technology
Universiti Teknologi MARA
40450 Shah Alam, Selangor, Malaysia
Email: khairulanwar@tmsk.uitm.edu.my

KEYWORDS

Clustering, Validity Indices, Consensus Clustering, Fuzzy Classification, Breast Cancer, Prognosis

ABSTRACT

Extracting usable and useful knowledge from large and complex data sets is a difficult and challenging problem. In this paper, we show how two complementary techniques have been used to tackle this problem in the context of breast cancer. Diagnosis concerns the identification of cancer within a patient; in contrast, *prognosis* concerns the prediction of the ongoing course of the disease, including issues such as the choice of potential treatments such as chemotherapy or drug therapy, in combination with estimation of chances (or length) of survival. Reliable prognosis depends on many factors, including the identification of the *type* of this heterogeneous disease. We first use a consensus clustering methodology to identify core, well-characterised sub-groups (or *classes*) of the disease based on a large database of protein biomarkers from over a thousand patients. We then use fuzzy rule induction and simplification algorithms to generate a simple, comprehensible set of rules for use in future model-based classification. The methods are described and their use is illustrated on real-world data.

INTRODUCTION

Breast cancer, the most common cancer in women (Parkin et al., 2001; Kamangar et al., 2006), is a complex disease characterized by multiple molecular alterations. Current routine clinical management relies on availability of robust clinical and pathologic prognostic and predictive factors to support decision making. Recent advances in high-throughput molecular technologies have supported the evidence of a biologic heterogeneity of breast cancer. We and others have applied protein biomarker panels with known relevance to breast cancer, to large numbers of cases using tissue microarrays, exploring the existence and clinical significance of distinct breast cancer classes (Abd El-Rehim et al., 2005; Ambrogi et al., 2006; Callagy et al., 2003; Jacquemier et al., 2005; Diallo-Danebrock et al., 2007).

Clustering has become a widely used approach to extrapolate important information from data and to separate different groups that share similar characteristics within them. Cluster analysis may be thought of as the discovery of distinct and non-overlapping sub-partitions within a larger population (Monti et al., 2003). Many different clustering techniques are known today, but often only a few selected methods are used in any given domain. Choosing which method to use is not an easy task, as different clustering techniques return different groupings. Consequently, it has been demonstrated (Ambrogi et al., 2006; Soria et al., 2010) that the use of several methods is preferable in order to extract as much information as possible from the data.

When using more than one algorithm, it is then common to define a consensus across the results (Kellam et al., 2001) in order to integrate diverse sources of similarly clustered data (Filkov and Skiena, 2003) and to deal with the stability of the results obtained from different techniques. Several approaches have been proposed for this task. Kellam and colleagues (Kellam et al., 2001) identified robust clusters by the implementation of a new algorithm called 'Clusterfusion'. It takes the results of different clustering algorithms and generates a set of robust clusters based upon the consensus of the different results of each algorithm. In essence, a clustering technique is applied to the clustering results. Another approach, suggested by Monti and colleagues (Monti et al., 2003), deals with class discovery and clustering validation tailored to the task of analysing gene expression data. The new methodology, termed 'consensus clustering', provides a method, in conjunction with resampling techniques, to represent the consensus across multiple runs of a clustering algorithm and to assess the stability of the discovered clusters. Filkov and Skiena suggested to exploit the popularity of cluster analysis of biological data by integrating clusterings from existing data sets into a single representative clustering based on pairwise similarities of the clusterings. Their proposed representative clustering was the one that minimised the distance to all the other partitions (Filkov and Skiena, 2003). In another approach, Swift and colleagues used consensus clustering to improve confidence in gene-expression analysis, on the assumption that microarray analysis using clustering algorithms can suffer from lack of inter-method consistency in assigning related gene-expression profiles to clusters (Swift et al., 2004).

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

We adopted an alternative approach, based on calculating a number of external cluster validity indices across a range of cluster solutions produced by alternative clustering algorithms, and using consensus across the cluster validity indices and across methods to reach the overall 'best' number of clusters (Soria and Garibaldi, 2010). This methodology results in a number of well characterised (separate and distinct) groups of breast cancer cases, which may be interpreted as different classes (or types) of breast cancer, with corresponding alternative treatment regimes.

There are many non-fuzzy classification algorithms currently available, for example (Witten and Frank, 2000). However, many of these classification algorithms may be very good in generalisation ability and so be very useful for classifying new instances, but lack of comprehensibility of the generated models. In fact, most of the models generated by non-fuzzy classification algorithms contain numerical values and may not be linguistically interpretable. This makes it harder for the user to utilise the models for decision making purposes. Note that an automated-system, or decision support system, is normally considered as a tool to assist experts or non-experts in decision making. Hence, interpretability of such a system is normally regarded as highly important (Castellano et al., 2006). With interpretability in mind, we recently proposed a novel algorithm to induce a simplified set of linguistic rules (Rasmani et al., 2009) suitable for use in a quantifier-based fuzzy classification system (Rasmani and Shen, 2004). This methodology was applied to the breast cancer classes obtained by our consensus clustering in order to obtain a model-based fuzzy classification system suitable for new cases.

CONSENSUS CLUSTERING

The three-step methodology for elucidating core, stable classes (groups) of data from a complex, multidimensional dataset was as follows:

1. A variety of clustering algorithms were run.

2. The most appropriate number of clusters was investigated by means of cluster validity indices.

3. Concordance between clusters, assessed both visually and statistically, was used to guide the formation of stable 'core' classes of data.

The methodology was applied to a well-known set of data concerning breast cancer patients (Abd El-Rehim et al., 2005) in order to obtain core classes. Once these core classes were obtained, the clinical relevance of the corresponding patient groups were investigated by means of associations with related patient data. All statistical analysis was done using *R*, a free software environment for statistical computing and graphics (Maindonald and Braun, 2003).

Clustering Algorithms
Five different algorithms were used for cluster analysis:

1. Hierarchical (HCA)
2. K-means (KM)
3. Partitioning around medoids (PAM)
4. Adaptive resonance theory (ART)
5. Fuzzy c-means (FCM)

Hierarchical clustering: The hierarchical clustering algorithm (HCA) begins with all data considered to be in a separate cluster. It then finds the pair of data with the minimum value of some specified distance metric; this pair is then assigned to one cluster. The process continues iteratively until all data are in the same (one) cluster. A conventional hierarchical clustering algorithm (HCA) was utilised, utilising Euclidean distance on the raw (unnormalised) data with all attributes equally weighted.

K-means clustering: The K-means (KM) technique aims to partition the data into K clusters such that the sum of squares from points to the assigned cluster centres is minimised. The algorithm repeatedly moves all cluster centers to the mean of their Voronoi sets (the set of data points which are nearest to the cluster centre). The objective function minimised is:

$$J(V) = \sum_{j=1}^{k} \sum_{i=1}^{c_j} ||x_i - v_j||^2$$

where x_i is the i-th datum, v_j is the j-th cluster centre, k is the number of clusters, c_j is the number of data points in the cluster j and $||x_i - v_j||$ is the Euclidean distance between x_i and v_j.

The j-th centre v_j can be calculated as:

$$v_j = \frac{1}{c_j} \sum_{i=1}^{c_j} x_i, \quad j = 1, ..., k.$$

K-means clustering is dependent on the initial cluster centres setting (which, in turn, determines the initial cluster assignment). Various techniques have been proposed for the initialisation of clusters (Al-Daoud and Roberts, 1996), but for this study we used a fixed initialisation of the cluster centres obtained with hierarchical clustering. The number of clusters is an explicit input parameter to the K-means algorithm.

Partitioning around medoids: The partitioning around medoids (PAM) algorithm (also known as the k-medoids algorithm) is a technique which attempts to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the K-means algorithm, PAM chooses data points as centers (the so-called medoids) and then assigns each point to its nearest medoid. A medoid is defined as the

object within a cluster for which the average dissimilarity to all other objects in the cluster is minimal, i.e. it is the most centrally located datum in the given cluster. Dissimilarities are nonnegative numbers that are close to zero when two data points are 'near' to each other and large when the points are very different (Kaufman and Rousseeuw, 1990). Usually, a Euclidean metric is used for calculating dissimilarities between observations.

The algorithm consists of two phases: the *build* phase in which an initial set of k representative medoids is selected and the *swap* phase in which a search is carried out to improve the choice of medoids (and hence the cluster allocations). The *build* phase begins by identifying the first medoid, the point for which the sum of dissimilarities to all other points is as small as possible. Further medoids are selected iteratively through a process in which the remaining points are searched to find that which decreases the objective function as much as possible. Once k medoids have been selected, the *swap* phase commences in which the medoids are considered iteratively. Possible swaps between each medoid and other (non-medoid) points are considered one by one, searching for the largest possible improvement in the objective function. Thiscontinues until no further improvement in the objective function can be found. The algorithm is described in detail in (Kaufman and Rousseeuw, 1990), pp.102–104. The number of clusters is an explicit input parameter to the PAM algorithm.

Adaptive resonance theory: The adaptive resonance theory (ART) algorithm has three main steps (Carpenter and Grossberg, 1987). First, the data are normalised to a unit hypersphere, thus representing only the ratios between the various dimensions of the data. Second, data allocated to each cluster are required to be within a fixed maximum solid angle of the group mean, controlled by a so-called 'vigilance parameter' ρ, namely $X_k \cdot P^i \leq \rho$. However, even when the observation profile and a prototype are closer than the maximum aperture for the group, a further test is applied to ensure that the profile and prototype have the same dominant covariates. This is done in a third step by specifying the extent to which the nearest permissible prototype allocation for the given observation must be on the same side of the data space from the diagonal comprising a vector of ones, $\hat{1}$, using a preset parameter, λ:

$$X_k \cdot P^i \leq \lambda X_k \cdot \hat{1}.$$

The ART algorithm is initialised with no prototypes and creates them during each successive pass over the data set. It has some, limited, sensitivity to the order in which the data are presented and converges in a few iterations. In the ART algorithm the clusters are determined automatically: the number of clusters is not an explicit parameter, although there are parameters that can adjust the number obtained.

Fuzzy c-means: The fuzzy c-means (FCM) algorithm is a generalisation of the K-means algorithm which is based on the idea of permitting each object to be a member of *every* cluster to a certain degree, rather than an object having to belong to only one cluster at any one time. It aims to minimise the objective function:

$$J(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{c}(\mu_{i,j})^m \|x_i - v_j\|^2$$

where n is the number of data points, x_i and v_j are the data points and cluster centres and $\mu_{i,j}$ is the membership degree of data x_i to the cluster centre v_j ($\mu_{i,j} \in [0,1]$). m is called the 'fuzziness index' and the value of $m = 2.0$ is usually chosen. An exhaustive description of this method can be found in (Bezdek, 1974). As for K-means, the number of clusters is an explicit input parameter to FCM.

Cluster Validity

Clustering validity is a concept that is used to evaluate the quality of clustering results. If the number of clusters is not known prior to commencing an algorithm, a cluster validity index may be used to determine the best number of clusters for the given data set. Although there are many variations of validity indices, they are all either based on considering the data dispersion in a cluster and between clusters, or considering the scatter matrix of the data points and the one of the clusters centers. In this study, the following indices were applied to those algorithms for which the number of clusters is an explicit parameter, over a range of number of clusters:

1. Calinski and Harabasz (Maulik and Bandyopadhyay, 2002)
2. Hartigan (Hartigan, 1975)
3. Scott and Symons (Scott and Symons, 1971)
4. Marriot (Marriot, 1971)
5. Trace$W$ (Edwards and Cavalli-Sforza, 1965; Friedman and Rubin, 1967)
6. Trace$W^{-1}B$ (Friedman and Rubin, 1967)

For each index, the number of clusters to be considered was chosen according to the rule reported in Table 1 where i_n is the validity index value obtained for n clusters (Weingessel et al., 1999).

CLUSTERING RESULTS

Patients and Clinical Methods

A series of 1076 patients from the Nottingham Tenovus Primary Breast Carcinoma Series presenting with primary operable (stages I, II and III) invasive breast cancer between 1986-98 was used to evaluate the methodology. Immunohistochemical reactivity for twenty-five proteins, with known relevance in breast cancer including those

Table 1: Different validity indices and their associated decision rules

Index	Decision rule
Calinski and Harabasz	$\min_n((i_{n+1} - i_n) - (i_n - i_{n-1}))$
Hartigan	$\min_n((i_{n+1} - i_n) - (i_n - i_{n-1}))$
Scott and Symons	$\max_n(i_n - i_{n-1})$
Marriot	$\max_n((i_{n+1} - i_n) - (i_n - i_{n-1}))$
Trace$W$	$\max_n((i_{n+1} - i_n) - (i_n - i_{n-1}))$
Trace$W^{-1}B$	$\max_n(i_n - i_{n-1})$

used in routine clinical practice, were previously determined using standard immunocytochemical techniques on tumour samples prepared as tissue microarrays (Abd El-Rehim et al., 2005). Levels of immunohistochemical reactivity were determined by microscopical analysis using the modified H-score (values between 0-300), giving a semiquantitative assessment of both the intensity of staining and the percentage of positive cells.

HCA, K-means, PAM and ART Clustering

The HCA results from our previous study (Abd El-Rehim et al., 2005) were utilised, unaltered. Both the K-means and PAM algorithms were run with the number of clusters varying from two to twenty, as the number of clusters is an explicit input parameter of the algorithms. Given that both algorithms can be sensitive to cluster initialisation and in order to obtain reproducible results, both techniques were initialised with the cluster assignments obtained by hierarchical clustering. For the ART algorithm, the parameters were adjusted to obtain six clusters to match the number of clusters previously obtained by HCA. The best validity index obtained for repeated runs of the algorithm with 20 random initialisations was used to select the final clustering assignment.

Fuzzy C-means Clustering

The fuzzy c-means algorithm did not perform as hoped. When the number of clusters was set as two and three, it appeared that reasonable results were obtained. However, from examination of the membership function of each point assigned to these clusters, it could be seen that it was very close to either $\frac{1}{2}$ or $\frac{1}{3}$, respectively. In other words, every data point was assigned to all the clusters with the same membership. Moreover, when the number of clusters was above three, non-zero memberships were evident for only three clusters and these memberships were similar to the three cluster solution — i.e. for $n > 3$, the $n = 3$ cluster solution was obtained, but with $n - 3$ empty clusters.

The fuzziness index m was altered in an attempt to improve the results obtained, but it was found that little difference in the results was observed until m was close to one. Given that when $m = 1$ fuzzy c-means is equivalent to K-means, this result was not useful. As there are many applications for which the fuzzy c-means technique has been successful (see, for example, (Wang and Garibaldi, 2005)), these results are not easy to explain, but they may have been caused by the fact that our data contains a lot of values close to the extremes of each variable. Although the fuzzy c-means algorithm is widely used in literature, we decided to drop it from further analysis due to its poor performance on our data.

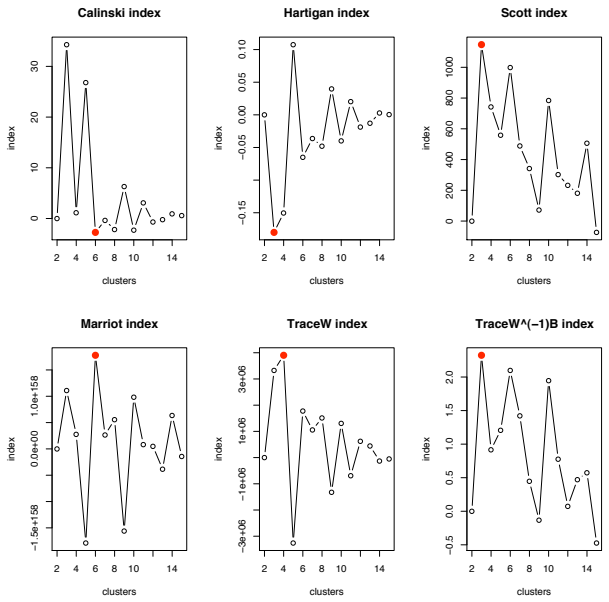


Figure 1: Cluster validity indices obtained for K-means for varying cluster numbers from 2 to 20.

Table 2: Optimum number of clusters estimated by each index for K-means and PAM methods

Index	K-means	PAM
Calinski and Harabasz	6	4
Hartigan	3	4
Scott and Symons	3	4
Marriot	6	4
Trace$W$	4	4
Trace$W^{-1}B$	3	4
Minimum sum of ranks	6	4

Cluster Validity

The values of the decision rule obtained for various values of the validity indices for K-means, for 2 to 20 clusters, are shown in Figure 1. The best number of clusters according to each validity index, for each clustering algorithm, is shown in Table 2, as indicated by the solid circle in Figure 1.

It can be seen that, while there was not absolute agreement among the indices as to which was the best number of clusters for the K-means method, there is good agreement that the best number of clusters for the PAM method is four. Although the best number of clusters varies according to validity index for K-means, on further inspection, it can be seen from Figure 1 that there is more agreement than might be immediately apparent. For example, the Scott and Symons index (which indi-

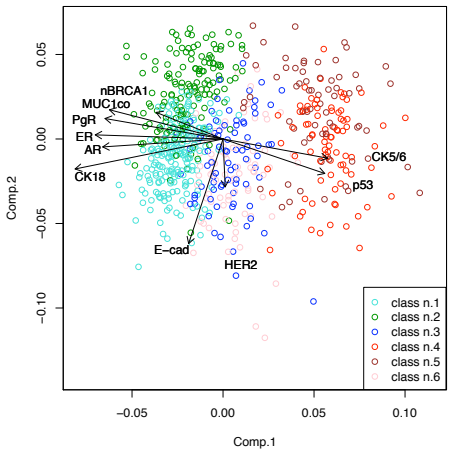


Figure 2: Biplot of classes projected on the first and second principal component axes

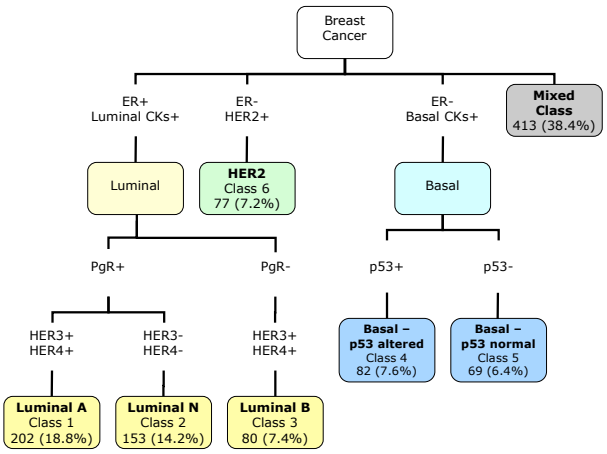


Figure 3: A summary of the classes of breast cancer obtained, with indicative class interpretations.

cated that the best number of clusters was three) indicated that the second best number of clusters was six. Consequently, the indices were used to rank order the number of clusters and the minimum sum of ranks was examined. It was found that the minimum sum of ranks (a form of consensus among the indices) indicated that the overall best number of clusters was six for K-means and four for PAM. However, it was subsequently found that the four cluster solution obtained by PAM was not as clinically interesting as the six cluster solution and it was dropped from further analysis.

Characterisation of Classes

Biplots of the six consensus classes were produced and are shown in Figure 2, in order to provide a visualisation of the separation of the classes. A proposed summary of the essential characterisations of the classes obtained is given in Figure 3, according to the available biopathological knowledge. It is worth noting that class 2, labelled as Luminal-N, and the split of the basal group into two different subgroups depending on p53 levels, appear to be novel findings not previously emphasised in literature.

FUZZY CLASSIFICATION

Fuzzy Subsethood Measures

A fuzzy subsethood measure was originally defined as the degree to which a fuzzy set is a subset of another. However, the definition of fuzzy subsethood value can be extended to calculate the degree of subsethood for linguistic terms in an attribute variable V to a decision class D (Yuan and Shaw, 1995). For linguistic terms $\{A_1, A_2, \ldots, A_n\} \in V$ and $(V, D) \subseteq U$:

$$S(D, A_i) = \frac{\sum_{x \in U} \nabla(\mu_D(x), \mu_{A_i}(x))}{\sum_{x \in U} \mu_D(x)} \qquad (1)$$

where ∇ can be any *t-norm* operator. It should be noted that, to be used for classification problems, both V and D must be defined under the same universe of discourse U (Yuan and Shaw, 1995). Although the decision class is represented by fuzzy sets, this definition allows the decision class with zero fuzziness where the membership value is either one or zero.

Rule Induction

FuzzyQSBA is a rule induction algorithm that was developed by extending the Weighted Subsethood-based Algorithm (WSBA) (Rasmani and Shen, 2006). WSBA has the significant advantage, as compared to previous subsethood-based methods, of not relying on the use of predefined threshold values in generating fuzzy rulesets. The development of WSBA was based on fuzzy subsethood values as defined in Equation (1). Given a training dataset, WSBA induces a fixed number of rules according to the number of possible classification outcomes. To avoid the use of any threshold values in the rule generation process, crisp weights generated using fuzzy subsethood values are created for each of the linguistic terms appearing in the resulting fuzzy rule antecedents. In FuzzyQSBA, fuzzy quantifiers are applied to replace the crisp weights within the rules learned by WSBA. As small changes in the training dataset might cause a change to the entire ruleset, developing a fuzzy model that employs continuous fuzzy quantifiers may be more appropriate compared to two-valued or multi-valued crisp quantifiers (Rasmani and Shen, 2004). (Vila et al., 1997) proposed a continuous fuzzy quantifier which applies linear interpolation between the two classical, extreme cases of the existential quantifier $\exists$ and the universal quantifier $\forall$. In particular, the quantifier was defined such that:

$$Q(A_{ij}, D_k) = (1 - \lambda_Q) T_{\forall, A/D} + \lambda_Q T_{\exists, A/D} \qquad (2)$$

where Q is the quantifier for fuzzy set A relative to fuzzy set D and λ_Q is the degree of neighbourhood of the two extreme quantifiers. The truth values of the existential quantifier $T_{\exists, A/D}$ and the universal quantifier $T_{\forall, A/D}$

were defined as:

$$T_{\exists,A/D} = \Delta_{k=1}^{N}\mu(a_k)\nabla\mu(d_k) \quad (3)$$

$$T_{\forall,A/D} = \nabla_{k=1}^{N}(1-\mu(d_k))\Delta\mu(a_k) \quad (4)$$

where a_k and d_k are the membership functions of fuzzy sets A and D respectively, ∇ represents a *t-norm* and Δ represents a corresponding *t-conorm*. By using fuzzy subsethood values as the *degree of neighbourhood* (λ_Q) of the quantifiers, any possible quantifiers that exist between the existential and universal quantifiers can be created in principle. Initially, all linguistic terms of each attribute are used to describe the antecedent of each rule. This may look tedious, but the reason for keeping this complete form is that every linguistic term may contain important information that should be taken into account. The continuous fuzzy quantifiers are created using information extracted from data and behave as a modifier for each of the fuzzy terms. The resulting FuzzyQSBA ruleset can be simply represented by:

$$R_k = \underset{i=1\ldots m}{\nabla}\Big(\underset{j=1\ldots n}{\Delta}(Q(A_{ij},D_k)\nabla\mu_{A_{ij}}(x))\Big), \quad k=1,2,\ldots,n \quad (5)$$

where $Q(A_{ij},D_k)$ are fuzzy quantifiers and $\mu_{A_{ij}}(x)$ are fuzzy linguistic terms. As both the quantifiers and the linguistic terms are fuzzy sets, choices of *t-norm* operators can be used to interpret $\nabla(Q(A_{ij},D_k),\mu_{A_{ij}}(x))$ whilst guaranteeing that the inference results are fuzzy sets. Based on the definitions of the fuzzy subsethood value, fuzzy existential quantifier and fuzzy universal quantifier (Equations (1,3,4)), it can be proved that if λ_Q is equal to zero then the truth-value of quantifier Q will also equal zero. Thus, during the rule generation process, the emerging ruleset is simplified as any linguistic terms whose quantifier has the truth-value of zero will be removed automatically from the fuzzy rule antecedents, reducing considerably the seeming complexity of the learned ruleset. As commonly used in rule-based systems for classification tasks, the concluding classification will be that of the rule whose overall weight is the highest amongst all.

Rule Extraction

Fuzzy quantifiers have been employed in FuzzyQSBA with the intention to increase the readability of the resulting fuzzy rules and to improve the transparency of the rule inference process. However, the structure of the rules is still very complex. Thus, although the use of quantifiers will make the rules more readable, it seems that it does not increase the comprehensibility of the fuzzy rules. As an alternative, a rule simplification process that is based on fuzzy quantifiers is proposed below. In (Bordogna and Pasi, 1997), fuzzy quantifiers are suggested to be used as a fuzzy threshold. The basic idea of a fuzzy threshold is extended here to conduct the rule simplification process for FuzzyQSBA. This is to offer flexibility in accepting or rejecting any particular linguistic term to represent a particular linguistic variable in a fuzzy rule. To employ the rule simplification, the following fuzzy quantifiers and fuzzy antonym quantifiers are proposed:

$$T_Q(\eta) = \begin{cases} 1 & \text{if } T_Q(\lambda) \geq \eta, \\ \frac{T_Q(\lambda)}{\eta} & \text{if } T_Q(\lambda) < \eta \end{cases} \quad (6)$$

$$T_{antQ}(\eta) = \begin{cases} 1 & \text{if } T_Q(\lambda) \leq 1-\eta, \\ \frac{1-T_Q(\lambda)}{\eta} & \text{if } T_Q(\lambda) > 1-\eta \end{cases} \quad (7)$$

where $T_Q(\lambda)$ is the truth value of quantifier (TVQ) associated with each linguistic term in Equation (5) and η is a threshold value that can be defined as:

$$\eta = p \times \omega \quad (8)$$

where p is a factor for the maximum TVQ, ω. In this technique, the decision to accept a particular linguistic term is made locally without affecting other variables. The aim of using a fuzzy threshold is to soften the decision boundary in the process of accepting or rejecting any terms to be promoted as antecedents of a fuzzy rule, whilst at the same time significantly reducing the number of terms in the induced fuzzy rules. The fuzzy quantifiers mentioned above can be interpreted as 'at least η' and its antonym 'at most $1-\eta$'.

Rule Simplification

The rule simplification algorithm is as follows:

1. For each variable, select the maximum TVQ and calculate $T_Q(\eta)$ and $T_{antQ}(\eta)$ for each linguistic term.
2. For $i = 1, 2, \ldots, l$ where l is the number of linguistic terms for a variable, and for $m \neq n$, calculate:

$$\delta(T_{Q_i}(\eta)) = |T_{Q_m}(\eta) - T_{Q_n}(\eta)|$$

$$\delta(T_{antQ_i}(\eta)) = |T_{antQ_m}(\eta) - T_{antQ_n}(\eta)|$$

3. Conduct the following test: if $\min_i\{\delta(T_{Q_i}(\eta))\} \geq \{\delta(T_{antQ_i}(\eta))\}$ then choose the negation of terms with the lowest TVQ to represent the conditional attribute; else choose the term with the highest TVQ.
4. Create a simplified rule using the accepted linguistic terms (or negation of the terms).

Note that when $\eta = 1$, the fuzzy quantifier and its antonym will become 'most' and 'least', and when $\eta = 0$ the quantifier and its antonym will become 'there exists at least one' and 'for all'. By using the technique proposed above, the primary terms with higher TVQs are accepted to represent the antecedents of the fuzzy rules. By lowering the value of η, the primary terms with a lower TVQ will gradually be accepted. The idea behind this technique is that only the dominant linguistic term (or its negation) will be chosen to represent a particular linguistic variable.

CLASSIFICATION RESULTS

The results of the automated rule induction and simplification obtained using the FuzzyQSBA algorithms described above is shown in Table 3. It can be seen that there is a very good correspondence between the automatically induced rules and the characterisation of the classes obtained from clinical experts shown in Figure 3. Note that the term 'luminal CKs' refers to CK5/6, CK14 and CK18, whereas 'basal CKs' refers to CK7/8 and others. However, in Table 3, the absence of luminal CKs defines membership of classes 4 and 5, as opposed to the presence of basal CKs as mentioned in Figure 3.

CONCLUSIONS

In this paper, we have illustrated the use of consensus clustering to elucidate six separate and distinct classes from the original data set. Further clinical investigations have confirm that these classes form well-chacterised sub-types of breast cancer with distinct clinical characteristics (Soria et al., 2010). We have then presented a rule simplification process (Rasmani et al., 2009) to accompany the FuzzyQSBA rule induction algorithms described previously (Rasmani and Shen, 2006) which results in a simple, comprehensible classification table for each of the six classes based on only ten biomarkers.

In future, we aim to implement the resultant fuzzy rule table in a model-based classification system that can be used to determine the type (class) of cancer in new patients presenting with breast cancer. We hope to thereby create a clinically useful decision support tool for assisting in the choice of treatment(s) for breast cancer, to improve patient survivability and quality of life (by ensuring appropriate treatments) and to reduce health service costs (by reducing unnecessary treatments).

REFERENCES

Abd El-Rehim, D., Ball, G., Pinder, S., Rakha, E., Paish, C., Robertson, J., Macmillan, D., Blamey, R., and Ellis, I. (2005). High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. *Int. Journal of Cancer*, 116:340–350.

Al-Daoud, M. and Roberts, S. (1996). New methods for the initialisation of clusters. *Pattern Recognition Letters*, 17(5):451–455.

Ambrogi, F., Biganzoli, E., Querzoli, P., Ferretti, S., Boracchi, P., Alberti, S., Marubini, E., and Nenci, I. (2006). Molecular subtyping of breast cancer from traditional tumor marker profiles using parallel clustering methods. *Clinical Cancer Research*, 12(3):781–790.

Bezdek, J. (1974). Cluster validity with fuzzy sets. *Journal of Cybernetics*, 3(3):58–73.

Bordogna, G. and Pasi, G. (1997). Application of OWA operators to soften information retrieval systems. In *The Ordered Weighted Averaging Operators: Theory, Methodology and Applications*. LNCS.

Callagy, G., Cattaneo, E., Daigo, Y., Happerfield, L., Bobrow, L., Pharoah, P., and Caldas, C. (2003). Molecular classification of breast carcinomas using tissue microarrays. *Diagn Mol Pathol*, 12:27–34.

Carpenter, G. and Grossberg, S. (1987). ART2: Stable self-organization of pattern recognition codes for analogue input patterns. *Applied Optics*, 26:4919–4930.

Castellano, G., Fanelli, A., and Mencar, C. (2006). Classifying data with interpretable fuzzy granulation. In *Proceedings of the 3rd International Conference on Soft Computing and Intelligent Systems*.

Diallo-Danebrock, R., Ting, E., Gluz, O., Herr, A., Mohrmann, S., Geddert, H., Rody, A., Schaefer, K., Baldus, S., Hartmann, A., Wild, P., Burson, M., Gabbert, H., Nitz, U., and Poremba, C. (2007). Protein expression profiling in high-risk breast cancer patients treated with high-dose or conventional dose-dense chemotherapy. *Clin Cancer Res*, 13:488–497.

Edwards, A. and Cavalli-Sforza, L. (1965). A method for cluster analysis. *Biometrics*, 21(2):362–375.

Filkov, V. and Skiena, S. (2003). Integrating microarray data by consensus clustering. In *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, pages 418– 426.

Friedman, H. and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62(320):1159–1178.

Hartigan, J. (1975). *Clustering Algorithms*. Wiley series in probability and mathematical statistics. Applied Probability and Statistics. New York: Wiley.

Jacquemier, J., Ginestier, C., Rougemont, J., Bardou, V.-J., Charafe-Jauffret, E., Geneix, J., Adélaïde, J., Koki, A., Houvenaeghel, G., Hassoun, J., Maraninchi, D., Viens, P., Birnbaum, D., and Bertucci, F. (2005). Protein expression profiling identifies subclasses of breast cancer and predicts prognosis. *Cancer Res*, 65:767–779.

Kamangar, F., Dores, G., and Anderson, W. (2006). Patterns of cancer incidence, mortality, and prevalence across five continents: Defining priorities to reduce cancer disparities in different geographic regions of the world. *J Clin Oncol*, 24:2137–2150.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: an Introduction to Cluster Analysis*. Wiley series in probability and mathematical statistics. Applied Probability and Statistics. New York: Wiley.

Kellam, P., Liu, X., Martin, N., Orengo, C., Swift, S., and Tucker, A. (2001). Comparing, contrasting and combining clusters in viral gene expression data. In *Proceedings of 6th Workshop on Intelligent Data Analysis in Medicine*.

Maindonald, J. and Braun, W. (2003). *Data Analysis and Graphics Using R — An Example-Based Approach*. Cambridge University Press.

Marriot, F. (1971). Practical problems in a method of cluster analysis. *Biometrics*, 27(3):501–514.

Maulik, U. and Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24(12):1650–1654.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52:91–118.

Table 3: Simplified ruleset created using automated FuzzyQSBA pruning

Classification	Variables									
	ER	PgR	CK18	CK5/6	CK14	HER2	HER3	HER4	P53	MUC1
Class 1	HIGH	HIGH	HIGH	LOW	-	LOW	HIGH	HIGH	-	-
Class 2	HIGH	HIGH	HIGH	LOW	-	LOW	-	LOW	-	HIGH
Class 3	HIGH	LOW	HIGH	LOW	LOW	LOW	-	-	-	-
Class 4	LOW	LOW	-	-	-	LOW	HIGH	-	HIGH	-
Class 5	LOW	LOW	-	-	-	LOW	-	-	LOW	-
Class 6	LOW	-	HIGH	-	-	HIGH	-	HIGH	-	-

Parkin, D., Bray, F., Ferlay, J., and Pisani, P. (2001). Estimating the world cancer burden: Globocan 2000. *Int J Cancer*, 94:153–156.

Rasmani, K., Garibaldi, J., Shen, Q., and Ellis, I. (2009). Linguistic rulesets extracted from a quantifier-based fuzzy classification system. In *Proceedings of IEEE International Conference on Fuzzy Systems*.

Rasmani, K. and Shen, Q. (2004). Modifying fuzzy subsethood-based rule models with fuzzy quantifiers. In *Proceedings of the 13th IEEE International Conference on Fuzzy Systems*.

Rasmani, K. and Shen, Q. (2006). Data-driven fuzzy rule generation and its application for student performance evaluation. *Applied Intelligence*, 24:305–309.

Scott, A. and Symons, M. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, 27(2):387–397.

Soria, D. and Garibaldi, J. (2010). A novel framework to elucidate core classes in a dataset. In *Proceedings of the World Congress on Computational Intelligence*.

Soria, D., Garibaldi, J. M., Ambrogi, F., Green, A. R., Powe, D., Rakha, E., Macmillan, R. D., Blamey, R. W., Ball, G., Lisboa, P. J., Etchells, T. A., Boracchi, P., Biganzoli, E., and Ellis, I. O. (2010). A methodology to identify consensus classes from clustering algorithms applied to immunohistochemical data from breast cancer patients. *Computers in Biology and Medicine*, 40:318–330.

Swift, S., Tucker, A., Vinciotti, V., Martin, N., Orengo, C., Liu, X., and Kellam, P. (2004). Consensus clustering and functional interpretation of gene-expression data. *Genome Biology*, 5:R94.

Vila, M., Cubero, J., Medina, J., and Pons, O. (1997). Using OWA operators in flexible query processing. In *The Ordered Weighted Averaging Operators: Theory, Methodology and Applications*. LNCS.

Wang, X.-Y. and Garibaldi, J. M. (2005). A comparison of fuzzy and non-fuzzy clustering techiques in cancer diagnosis. In *Proceedings of the 2nd International Conference in Computational Intelligence in Medicine and Healthcare*, pages 250–256, Lisbon, Portugal.

Weingessel, A., Dimitriadou, E., and Dolnicar, S. (1999). An examination of indexes for determining the number of clusters in binary data sets. Working Paper No.29.

Witten, I. and Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco.

Yuan, Y. and Shaw, M. (1995). Induction of fuzzy decision trees. *Fuzzy Sets and Systems*, 69(2):125–139.

AUTHOR BIOGRAPHIES

JONATHAN M. GARIBALDI received the BSc (Hons) degree in Physics from Bristol University, UK, and the MSc and PhD degree from the University of Plymouth, UK, in 1984, 1990, and 1997, respectively. He is currently an Associate Professor and Reader within the Intelligent Modelling and Analysis (IMA) Research Group in the School of Computer Science at the University of Nottingham, U.K. Dr Garibaldi has published over 80 papers on fuzzy expert systems and fuzzy modelling, including three book chapters, and has edited two books. His main research interests are modelling uncertainty in human reasoning and especially in modelling the variation in normal human decision making, particularly in medical domains. He has created and implemented fuzzy expert systems, and developed methods for fuzzy model optimisation. His email is `jmg@cs.nott.ac.uk` and his personal webpage is at `http://ima.ac.uk/garibaldi`.

DANIELE SORIA is currently a post-doctoral research associate within the Intelligent Modelling and Analysis Research Group, School of Computer Science, University of Nottingham. Daniele Soria received his BSc and MSc in Applied Mathematics from the University of Milan, Italy, in 2004 and his PhD, entitled 'Novel Methods to Elucidate Core Classes in Multi-Dimensional Biomedical Data', in Computer Science from the University of Nottingham, UK, in 2010. His research interests include data mining, bioinformatics and medical applications. His email is `dqs@cs.nott.ac.uk` and his personal webpage is at `http://ima.ac.uk/soria`.

KHAIRUL A. RASMANI is a lecturer at the Faculty of Information Technology and Quantitative Sciences, Universiti Teknologi MARA, Malaysia. He received his Masters Degree in Mathematical Education from University of Leeds, UK in 1997 and his Ph.D. degree from University of Wales, Aberystwyth, UK in December 2005. His research interests include fuzzy approximate reasoning, fuzzy rule-based systems and fuzzy classification systems. In 2009, he was a visiting researcher in the IMA Research Group, University of Nottingham, UK, during which time the majority of the work on fuzzy rule simplification was carried out.

SCIENTIFIC RESEARCH FUNDING FROM THE EU

David Crowley
University of Nottingham Malaysia Campus
David.Crowley@nottingham.edu.my

KEYWORDS European Community, FP7 Seventh Framework Programme, European Research Council (ERC), The People Programme

ABSTRACT
The *Seventh Framework Programme* for research and technological development (FP7) is the European Union's main instrument for funding research in Europe. FP7, which applies to the years 2007-2013, is the natural successor to the *Sixth Framework Programme* (FP6), and is the result of years of consultation with the scientific community, research and policy making institutions, and other interested parties.
Since their launch in 1984, the Framework Programmes have played a lead role in multidisciplinary research and cooperative activities in Europe and beyond. FP7 continues that task, and is both larger and more comprehensive than earlier Framework Programmes. Running from 2007 to 2013, the programme has a budget of 53.2 billion euros over the seven-year lifespan, the largest funding allocation yet for such programmes.

FP7 has some key differences to earlier EU research programmes, including:

Increased budget – the FP7 budget represents a 63% increase from FP6 at current prices, which means additional resources for European research.
.
Focus on themes – a strong focus on major research themes (e.g. health, ICTs, space, etc.) within the largest component of FP7 – Cooperation – makes the programme more flexible and responsive to the needs of industry.

European Research Council (ERC) – the first pan-European agency for funding research, the newly created *European Research Council*, aims to fund more high-risk yet potentially high-gain European research at the scientific frontiers. The priorities in FP7 are contained within several specific programmes, as follows:

Cooperation programme – the core of FP7
The core of FP7 and its largest component by far, the *Cooperation programme* fosters collaborative research across Europe and other partner countries, according to several key thematic areas. These themes are: health; food, agriculture and fisheries, and biotechnology; information and communications technologies; nanosciences, nanotechnologies, materials and new production technologies; energy; environment (including climate change); transport (including aeronautics); socio-economic sciences and the humanities; space and security. Special attention is also being paid to multi-disciplinary and cross-theme research, including joint calls for proposals between themes.

Ideas programme – and the European Research Council (ERC)
The *Ideas programme* is the first time an EU Framework research programme has funded pure, investigative research at the frontiers of science and technology, independently of thematic priorities. As well as bringing such research closer to the conceptual source, this flagship FP7 programme is a recognition of the value of basic research to society's economic and social welfare.

People programme – boosting European research careers
The *People programme* provides significant support for research mobility and career development, both for researchers inside the European Union and externally. It is being implemented via a coherent set of Marie Curie actions, designed to help researchers build their skills and competences throughout their careers. The programme includes activities such as initial researcher training, support for lifelong training and development via trans-national European fellowships and other actions, and industry/academia partnerships. An international dimension with partners outside the EU is to further develop the careers of EU researchers, by creating international outgoing and incoming fellowships to foster collaboration with research groups outside Europe.

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

Agent-Based Simulation

Simulation of Complex Systems and Methodologies

Electrical and Electromechanical Engineering

Simulation, Experimental Science and Engineering

Finance and Economics

Simulation in Industry, Business and Services

Simulation Applications in Industry

Simulation of Intelligent Systems

Modelling, Simulation and Control of Technological Processes

Multi-Resolution and Granular Modelling

EPMAS: EVOLUTIONARY PROGRAMMING MULTI-AGENT SYSTEMS

Ana M. Peleteiro, Juan C. Burguillo
Telematics Engineering Department
University of Vigo
Campus de Lagoas-Marcosende, 36310 Vigo, Spain
Email: {apeleteiro, J.C.Burguillo}@det.uvigo.es

Zuzana Oplatkova, Ivan Zelinka
Faculty of Applied Informatics
Tomas Bata University in Zlin
Nad Stranemi 4511, 76001 Zlin, Czech Republic
Email: {oplatkova, zelinka}@fai.utb.cz

KEYWORDS

Evolutionary Programming, Grammatical Evolution, Multi-agent Systems

ABSTRACT

Evolutionary Programming (EP) seems a promising methodology to automatically find programs to solve new computing challenges. The Evolutionary Programming techniques use classical genetic operators (selection, crossover and mutation) to automatically generate programs targeted to solve computing problems or specifications. Among the methodologies related with Evolutionary Programming we can find Genetic Programming, Analytic Programming and Grammatical Evolution. In this paper we present the Evolutionary Programming Multi-agent Systems (EPMAS) framework based on Grammatical Evolution (GE) to evolutionary generate Multi-agent systems (MAS) ad-hoc. We also present two case studies in MAS scenarios for applying our EPMAS framework: the predator-prey problem and the Iterative Prisoner's Dilemma.

INTRODUCTION

In our evolving society, the computing and engineering problems we have to solve are becoming more difficult and intractable every day. Future computers and software systems should be able to automatically deal with these new challenges. This is a central topic in Artificial Intelligence discipline (Russell and Norvig, 2003).

Evolutionary algorithms and multiagent systems seem two promising methodologies to automatically find solutions for these new challenges. These algorithms use the classical genetic operators (selection, crossover and mutation) to find an optimal solution, and they have been successfully applied in the automatic management of isolated problems, i.e., evolutionary methods can lead to the realization of artificial intelligent systems (Fogel et al., 1966). Among the methodologies related with Evolutionary Programming we can find Genetic Programming (Koza, 1994), Analytic Programming (Zelinka and Oplatkova, 2003) and Grammatical Evolution (O'Neill and Ryan, 2003). The first methodology (GP) is based on the LISP language and its tree syntax, while the last two ones have the advantage of being language independent. There are several approaches in the literature applying GP (Haynes and Sen, 1996; Calderoni and Marcenac, 1998) to try to solve problems in Multi-agent Systems. However, to our knowledge, this is the first approach using a GE framework for solving general problems in MAS.

In this paper we present an analysis of the conditions needed to apply Evolutionary Programming techniques to create a Multi-agent System able to solve a given problem. Then we propose the use of Grammatical Evolution (GE) to provide 'good' solutions in MAS scenarios. Finally, we present two case studies based on two typical MAS scenarios. The first one is the well known predator-prey problem, in which some predators aim at surrounding the prey. The second one is the Iterative Prisoner's Dilemma, a classical Game Theory scenario, where we aim at finding an optimal solution for several configurations of the famous Axelrod's tournament (Axelrod, 1984).

The rest of the paper is organized as follows: in section GRAMMATICAL EVOLUTION we give a brief introduction to this evolutionary technique. In section A FRAMEWORK FOR EVOLUTIONARY PROGRAMMING MULTI-AGENT SYSTEMS we present our theoretical framework for evolving Multi-agent Systems. Section CASE STUDIES presents two experiments where we test our framework with classical MAS scenarios. Finally, we present the conclusions and our future work.

GRAMMATICAL EVOLUTION (GE)

Grammatical Evolution (GE) (O'Neill and Ryan, 2003) is an evolutionary computation technique based on Genetic Programming (GP) (Koza, 1994). GP provides a systematic method to allow computers to automatically solve a given problem from a high-level statement. The idea is to use evolutionary or genetic operations (selection, crossover and mutation) to generate new populations of computer programs, measuring the performance (i.e., the fitness) for each one, and obtaining the one with the best fitness as the result.

One of the main advantages of GE compared to GP is that it can evolve complete programs in any arbitrary programming language (O'Neill and Ryan, 2003), which provides flexibility to the system under development.

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

GE is a grammar-based form of GP, and this means that we can apply genetic operators to an integer string, subsequently mapped to a program, through the use of a grammar. The grammar describes the output language, using the Backus Naur Form (BNF) (Garshol, 2003), in the form of production rules, and it governs the creation of the program itself. Changing the grammar we can radically change the behavior of the generation strategy.

GE has been applied in many fields, for instance, in technical trading (predicting corporate bankrupt, bond credit rating) (Brabazon and Oneill, 2004), to find trigonometric identities (Ryan et al., 1998), or to solve symbolic regression problems (Ryan and O'Neill, 1998).

A FRAMEWORK FOR EVOLUTIONARY PROGRAMMING MULTI-AGENT SYSTEMS

In this section we provide a description of our framework for using Evolutionary Programming to automatically generate code for Multi-agent Systems (MAS). In order to do this, we need to guarantee that the properties that characterize a MAS are preserved, that we have an iterative process to successively refine the generated agents and that we are able to test somehow the behavior of the MAS to evaluate its performance.

But, first we need to define our concept of agent. Unfortunately, there is no general agreement in the research community about what an agent is. Therefore we cite a general description (Wooldridge and Jennings, 1995), and according to it, the term agent refers to a hardware or (more usually) software-based computer system characterized by the well-known properties:

- **Autonomy**: *agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state*. To achieve this property, we need to obtain a different code for the different agents of the MAS and include, within the code properties, the capability to decide when and how to act depending on internal states and external perceptions.

- **Social ability**: *agents interact with other agents (and possibly humans) via some kind of agent-communication language*. This can be obtained providing properties for explicit or implicit interaction, or communication, among the agents of the system.

- **Reactivity**: *agents perceive their environment, and respond in a timely fashion to changes that occur in it*. To do this we need to generate code for the agents able to react to the different inputs provided by the environment.

- **Pro-activeness**: *agents do not simply act in response to their environment; they are able to exhibit goal-directed behavior by taking the initiative*. To achieve this property, the agent must include code for deciding how to act depending on internal states or external inputs.

However, there are other desirable, although not mandatory, attributes that can be present: *benevolence, rationality* and *adaptability (or learning)* (Wooldridge, 2002). Now we can define a Multi-agent System as a system consisting of an interacting group of agents.

The simulation of agents and their interactions, usually known as agent-based modeling, is the basis for the iterative approach presented in this section in order to successively refine a MAS and achieve the best possible solution for a given problem or scenario.

We have described a MAS as a set of multiple autonomous entities (agents) that interact, and the right interaction is the key to achieve the desired MAS behavior. We need a framework to model distributed problems, and we must define a procedure to solve them by means of successive iterations that shift the MAS behavior to find alternative solutions, and eventually the optimum.

We consider that we can get closer to this conceptual framework by mixing the conceptual modeling of the problem at one level and its actual performance at another one. The algorithm in Table 1 describes the basic steps needed to generate MAS solutions and to evaluate its performance.

Table 1: Generating Evolutionary Solutions for a MAS

1 Generate a new solution for the MAS.
2 Simulate the new solution.
3 Evaluate the results of the simulation, and get the fitness.
4 If a STOP criteria does not hold, go back to (1).

This cycle is an iterative process that successively generates, and ideally improves, the solutions for the MAS behavior. To obtain a final solution we need to describe how to manage the three phases that compose this iterative process:

1. **Generate a solution for the MAS**: to do this we need to be able to generate evolutionary code, different for every agent of the MAS. Thus, we need multiple instances of an Evolutionary Programming tool in order to generate different code for the different agents.

2. **Simulate the MAS behavior**: for this we need to simulate the agents generated in the step 1 within the MAS environment, being able to produce some final result (fitness) to be used as input for the evaluation phase.

3. **Evaluate the results of the simulation and get the fitness**: for this we need to define some type of fitness for the different agents as a measure to indicate

how well they have achieved their individual objectives. We may also define a global fitness to describe if the system as a whole has reached the pursued objective. This individual and global fitness must act as a feedback for the whole system in the generation of a new solution, ideally better than the previous ones.

Now, we propose the use of two open source tools, freely available on the Internet, to generate solutions by means of Evolutionary Programming in MAS (EPMAS).

Grammatical Evolution in Java (GEVA)

Grammatical Evolution in Java (GEVA) (O'Neill et al., 2008) is a free and open source environment for using GE, developed at the UCD's Natural Computing Research & Applications group. This software provides a search engine framework to build the functions, as well as a simple GUI and the genotype-phenotype mapper of GE.

Netlogo

NetLogo (Wilensky, 2007) is a free Multi-agent modeling environment for simulating natural and social phenomena. It is particularly well suited for modeling complex systems that evolve. Modelers can give instructions to hundreds of agents operating independently. This makes it possible to explore the connection between the micro-level behavior of individuals, and the macro-level patterns that emerge from the interaction of many individuals.

This environment has been used to carry out simulations in areas such as biology, medicine, physics, chemistry, mathematics, computer science, economics and social psychology.

NetLogo provides a 2-D world made of agents that simultaneously carry out their own activity. We can model the behavior of these agents, individually and also in group. This makes Netlogo a really useful tool for Multi-agent simulations. Besides, with its user-friendly graphical interface, a user may interact with the system and obtain real-time feedback from the simulations.

A Framework for EPMAS

The two previous tools, i.e., GEVA and Netlogo, solve the needs expressed in Table 1 to create a framework for Evolutionary Programming Multi-agent Systems. GEVA environment has been designed to obtain solutions for a particular function using a pre-defined function population. To generate different functions for the agents in the system we need to create multiple instances of GEVA, each one managing the code for every particular agent (or type of agents) in the MAS. Unfortunately, GEVA has not included the possibility to run several instances of the tool in a shared memory, and this means that we need to use external facilities (files at the OS level) to communicate those processes.

Once we have created the code for all the agents in the system with several GEVA instances, we need to simulate the actual solution for the MAS, and test its behavior in the particular problem it is aimed to solve (step 2 in Table 1). This can be done using the Netlogo simulation tool. In this case, the connection between GEVA and Netlogo has been also achieved by means of file interactions. In Fig. 1 we show a schema of the communication between GEVA and Netlogo to implement the model described in Table 1.

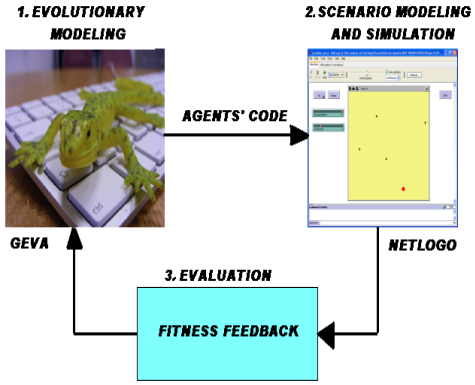


Figure 1: Communication between GEVA and Netlogo.

Finally, we have to define a fitness measurement, dependent from the problem definition, to provide the fitness feedback for the different GEVA instances, in order to generate a new solution for every agent of the MAS.

The algorithm of Table 1 stops when some criterion selected by the system designer has been achieved. This means that the solution complies with certain criteria, for instance, that a number of iterations have been performed, or that a particular set of expected values in the solution has been accomplished.

CASE STUDIES

In this section we describe two example scenarios for MAS, and we present the results of the experiments performed using GE to obtain a solution for each of them. In the following two subsections we describe the experiments, as well as presenting the results obtained.

Predator-prey Scenario

The predator-prey pursuit problem (Benda et al., 1986) is one of the first and well-known testbed for learning strategies in Multi-agent Systems. It consists of a set of agents, named predators, that aim at surrounding another agent, named prey, that must escape from them. This toy problem has been addressed many times by the Artificial Intelligence (AI) Community. Initially, Korf (Korf, 1992) proposed a solution without Multi-agent communication. Since then, a great number of alternatives have emerged usually involving reinforcement learning (Tan, 1997) techniques.

We use Netlogo to simulate the scenario where the prey and the predators perform their chase (Fig. 2). We have

four predators (black arrows) that aim at surrounding and catching one prey (a red cicle).

The predators can move to the North (N), South (S), East (E) and West (W), and they can only see the prey if they are closer than a quarter of the maximum distance of the map scenario. Our predators may use a communication function (*broadcast();*) that allows to communicate the position of the prey to the rest of the predators. But if any predator cannot see the prey, then they move randomly. A predator catches the prey if both are located in the same position. The prey behavior is really simple: it randomly moves to the N, S, E or W all over the map (it is a non-toroidal world).

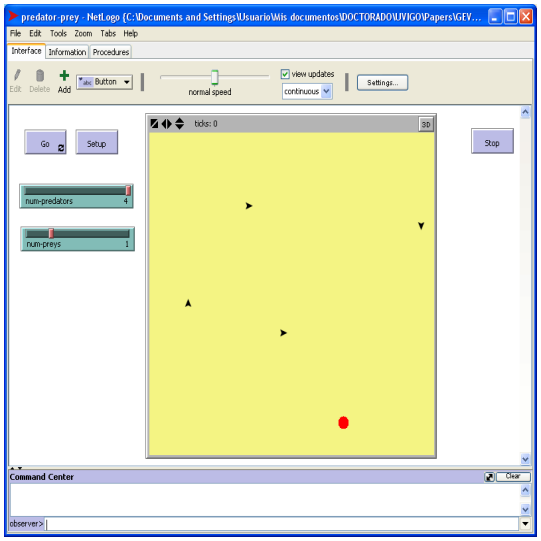


Figure 2: Netlogo Scenario where the Chase is Performed. Predators are Depicted as Black Arrows and the Prey as a Red Circle.

Our predators' program should evolve to find the best function to catch the prey. To do this, we use GE, and we generate the function (from Eq. 1) with GEVA and pass it to Netlogo. Then, Netlogo simulates a run of the MAS where the predators try to catch the prey with the given evolutive programs.

In this experiment, the parameters used in GEVA are a population size of 100 programs, 100 generations to run and a mutation probability equal to 0.01. In Netlogo, we run every simulation ten times, and each run ends when a predator catches the prey, or when the maximum time for catching the prey has finished. After these runs, Netlogo returns to GEVA an average fitness value which indicates how well the generated code has worked. GEVA stores this value with the generated code, and provides a new program (generated from Eq. 1), repeating these steps until the evolutionary process finishes.

To measure the fitness value in Netlogo, we give a positive reward every time the predator gets closer to the prey in the last move it has performed, or we do not give a reward if the predator gets away (we do this valuation before the prey moves). We repeat this process until the end of the run, where we have two possible outcomes: if a predator has caught the prey, then we divide the reward obtained by the total distance it has traveled. If not, we assign the smallest fitness value, and we invert this values (since GEVA finds minimums). Finally, Netlogo passes this value to GEVA, which stores it, generates another function and repeats the whole process again. The algorithm stops when GEVA does not generate more functions, i.e., the evolutionary process stops, and gives as a result the agent program that has the best performance.

$$\begin{aligned} &< opbroad >< op >< op >< op >< op > \\ &< op >= ifPreyNorth\{< dir >\}; \\ &ifPreySouth\{< dir >\}; \\ &ifPreyEast\{< dir >\}; \\ &ifPreyWest\{< dir >\}; \\ &< dir >= \ \{ \ N, \ S, \ E, \ W \ \} \\ &< opbroad >= \{broadcast(); notBroadcast(); \} \end{aligned} \quad (1)$$

After executing the chase in our EPMAS framework, we obtained that the program with best performance is the one shown in Eq. 2. It is the reasonable result, since first predators broadcast the prey position to their mates (if they see the prey), and then move according to the prey's position. We can see the evolution of the fitness in Fig. 3.

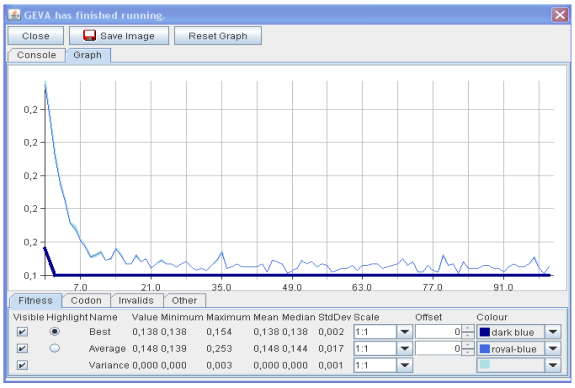


Figure 3: Representation of the Fitness Evolution in GEVA for the Predator-prey Model.

$$\begin{aligned} &broadcast(); \\ &ifPreyNorth\{N\}; \\ &ifPreySouth\{S\}; \\ &ifPreyEast\{E\}; \\ &ifPreyWest\{W\}; \end{aligned} \quad (2)$$

Game Theory Scenario

Game Theory (Binmore, 1994) is a branch of applied mathematics that helps to understand the strategies that selfish individuals may follow when competing or collaborating in games and real scenarios (Osborne, 2003).

The concept of cooperation evolution has been successfully studied using theoretical frameworks like the Prisoner's Dilemma (PD) (Axelrod, 1984), which is one of the most well-known strategy games. It models a situation in which two entities have to decide whether cooperate or defect, but without knowing what the other is going to do.

Nowadays, the PD game has been applied to a huge variety of disciplines: economy, biology, artificial intelligence, social sciences, e-commerce, etc. Table 2 shows the general PD form matrix, which represents the rewards an entity obtains depending on its action and the opponent's one. In this matrix, T means the Temptation to defect, R is the Reward for mutual cooperation, P the Punishment for mutual defection and S the Sucker's payoff. To be defined as a PD, the game must accomplish that:

$$\begin{aligned} T > R > P > S \\ 2R > T + S \end{aligned} \tag{3}$$

Table 2: General Prisoner's Dilemma Matrix.

	Player B Cooperates	Player B Defects
Player A Cooperates	R, R	S, T
Player A Defects	T, S	P, P

There is a variant of PD, which is the Iterated Prisoner's Dilemma (IPD) (Axelrod, 1984), in which the game is played repeatedly. In it, the players can punish their opponents for previous non-cooperative behavior, remembering their opponent previous action and adapting their strategy. Game Theory shows that the optimal action if both players know that they are going to play exactly N times is to defect (it is the Nash equilibrium of the game) (Binmore, 1994). But when the players play an indefinite or random number of times, cooperation can emerge as a game equilibrium. The IPD game models transactions between two persons requiring trust and cooperative behavior, and that is why this game has fascinated researchers over the years.

Axelrod organized a tournament in the eighties, where famous game theorists sent their strategies to play the IPD. The winner was Tit-For-Tat (TFT), a simple strategy where the player cooperates on the first move, and in the rest of iterations it reciprocates what the other player did on the previous move. TFT is considered to be the most robust basic strategy. Although, for a certain range of parameters, and in presence of noise, it was found that a strategy, named Pavlov, that beats all the other strategies by giving preferential treatment to co-players which resemble Pavlov (Nowak and Sigmund, 1993).

In our example, we want to emulate Axelrod's tournament by means of evolutionary computation. To do that, we 'organize' a tournament, where every player individually plays against all the rest. Within the set of players, some of them play fixed strategies, and others play as evolutionary agents.

The values we use for the payoff matrix are the classical ones to play the Prisoner's Dilemma, i.e., R=3, S=0, P=1 and T=5. The idea here is to perform experiments in a controlled environment, to see if GE provides 'good' solutions for several well-known configurations. For that we have defined some tournaments with concrete players. The main GEVA parameters we have used are: a population size of 300 programs, 300 generations, and a mutation probability of 0.05. The players play 1000 rounds against every opponent. The strategies considered for the fixed players are:

- **all-D**: the player always defects.
- **all-C**: the player always cooperates.
- **TFT**: the player uses Tit-For-Tat that cooperates at the first iteration and then reciprocates what the opponent did on the previous move.
- **Pavlov (P)**: the player cooperates at the first iteration, and whenever the player and co-player did the same thing at the previous iteration; otherwise, Pavlov defects.
- **Random (R)**: the player cooperates with probability 0.5.
- **ITFT**: the player uses inverse TFT.

Next, we present the results in tables, where *Initial conditions* refers to the initial action taken for the first iteration, *Type of players* are the strategies that the fixed players play, and finally *Results* describe the agent program (denoted as EA, Evolutionary Agent) obtained with the best performance for the tournament configuration.

In Table 3, we can observe that the best behavior against strategies that do not take into account the opponent's last action is to defect, since we obtain maximum gain without being punished.

In Table 4, we present how the evolutionary agent (EA) behaves against more efficient strategies, in fact, against the two strategies with best behavior in Axelrod's tournament.

In the first tournament, we see that TFT is the best strategy if we have players playing all-D, P and TFT, which is normal since the number of defectors is low, the EA defects against them, and TFT works well if playing against TFT and P.

Table 3: Three Tournaments: Players all-D; all-D and all-C and all-D, all-C and rand.

	Tourn. 1	Tourn. 2	Tourn. 3
Initial conditions	D	D, C	D, C, R
Type of players	20 all-D	10 all-D, 10 all-C	7 all-D, 7 all-C, 7 rand
Results	all-D	all-D	all-D

Table 4: Three Tournaments: Players all-D, P, TFT; P, TFT, all-D, all-C, ITFT and P, TFT.

	Tourn. 1	Tourn. 2	Tourn. 3
Initial conditions	D, C, C	C, C, D, C, D	C, C
Type of players	5 all-D, 10 P, 10 TFT	5 P, 5 TFT, 5 all-D, 5 all-C, 5 ITFT	10 P, 10 TFT
Results	TFT	all-D	P, TFT, all-C

In the second tournament the EA plays against all the strategies (except for random). The result is that the best behavior is to always defect. This makes sense since as we have five all-D, to obtain the maximum gain against them the player should defect. Besides, against ITFT, if the player defects the opponent will cooperate the following round, thus defector obtains maximum gain. Against Pavlov, the EA gets the minimum gain in one round, and the maximum one in the other, since initially the Pavlov player cooperates, while the evolutionary agent defects, thus in the following round they both defect, and in the next Pavlov will cooperate again. With TFT, the player gets the minimum gain every time, but still is more than zero.

Finally, in the third tournament of Table 4, the EA plays against the two more efficient strategies, P and TFT, obtaining that the best to do in this case is to use one of those strategies or to be an all-C, i.e., all of them are equivalent.

Table 5 shows the results when two evolutionary agents play a tournament, with other 25 players that use several fixed strategies. As stated before, GEVA does not consider multi-threading, thus the evolutionary players have to communicate via system files, because each of them is an instance of GEVA, and this takes an important amount of time. We perform a tournament in a Pentium(R) Dual Core CPU, E5200 @ 2.50GHz, 3.50GB RAM, lasting two hours for finishing. In Table 5 we observe that we obtain all-D for both evolutionary players, which is a coherent result comparing it with the second tournament of Table 4.

Table 5: Two Evolutionary Player Playing against P, TFT, all-D, all-C and ITFT.

	Tournament
Initial conditions	C, C, D, C, D
Type of players	5 P, 5 TFT, 5 all-D, 5 all-C, 5 ITFT
Results	all-D, all-D

CONCLUSIONS AND FUTURE WORK

In this paper we have presented the use of Evolutionary Programming to evolve Multi-agent problems by means of Grammatical Evolution. The main contribution of the paper is an evolutionary framework to allow the emergence of Multi-agent Systems adapted to the particular conditions and the scenario to model. We present two case studies with classical MAS scenarios to show that the combination of GEVA and Netlogo is a good option to evolve Multi-agent Systems from scratch by combining system generation and simulation.

While some approaches have considering the use of GP in MAS, to the best of our knowledge, this is the first framework combining GE with Multi-agent Systems for solving general scenarios. Among the conclusions obtained we have found that GE can be a good candidate to automatically solve Multi-agent problems. Nevertheless, due to the novelty of this approach, we have found the limitations of the present version of GEVA and Netlogo to simulate complex MAS scenarios. This happens due to lack of support for multi-threading operations, and therefore we had to communicate the different GEVA instances, and the Netlogo simulator, by means of operating system files, which delays a lot the whole process.

As future work, we plan to find a sound and complete formal description of our MAS framework, to create our own evolutionary simulation environment for avoiding the problems found, and finally to validate such new formal framework with more complex examples from different disciplines.

REFERENCES

Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books.

Benda, M., Jagannathan, V., and Dodhiawala, R. (1986). On optimal cooperation of knowledge sources - an empirical investigation. Technical Report BCS–G2010–28, Boeing Advanced Technology Center, Boeing Computing Services, Seattle, Washington.

Binmore, K. (1994). Game theory and the social contract volume i: Playing fair. *The MIT Press: Cambridge, MA*.

Brabazon, A. and Oneill, M. (2004). Evolving technical trading rules for spot foreign-exchange markets using grammatical

evolution. *Computational Management Science*, 1(3-4):311–327.

Calderoni, S. and Marcenac, P. (1998). Genetic progamming for automatic design of self-adaptive robots. In *EuroGP '98: Proceedings of the First European Workshop on Genetic Programming*, pages 163–177, London, UK. Springer-Verlag.

Fogel, L. J., Owens, A. J., and Walsh, M. J. (1966). *Artificial Intelligence through Simulated Evolution*. John Wiley, New York, USA.

Garshol, M. (2003). *BNF and EBNF: What are they and how do they work?*

Haynes, T. and Sen, S. (1996). Evolving Behavioral Strategies in Predators and Prey. In *Adaptation and Learning in Multiagent Systems*, pages 113–126.

Korf, R. (1992). A simple solution to pursuit games. In *Proceedings of the 11th International Workshop on Distributed Artificial Intelligence. Glen Arbor, MI.*

Koza, J. R. (1994). *Genetic programming II: automatic discovery of reusable programs*. MIT Press, Cambridge, MA, USA.

Nowak, M. and Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature*, 364(6432):56–58.

O'Neill, M., Hemberg, E., Gilligan, C., Bartley, E., McDermott, J., and Brabazon, A. (2008). Geva: grammatical evolution in java. *SIGEVOlution*, 3(2):17–22.

O'Neill, M. and Ryan, C. (2003). *Grammatical Evolution: Evolutionary Automatic Programming in an Arbitrary Language*. Kluwer Academic Publishers, Norwell, MA, USA.

Osborne, M. J. (2003). *An Introduction to Game Theory*. Oxford University Press, USA.

Russell, S. J. and Norvig, P. (2003). *Articial Intelligence: A Modern Approach, 2nd Ed.* Prentice Hall, Englewoo.

Ryan, C. and O'Neill, M. (1998). Grammatical evolution: A steady state approach. In *In Late Breaking Papers, Genetic Programming*, pages 180–185.

Ryan, C., O'Neill, M., and Collins, J. (1998). Grammatical evolution: Solving trigonometric identities. In *Proceedings of Mendel '98: 4th International Conference on Genetic Algorithms, Optimization Problems, Fuzzy Logic, Neural Networks and Rough Sets*, pages 111–119.

Tan, M. (1997). Multi-agent reinforcement learning: Independent vs. cooperative learning. In Huhns, M. N. and Singh, M. P., editors, *Readings in Agents*, pages 487–494. Morgan Kaufmann, San Francisco, CA, USA.

Wilensky, U. (2007). Netlogo: Center for connected learning and computer-based modeling.

Wooldridge, M. (2002). *Introduction to MultiAgent Systems*. John Wiley & Sons.

Wooldridge, M. and Jennings, N. R. (1995). Intelligent agents: Theory and practice.

Zelinka, I. and Oplatkova, Z. (2003). Analytic programming comparative study. In *CIRAS03, The second International Conference on Computational Intelligence, Robotics, and Autonomous Systems*.

AUTHOR BIOGRAPHIES

ANA PELETEIRO received the M.Sc. degree in Telecommunication Engineering in 2009 (Honor mention) at University of Vigo. She is currently a PhD student in the Department of Telematic Engineering at the same university. Her research interests include intelligent agents and multi-agent systems, self-organization, evolutionary algorithms and game theory. Her email is `apeleteiro@det.uvigo.es` and her personal webpage `http://www-gti.det.uvigo.es/~apeleteiro`

JUAN C. BURGUILLO received the M.Sc. degree in Telecommunication Engineering in 1995, and the Ph.D. degree in Telematics (cum laude) in 2001; both at the University of Vigo, Spain. He is currently an associate professor at the Department of Telematic Engineering in the University of Vigo. His research interests include intelligent agents and multi-agent systems, evolutionary algorithms, game theory and telematic services. His email is `J.C.Burguillo@det.uvigo.es` and his personal webpage `http://www.det.uvigo.es/~jrial`

ZUZANA OPLATKOVA was born in Czech Republic, and went to the Tomas Bata University in Zlin, where she studied technical cybernetics and obtained her MSc. degree in 2003 and Ph.D. degree in 2008. She is a lecturer (Artificial Intelligence) at the same university. Her research interests are: evolutionary computing, neural networks, evolutionary programming, metaevolution. Her e-mail address is: `oplatkova@fai.utb.cz`

IVAN ZELINKA was born in Czech Republic, and went to the Technical University of Brno, where he studied technical cybernetics and obtained his degree in 1995. He obtained his Ph.D. degree in technical cybernetics in 2001 at Tomas Bata University in Zlin. He is now professor (artificial intelligence, theory of information) at the Department of Informatics and Artificial Intelligence. His e-mail address is: `zelinka@fai.utb.cz` and his Webpage can be found at `http://www.ivanzelinka.eu`

ANALYSIS AND CLASSIFICATION OF TASK KNOWLEDGE PATTERNS

Cheah WaiShiang, Leon Sterling
Department of Computer Science and Software Engineering
University of Melbourne, Australia
E-mail: w.cheah@pgrad.unimelb.edu.au, leonss@unimelb.edu.au

KEYWORDS
Classification scheme, attribute-based analysis, patterns

ABSTRACT

Patterns have recorded the experience of engineering software systems. Various patterns have been introduced and described in several domains. A particular class of patterns, *task knowledge patterns* that are important for agent oriented software development, is investigated in this paper. It has been reported that descriptions of task knowledge given in the pattern literature is not clearly structured, and some of the useful description is left implicit to the developer. This paper reports our investigation on task knowledge that has been shared across a wider spectrum of articles to showcase such scenario. We present a new way to analyse and classify task knowledge patterns. Furthermore, we demonstrate how the experience of different dvelopment groups is described very differently based on the classification result. The investigation shows the need to present task descriptions in a structured manner. The analysis and classification techniques demonstrated are applicable for agent-based simulation.

INTRODUCTION

Over the past several years, various agent patterns have been introduced, for example as listed in (Oluyomi, Karunasekera et al. 2007) and (WaiShiang 2010). Patterns record experience of engineering software system. While various patterns have been introduced and patterns have been described in several domains by Oluyomi, a particular pattern, *task knowledge patterns*, that is important but paid less attention to by Oluyomi, is investigated in this paper.

In (WaiShiang 2010), we report on task knowledge patterns known as CommonKADS template knowledge models (Schreiber 2000) that describe the problem solving method or task knowledge within task template description. Task knowledge is being shared by describing the experience of dealing with particular task types across a wider spectrum of articles (De Wolf and Holvoet 2006). For example, various articles have been presented in dealing with mediating user requests and services through engineering a brokering system.

It has been reported that the description for the task knowledge given in the articles is not clearly structured, has indirect descriptions and some of the useful description seems implicit to the developer (De Wolf and Holvoet 2006). The investigation to showcase such scenario is presented in this paper. The investigation raises the need to present task descriptions that are spread across various articles in a structured manner. We present a new way to analyse and classify task knowledge patterns. Furthermore, we demonstrate how the experience of different dvelopment groups is described very differently based on the classification result. The investigation raises the need to present task descriptions that are spread across various articles in a structured manner. The analysis and classification techniques are applicable for agent-based simulation.

This paper consists of five sections. Section 2 introduces a Two-ways classification scheme that will be used to analyse and classify task knowledge patterns. An analysis method, attribute-based analysis, is presented in Section 3. An example of classification, brokering systems, is presented in Section 4. Altogether, we present the result of classifying 58 articles that describe the experience of engineering a brokering. The sample collection of 58 articles is first described, followed by the result of the classification and observations. The paper's conclusions are given in Section 5.

TWO-WAYS CLASSIFICATION SCHEME

Viewpoint abstraction layers	Viewpoint aspects		
Conceptual independent modelling level	Information	Interaction	Behaviour
Platform independent design & modelling level	Definitional	Structural	Interaction
Platform specific design & modelling level	Structural	Interaction	Strategy

Figure 1: Two-ways classification scheme

We can describe the task knowledge that has been shared using a Two-ways classification scheme adapted from (Sterling and Taveter, 2009) as shown in Figure 1. The Two-ways classification scheme is used to classify knowledge according to viewpoint abstraction layers and viewpoint aspects. The viewpoint abstraction layers presents the stages of developing sociotechnical systems. The viewpoint aspects presents aspects for each stage of the development.

The viewpoint abstraction layers constitute three levels. They are conceptual domain modelling (CIM) level, platform independent design and computation design (PIM) level and the platform specific design and

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

implementation (PSM) level. The conceptual domain modelling level records the experience in analyzing the system. The description given is presented at a high level of abstraction. For example, the listing of players and tasks related to them within the insurance brokerage; listing of information that is required to be collected by the broker and such description does not describe the detail elements like association among the information, attributes; listing of brokering activities. The PIM level records experience to provide a detailed design of the system. The PIM shows part of the system design specification that does not change from one platform to the other. The design activities will range from conceptualization, arrangement of external functions that rely upon, arrangement of inferences steps or internal structure and method control. The PSM level will complement the PIM level with the lowest detail design that specifies how the system is to be implemented in a specific platform, tool or programming language. For example, the functionality of the mediator is designed at code level, the service description for matchmaking is designed through MONET description (Ludwig, Rana et al. 2006).

The viewpoint aspects represent the software development task that is split into three categories. They are information category, interaction category and behaviour category. The information category reflects the terms used. The interaction category reflects who are involved in solving the problem. It is related to the modelling of transfer function with the external world when solving a problem with CommonKADS. The behaviour category reflects the inference steps or actions, and the sequences of actions together with the control behaviour of the actions. For example, an online housing system will obtain information from the database system. The system will use certain resources like user identification, price list and so on for task execution. It performs tasks like assessment of user application, and monitoring housing applications. Typically it will rely on an external application like a banking system for validating user financial information.

To classify and analyse the shared experience through the Two-ways classification scheme, attribute-based analysis is introduced. Attribute-based analysis is introduced to characterize and categorize the shared experience through the Two-ways classification scheme.

ATTRIBUTE-BASED ANALYSIS

Attribute-based analysis is introduced to characterize and categorize the patterns through the Two-ways classification scheme (Oluyomi, Karunasekera et al. 2007). In other words, the attribute-based analysis is used for classifying agent patterns according to the dimensions of the Two-Ways classification scheme. The attribute-based analysis will determine into which dimension (e.g. level and category of Two-ways classification scheme) a pattern belongs to based on the identification of pattern level attributes and identification of pattern category attributes. For example, a pattern is classified at a conceptual independent modelling level and behaviour category, if the pattern has been described at a high level of abstraction like presenting a collection of goals with the notion of the goal reflecting the behaviour aspect within the Two-Ways classification scheme.

The attributes represent the features or common knowledge elements that fall within a particular level and category in the Two-Ways classification scheme. They represent the agent concepts that people use during the engineering of software system (e.g. agent oriented software system or non-agent oriented software system).

Figure 2 shows the attributes that we identified for Two-ways classification scheme. Each dimensions (e.g. level and category of Two-ways classification scheme) cover sets of attributes that will be used to classify and analyse the agent patterns through attribute-based analysis. The attributes are identified by studying the conceptual space for social technical system, various agent oriented methodologies and agent oriented software development projects (Sterling and Taveter 2008). Due to space limit, we only present the informal definition for the attributes at viewpoint abstraction layers for the following description. This will be followed by the description of attribute analysis tools and process.

Two-ways classification scheme level attributes

This section presents the informal definition for the attributes at viewpoint abstraction layers.

Conceptual Domain Modelling level.

Identify Role/Multiple roles Identification of roles have been played for solving a particular problem and opportunity within a real life organization. This attribute looks at the activity to analyse the roles that have been played and the responsibilities of the role in solving a problem.

Assigning Goal Assigning goals and sub-goals appears in the analysis phase of agent oriented software development. In addition, the attribute looks at the activity to identify the goal(s) achieved by a particular role and resource usage.

Identify domain entities Identifying the resources or sources that are required for the problem and opportunity given is the activity that is described by this attribute. Activities like understanding the scope and the aim of the domain entities and identifying the glossary are the focus of this attribute.

Viewpoint Abstraction layer	***Viewpoint aspect***		
	Information	***Interaction***	***Behaviour***
Conceptual domain modelling • Identify roles • Assigning goals • Identify domain entities • Organization of roles	• Domain entities	• Roles • Organization structure • Social policies/authority power • Responsibility	• Goal • Quality goal • Goal dependency • Roles/actors • Resource
Platform-independent computational design • Ontology/ domain conceptualization • Arrangement of agents • Agents message exchange • Agent or module internal activity	• Concepts • Slot • Relation	• Agents/ multiple agents • Communication state/ social order/communication strategy • Message exchanged	• Agent type • Information type • Reasoning/strategy • Proactive • Reactive
Platform-specific design and implementation • Plan specification • Interaction specification • Agent instantiation, aggregation, inheritance • Information specification	• Concrete object (i.e. Belief) • Syntax of Knowledge representation in language specific	• Message scheme • Communication support • Concrete organization structure	• Behaviour construct (i.e. Plan, event) • Concrete agents/ agent instances • Perception • Resource Utilization

Figures 2: Attributes of Two-ways classification scheme

Organization of role Organization of role describes the arrangement of role in dealing with a problem and opportunity in an organization. This attribute looks at describing an organization structure such as peer to peer, flat, hierarchical, etc.

Platform Independent Computational Design level.
Ontology/ domain conceptualization The attribute captures activities for modelling concepts of roles, class, property and concept hierarchies.
Arrangement of agents Agent acquisition is the design activity that is captured in this attribute. The agent in this context can refer to artifact like humans or people, functional roles, software components, modules, information systems or software agents.
Agent interaction Activity in determining the flow of information in exchanging the information among external software components or designing the interaction protocol is captured in this attribute.
Agent internal activity The agent internal activity captures the designing of agent internal state(s) corresponding to decision making, responding to changes that occur and goal achievement. Some examples of designing an agent internal structure are determining post condition, pre-condition or execution control (e.g. looping, recursive) of agent behaviour, assigning actions and inference steps for agent reasoning.

Platform Specific Design and Implementation level.
Behaviour specification Behaviour specification describes the lowest level detailed design of the event and control behaviour of an event that is used in various platforms like Jason, JACK, .Net and so on.
Interaction specification Interaction specification describes the lowest level message exchanged among agents. In addition, the designing of a message transport protocol as well as a communication support like bandwidth, communication hardware determination will be captured in this attribute.
Agent instantiation, aggregation, inheritance An agent in this level involves the designing of a low level detail of concrete agent like designing the agent through an object oriented class diagram, with OO class representation as agent type, methods, attributes of the class and inheritance, association among the classes as agent behaviour.
Information specification In ontology engineering (Uschold and Gruninger 1996), this is known as the process of ontology coding.

Attribute-based analysis: Tool and processes

Tool and processes are introduced for attribute-based analysis. The attribute-based analysis tools are analysis tables used for identification of pattern level attributes to determine the level that the pattern belongs to and identification of pattern category attributes to determine the category that the pattern belongs to.
The tool is a tablet form that is populated with attributes from within the Two-Ways classification scheme. Altogether, four sets of attribute-analysis tables are introduced as shown in Table 1 to Table 4. Each of the analysis tables outlines three sections. They are the *knowledge source* that records the input for analysis purpose; the *viewpoint abstraction layer or dimension* and the *attributes* underneath the dimension and the *analysis outcome* of the analysis.
The knowledge source consists of a pattern description used for evaluation. The dimension and attributes cater the elements used for identification of pattern level attributes and pattern category attributes.

Table 1: Identification of viewpoint abstraction layer attributes for patterns

Knowledge source	Information	Interaction				Behaviour					
	Domain entities	Roles	Responsibility	Organization Structure	Social policy	Goal	Quality goal	Role/ actors	Goal dependency	Resource	Analysis conclusion

Knowledge source	Conceptual domain modelling				Platform independent modelling		Platform specific modelling						
	Identify role	Assigning goal(s)	Identify domain entities	Organi-zation of role	Ontology / Conceptu alization	Arrangement of agents	Agent internal activity	Agent message exchange	Plan/ event spec.	Inter-action spec.	Agent In-stantiation	Info spec	Viewpoint abstraction layer

Table 2: Identification of category attributes for conceptual domain modelling level

Know. source	Information			Interaction			Behaviour					Analysis Conclusion
	Concept	Slot	Relation	Agents/ multiple agents	Social order	Message	Agent	Resource/ Service	Reasoning	Proactive	Reactive	

Table 3: Identification of category attributes for platform independent modelling level

Know. source	Information		Interaction			Behaviour				
	Concrete object	Syntax of KR	Message content	Communication Support	Concrete organization structure	Behaviour construct	Concrete Agent	Perception	Resource utilization	Analysis Conclusion

Table 4: Identification of category attributes for platform dependent and implementation modelling level

Table 1 is designed to classify and analyse the patterns according to a particular vertical dimension. It has been populated with attributes that can be identified within the vertical dimension. For example, analysis table of Table 1 is outlined with levels of viewpoint abstraction layers and the attributes in each of the levels (e.g. *identify roles, assigning roles, identify domain entities, organization of role* at conceptual independent modelling level). Other analysis tables are designed to classify and analyse the patterns according to a particular viewpoint aspect. Table 2 is the analysis table that is used to identify which viewpoint aspect the pattern belongs at the conceptual domain modelling level. Table 3 is the analysis table that is used to identify which viewpoint aspect the pattern belongs at the platform independent design level. Table 4 is the analysis table that is used to identify which viewpoint aspect the pattern belongs to at the platform specific design and implementation level.

To analyse the pattern using attribute-based analysis, the attribute-based analysis processes are described as follows. We first examine the shared experience towards discovery of attributes. We review and observe the shared experience in engineering the system, to review what might be shareable from the description given, *step1*. From the observation, we proceed to viewpoint abstraction layer attributes analysis and viewpoint aspect attributes analysis. This involves activity in placing the patterns on the attribute analysis table (e.g. Table 1 to Table 4) and to indicate which of the viewpoint abstraction layer attributes or viewpoint aspect attributes listed in the analysis table are presented in the pattern, based on the examination of the pattern. Finally, we conclude the analysis result.

Concluding the analysis result is based on the frequent occurrence of attributes within dimensions of Two-ways classification scheme. In other words, we conclude that a pattern falls in a particular dimension if the number of occurrence attributes for the shared experience in one dimension is higher than the number of occurrence of attributes in other dimension.

In the previous description, we presented the dimensions of Two-ways classification scheme and the informal description on those dimensions. This is followed by our description on attributes for attribute-based analysis as well as informal definition on the attributes at viewpoint abstraction layers. In addition, we presented the tool and processes for attribute-based analysis. To demonstrate the feasibility of the classification scheme, an example is presented in the following sections. We conduct an analysis and classification of 58 ‘brokering patterns’ and present our observation from the result of the classification.

ATTRIBUTE-BASED ANALYSIS FOR ‘BROKERING PATTERNS’

In this section, we demonstrate the analysis and classification of ‘brokering pattern’ through Two-Ways classification scheme using attribute-based analysis. In other words, the pattern description is analysed using the attribute analysis table of Table 1 to Table 4. A sample to demonstrate the analysis and classification of brokering pattern through the Two-ways classification scheme using attribute-based analysis is presented first. This is followed by the result of the analysis of 58 ‘brokering patterns’. At the end of this section, we present observations on the result of the classification.

In the example, we present the source of the 'brokering pattern', overview of the pattern, justification of attribute-based analysis and result of the classification. In addition, the example presents the experience that has been shared in the pattern documentation.

Source: (Lau and Goh 2002)

Overview: The system presents the brokering of the e-marketplace through the broker model. The broker model consists of formulization of the brokering problem by treating it as a set packing problem and solved through a proposed iterative greedy approximation algorithm.

Solution:

```
FUNCTION LOCAL_SEARCH(set S)
    TempSBest ← S;
    FOR EACH COLUMN j in Candidates DO
        TempS ← RemoveConflict(S, j);
        TempS ← TempS + {j};
        TempS ← GREEDY(TempS);
    IF (Value(TempS) ≥ Value(TempSBest)) THEN
            TempSBest ← TempS;
    ENDIF
    ENDFOR
    RETURN TempSBest;
```

Figure 3: Functions that present the algorithm in handling the brokering problem as described.

Viewpoint abstraction layer attributes analysis: The pattern presented in Figure 3 was analyzed using the level attribute table (Table 1) in order to determine the level that the pattern belongs to. Figure 3 records the detail description to identify best fit for a particular agent. From the analysis, we can identify the attribute for this pattern include 'agent internal activity'. Although it does not indicate as a software agent, the function component that presented a detail design on part of the system internal activity can relate to the attribute of agent. As a result, we conclude that the pattern belongs to the PIM-level.

Viewpoint aspect attributes analysis: The pattern presented in Figure 3 was analyzed using the category attribute table (Table 2, Table 3 or Table 4) in order to determine the viewpoint aspect that the pattern belongs to. At the PIM-level, the set of attributes for the pattern in Figure ks17c include agent, resource (e.g. input on the function), reasoning (e.g. influence step, control structure of for loop, if condition) and proactive (internal influence, if value). Hence, it belongs to the behaviour category.

From the analysis and classify result, the knowledge and experience that have been shared in this 'brokering pattern' belongs to the PIM level behaviour category as it contain a detail description of broker agent activity

Table 5: Result of the classification of 'brokering pattern'

Viewpoint abstraction layer	Viewpoint aspect		
	Information	Interaction	Behaviour
CIM			
PIM			Figure 3
PSM			

Continuing the above practice, we conduct a further study to analyse and classify 58 'brokering patterns'. The reader can request the sample collection of the 'brokering patterns' by contacting the authors. The document contain the references for the 'brokering patterns'.

Result of the analysis of 'brokering patterns' and observation

Table 6 shows the comprehensive view of the 'brokering patterns'. From the table, each figure represents the identification of a brokering pattern that has been shared within a particular article or pattern documentation. Each experience that has been shared will label as 'X', in which each figure contains the pattern description that has been structured in the article accordingly. In addition, each figure will capture different representation (e.g. algorithm, architecture etc.) that has been used by pattern designer to describe the shared knowledge. For example, from the Table 6, we compiled the shared experience with figures 1a and 1b. Both figures shows the experience in engineering the brokering system in different sections of the article accordingly. In addition, both figures show different representations that have been used by pattern designer in sharing the development experience of brokering system.

From the result of the analysis, it presents what might be shareable from the 'brokering patterns'. On the other hand, it is interesting to observe that the knowledge and experience in engineering the brokering system that are shared across the various articles do not present in clear structure, indirect description and some of the useful description is seem implicit to the developer. Our observation can further describe with the following description.

- Within a particular 'brokering pattern', the pattern designer will share the analysis of the brokering system, followed by the design view and continue to share the analysis view and so on. To link from one representation to the others and switch the description from one view to the others within the pattern description has created some confusion and difficulty. The patterns that happen in these kinds are knowledge source of ks2, ks3, ks4, ks5, ks8, ks9, ks11, ks12, ks14, ks15, ks16, ks17, ks22, ks27, ks30, ks33, ks37, ks38, ks39, ks40, ks42, ks43, ks45, ks47, ks51, ks52, ks53 and ks57.

Table 6: A Comprehensive view of 58 ‘brokering patterns’ in Two-ways classification scheme

Vertical dimension/	**Horizontal dimension/ Viewpoint aspect**																							
	Information								***Interaction***								***Behaviour***							
Conceptual domain modelling			3c	4c						2a						8a								
					13d		15c										9b		11e	12b	13a	14b		16a
											11d	20a		14a		16c			19a					
														30a			25a		27a					
										34a				38a		40c								
	41a																							
																						54a		
																	57a							
Platform-independent computational design						6b	7a		1a		3a		5b			8c	1b	2b	3c	4a, b	5a, c	6a	7b	8a, b
	9c				13d					10a	11b				15a		9a, d	10b ,c	11a	12a	13c	14b ,c,d	15b	16b
							23a		17a				21b	22a	23c		17b ,c	18a	19b	20b	21a		23b	24a
	25c			28c			31b			26a		28b				32a	25b	26b	27c	28a	29a	30b ,d	31a	32b
	33b	34c							33a, b					38c	39a		33c ,d	34b		36a	37a ,b	38b ,d	39c	40b
										41b				46b		48a		42b ,c	43b	44a	45a ,b,c	46a	47a ,b	
						54b			49a					54e			49b		51a ,b	52a ,b	53a	54 d		56a
	57b																57a ,c	58a						
Platform-specific design and implementation																8d								
											11c														
														22b, c						20c				
								32b						30e					27b ,d,e			30c	31c	
														38e		40a			35a				39b	
	43 a															48b		42a						
					53b	54c	55a										49c	50a			53b ,c		55b	56 b

- Some knowledge is shared explicitly and some is kept implicitly. This can look to the distribution of the knowledge has been shared across the dimensions of the Two-ways classification scheme as shown in Table 6. For example, some patterns explicitly shared the information that was required for the system; internal structure or behaviour of the system; external component or person that it relies upon for task accomplishment. As a result, we face difficulty in understanding the implicit knowledge that has been described in the patterns.
- A similar aspect of the system is shared in different representations in a pattern. For example, pattern of knowledge source, ks10 has described the behaviour of the system in the architecture view and pseudo-code. Other patterns are ks8, ks10, ks14, ks17, ks22, ks27, ks33, ks37, ks38, ks42, ks45, ks47, ks51, ks52, ks53 and ks54. This will lead to some confusion and extra effort is needed to learn the representations that are used for describing the system.

DISCUSSION AND RELATED WORKS

Various pattern classifications have been introduced for patterns. Gamma et al introduce several pattern catalogs for design patterns (Gamma, Helm et al. 1995). A comprehensive pattern classification scheme for agent patterns was introduced by Oluyomi in the University of Melbourne AgentLab (Oluyomi, Karunasekera et al. 2007). We extended Oluyomi's pattern classification to introduce a more comprehensive and simpler classification scheme for patterns. The pattern classification allows the ease of accessibility of patterns by user and pattern designer (Oluyomi, Karunasekera et al. 2007). From the proposed classification scheme, we demonstrate how the software practitioners are able to recognize what state of software engineering the patterns relate to and are able to relate the patterns from one dimension to the others. In addition, the result of the classification enable us to justify that the experience of different development groups for the case of brokering mechanism is described very differently.

We believe that it is hard to evaluate each of the classification scheme. The best we can offer is to demonstrate how the classification scheme is used to analyse and classify 58 brokering patterns.

CONCLUSIONS

An investigation of task knowledge patterns that have been shared across various articles is conducted in this paper. It is interesting to observe that the experience in engineering the brokering system given in the 'brokering patterns' documentations is not presented with clear structure, has indirect description and some of the useful description seems implicit to the developer.

We presented a comprehensive view of the 'brokering patterns' based on the analysis and classification of the 'brokering patterns' through the Two-ways classification scheme using attribute-based analysis. We demonstrated how the Two-ways classification scheme can be adopted to analyse and classify the 'brokering patterns' that ranged from an agent oriented software system to a non-agent related system. Hence, it can turn into a simulation platform to visualize the knowledge and experience that have been shared in a wider spectrum of articles. Continuing studies are being conducted to demonstrate the Two-ways classification scheme to analyse and classify agent patterns; to demonstrate the usage of the result's classification in guiding the pattern writing (WaiShiang 2010).

REFERENCES

De Wolf, T. and T. Holvoet (2006). A catalogue of decentralised coordination mechanisms for designing self-organising emergent applications. Technical Report, Department of Computer Science, K.U. Leuven.

Gamma, E., R. Helm, et al. (1995). Design patterns: elements of reusable object-oriented software, Addison-Wesley.

Lau, H. C. and Y. G. Goh (2002). An intelligent brokering system to support multiagent web-based 4th-party logistics. Proceedings 14th IEEE International Conference on Tools with Artificial Intelligence.

Oluyomi, A., S. Karunasekera, et al. (2007). "A comprehensive view of agent-oriented patterns." Autonomous Agents and Multi-Agent Systems **15**(3): 337-377.

Schreiber, G. (2000). Knowledge engineering and management: the CommonKADS methodology, MIT press.

Sterling, L. and K. Taveter (2008). The Art of Agent Oriented Modelling, MIT Press, Cambridge.

Uschold, M. and M. Gruninger (1996). "Ontologies: Principles, methods and applications." The Knowledge Engineering Review **11**(2): 93-136.

WaiShiang, C. (2010). Patterns in agent oriented software development. Department of Computer Science and Software Engineering, The University of Melbourne PhD thesis.

AUTHORS' BIOGRAPHIES

Leon Sterling was born in Melbourne, Australia. He completed a B.Sc.(Hons.) at the University of Melbourne, and a Ph.D. in computational group theory at the Australian National University. He has worked at universities in the UK, Israel and the US. In 1995, he returned to the University of Melbourne, and founded the University of Melbourne Intelligent Agent Lab. He moved in 2010 to become Dean of the Faculty of Information and Communication Technologies at Swinburne University of Technology.

Cheah WaiShiang is currently a PhD student at the University of Melbourne, Australia.

A SPATIAL SIMULATION MODEL FOR THE DIFFUSION OF A NOVEL BIOFUEL ON THE AUSTRIAN MARKET

Elmar Kiesling, Markus Günther, Christian Stummer, Rudolf Vetschera, and Lea M. Wakolbinger
University of Vienna
Department of Business Administration
Vienna, 1210, AUSTRIA
Email: christian.stummer@univie.ac.at

KEYWORDS

Agent-Based Modeling, Innovation Diffusion, BTL-Biofuel, Spatial Social Network

ABSTRACT

To ensure long-term security of energy supply and to mitigate climate change, sustainable low-carbon alternatives to fossil fuels are needed. Second generation biomass-to-liquid biofuels provide such an alternative and are widely considered a promising technology that may overcome the environmental and economic problems associated with first generation biofuels. In Austria, a research group at the Vienna University of Technology is developing such a biofuel based on Fischer-Tropsch synthesis. While the remaining technical challenges are expected to be overcome in due time, the market introduction also requires substantial investments. In this work, we introduce an agent-based simulation model that can provide potential investors with forecasts for the biofuel's market diffusion. The model considers initial and repeat purchases, multiple competing products, and the spatial dispersion of both consumers and potential points of sale. Given limited production capacity, the latter feature supports decision makers in choosing (initial) points of sale with respect to rich sources of biomass as well as the geographic concentration of consumers.

INTRODUCTION

Concerns over price and long-term supply of oil as well as the considerable environmental impact of fossil fuel use have led to surging interest in alternatives to petroleum-based transportation fuels. This issue has become particularly pressing because carbone dioxide (CO_2) emissions have to be drastically reduced in order to mitigate climate change (O'Neill et al., 2010). Note that currently the combustion of transportation fuels accounts for more than 20% of net CO_2 emissions in the EU, with strong upward emission trends (European Environment Agency, 2008).

In the long run, hydrogen-powered, plug-in hybrid and electric cars may provide a low-carbon alternative given that production of hydrogen and electricity is decarbonised. In the short run, however, substitutes for fossil fuels have to ensure full compatibility with the existing petrol infrastructure in order to become adopted by a considerable share of consumers. Second generation biofuels gained from biomass-to-liquid (BTL) processes based on the Fischer-Tropsch (FT) synthesis, such as the biofuel "BioFiT" that is currently under development at the Institute of Chemical Engineering at the Vienna University of Technology (Fürnsinn, 2007), constitute a particularly interesting alternative since they are not only fully compatible with the existing infrastructure (thus eliminating technical barriers to their adoption), but they can also use a wide range of inputs such as wood, sawdust, switchgrass, and agricultural residues and by-products. Thereby, they overcome many issues associated with the agricultural production of first generation biofuels, including reservations concerning their overall carbon advantage (Fargione et al., 2008), negative impact on water resources (Logan, 2008) and biodiversity (Inderwildi and King, 2009), unfavorable net energy return (Pimentel and Patzek, 2005), and the increased competition for land, which may exacerbate food security issues in the developing world (Inderwildi and King, 2009). At present, the application case biofuel is produced on a laboratory scale (Hofbauer et al., 2005). Industrial scale up and market introduction are expected within the next three to five years. This paper addresses the question whether or not and how fast the novel BTL-FT biofuel would be adopted by consumers.

So far, innovation diffusion research has investigated various factors that determine the speed and degree with which new products, practices, and ideas propagate through a society (Rogers, 2003). In his seminal contribution, Bass (1969) introduced a mathematical model of the innovation diffusion process on an aggregate level and provided a closed formula describing the effects of (external) mass communication and (internal) word-of-mouth communication on the diffusion process. The Bass model has been extended in several respects later on (cf., e.g., Robinson and Lakhani, 1975; Easingwood et al., 1983; Norton and Bass, 1987; Jain and Rao, 1990; Bass et al., 1994). However, none of these approaches takes the individual characteristics of consumers into account and they thus widely neglect consumers' heterogeneity with regard to preferences and behavior.

By contrast, agent-based models (ABMs) account for the consumer's individual decision-making and model

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

the diffusion as an aggregate process of individual adoption decisions. The main merit of such an approach lies in its capability to capture the complex structures and dynamics of diffusion processes without knowing the exact global interdependencies (Borshchev and Filippov, 2004). Moreover, an agent-based approach makes it possible to account for micro-level drivers of innovation adoption by modeling how consumers' attitudes and behaviors are affected by, for instance, the perception of product characteristics or information exchanged in a social network.

In this paper, we discuss the design and implementation of an ABM for simulating the diffusion of a second generation biofuel on the Austrian market. Our paper's contribution with respect to methodological advances is threefold. First, in contrast to other innovation diffusion models that are typically concerned with initial adoption only, our model explicitly considers both initial and repeat purchases. This expanded focus becomes particularly relevant when modeling the diffusion of frequently purchased products such as fuel.

Second, our model considers multiple competing products and, thus, is not limited to applications in which no preexisting substitutes are available. This innovative feature is essential when simulating the diffusion of a second generation biofuel.

Third, space is modeled explicitly, which is a distinguishing feature compared to many other ABMs of innovation diffusion, including a previous version of our model (Kiesling et al., 2009). An explicit spatial model enables us to model a geographically dispersed social network and to explicitly consider multiple points of sale (POS) and their locations.

From an application point of view, we provide decision makers with insights into the marketability of an innovative biofuel, enable them to assess the effectiveness of various pricing strategies, and provide them with a tool for selecting the points of sale at which the biofuel should be made available.

The remainder of the paper is structured as follows. The next section introduces our agent-based model. Then, we discuss the design of the simulation study and present results of selected simulation experiments. Finally, we summarize key findings and provide suggestions for further research.

AGENT-BASED MODEL

Agent-based modeling has been used widely in innovation diffusion research (Garcia, 2005). Typically, ABMs proposed in the literature (Delre et al., 2007a,b; Schenk et al., 2007; Schwoon, 2006) divide time into periods (i.e., they follow a discrete time approach). Our model, by contrast, conceptually treats time as continuous and is thus based on a discrete event approach. Accordingly, we use continuous interarrival time distributions (e.g., Poisson processes) for events and thus avoid the problem of determining the sequence of events scheduled for the same time period. Furhtermore, the approach allows for more general models without a loss in computational efficiency.

In the following section, we outline the major model elements, describe the spatial model, and discuss our modeling of agent behavior before concluding this section with comments on implementation issues.

Model elements

Products Each product is characterized in terms of multiple attributes such as price, quality, or environmental friendliness. We assume that consumers do not generally know the "true" product attribute values with certainty, but rather estimate product characteristics based on the limited local information that they possess. This information may stem from two separate sources, namely (i) the estimations of other consumers as obtained through word-of-mouth and (ii) personal experience. The degree to which consumers can draw upon the latter source may, however, be limited for certain attributes due to the fact that not all of a product's characteristics can be easily evaluated by using it. For instance, a consumer is unlikely to be able to precisely assess a fuel's combustion properties or environmental impact. In order to incorporate these limitations, we introduce an attribute-specific observability parameter that controls the influence of personal experience on the formation of product attribute estimates.

Points of sale We focus on the diffusion of innovations that belong to the category of frequently purchased goods being distributed by means of physical POS. Each POS is characterized by a location and a product range.

Consumer agents Consumers form the pivotal elements of the model. In contrast to aggregate models of innovation diffusion, ABMs recognize consumers as independent entities with heterogeneous preferences that are capable of learning and adapting their behavior. In our model, each consumer agent is characterized by a number of individual parameters, most importantly partial utility functions for each product attribute and an "innovativeness threshold". The latter parameter is used to incorporate individualized adoption behavior based on the concept of innovativeness (Rogers, 2003): some agents are willing to adopt the innovation once its utility exceeds that of existing alternatives, whereas others will not purchase the new product before they are completely convinced that it offers considerable advantages. For the application case presented in this paper, additional parameters that characterize the agent's mobility behavior are also taken into account.

Social network Social interactions play a crucial role in the spread of innovations across societies (Alkemade and Castaldi, 2005; Deffuant et al., 2005). ABM approaches allow for an explicit micro-level representation

of these interactions. In our model, consumer agents are embedded in a social network that represents communication links between individual members of the social system. The algorithm used for constructing this social network is covered in the following section. An arrival process is attached to each link in the network and used to generate communication events. We applied a Poisson stream in all our experiments to reflect our assumption of exponentially distributed interarrival times for communication events.

Spatial model

In our spatial model, both consumer agents and POS are embedded in geographic space. Consumer agents are distributed according to a region's measured population density; POS are placed according to their actual geographical location.

Once consumer agents have been assigned a geographical location, the next step is to construct links among them to create a network. The algorithm used to generate the synthetic network should reproduce characteristic features of real world social networks, including (i) small diameter, (ii) high clustering, and (iii) scale-freeness.

Small diameter is the property that the largest number of links on the shortest path between any two nodes is small; more specifically, the diameter of small-diameter networks scales logarithmically with the number of nodes. This distinct feature of social networks has been demonstrated in many empirical studies since Milgram's famous experiment first showed that any two persons in a society are separated by an average distance of about six steps (Travers and Milgram, 1969).

Next, social networks tend to be highly clustered, which means that A being linked to B as well as to C implies a strong likelihood that C is also linked to B. Networks that are both small in diameter and highly clustered are called 'small-world' networks (Watts and Strogatz, 1998).

Finally, a notable characteristic of many real world social networks is the relatively high number of nodes with a degree that greatly exceeds the average (where "degree" refers to a node's number of links). This corresponds to the notion that some people have a much larger number of acquaintainces than others. More specifically, many (but not all) social networks show the scale-freeness property (Sen, 2006), i.e., the probability $P(k)$ that a node in the network is connected to k other nodes decays as a power law (Barabási et al., 1999).

Several generative models that mimic these properties of real networks have been suggested in the literature. The small-world model proposed by Watts and Strogatz (1998) generates networks that exhibit small diameter and high clustering. However, they are typically not scale-free.

Barabási et al. (1999) proposed an evolving model based on a preferential attachment mechanism. In this model, one starts with a few nodes linked to each other. Nodes are then added one by one and attached to existing nodes with probabilities according to the degree of the target node. Therefore, the more connected a node is, the more likely it is to receive new links. The resulting networks are scale-free, but typically not highly clustered.

Successful approaches to capture all the desirable features in a single model have been proposed later on (e.g., Dorogovtsev et al., 2002). Our simulation model uses a social network model that is defined in geographical space. Since people in the same locality are typically more likely to know and influence each other (Latane et al., 1995), the spatial distance between nodes needs to be accounted for when constructing links. Therefore, we use an Euclidean network model (Manna and Sen, 2002; Yook et al., 2002) in which the usual attachment probability in the Barabási and Albert model is modulated by a factor related to the geographical distance of the two nodes. Each node is added one by one and connected to its ith predecessor of degree k_i with a link of length ℓ using a probability proportional to $k_i^{\beta}\ell^{\alpha}$. For a limited range of parameters α and β, this model exhibits all three characteristic features mentioned above (Sen and Manna, 2003). In our implementation of the model, we do link each incoming node not only to a single, but to n existing nodes.

Model behavior

Several micro-level mechanisms determine the emergent macro-level behavior of our model, namely communication events, need events, and experience events.

Communication events are generated and scheduled with respect to the arrival processes attached to each edge in the social network. Whenever two agents communicate, they may discuss multiple attributes of a single or multiple products. By doing so, they exchange their respective attribute value estimations, which are based on both information obtained in prior communications and personal experience. In the following, we will call each combination of attribute and product (e.g., "quality of conventional fuel" or "environmental friendliness of the second generation biofuel") a "topic".

Before actually exchanging information, the two agents decide on the topics to discuss by randomly choosing from the sets of products and attributes they are currently aware of. The selection probability of each topic corresponds to the agents' preferences for the respective attribute, since consumers are expected to more often talk about attributes that are important to them than about those playing only a minor role in their individual purchase decisions.

Then, a random number of topics is drawn from the union of the topics chosen individually by each agent to finally establish a set of topics for the communication event. For each of these topics, agents exchange their current estimate (i.e., the expected value based on the information available) of the respective product attribute valuation. The information received is weighted with a

credibility parameter that may vary for different communication partners. This parameter is not necessarily symmetric, i.e., it can be set individually for each end of a bidirectional link connecting acquaintances. Its interpretation is related to the concept of trust, since a receiver will more likely believe the information obtained from a highly credible source. In the context of our model, for example, it is possible to regard a source (e.g., an expert) as particularly credible with respect to a specific issue without having a particularly high level of interpersonal trust. We therefore use the term credibility and define it broadly in the sense of "influence", rather than as an attribute describing the "quality" of the relationship. An exponential factor is applied with the effect that the importance of old information gradually decays as new information arrives.

Furthermore, consumers learn about new products by communicating with agents that are already aware of these products, given that the respective product is chosen as a topic. Finally, when attributes an agent was previously unaware of occur in the list of topics, agents widen the set of attributes that they consider when making purchasing decisions. This mechanism incorporates the effect that innovations may change the way products are evaluated. As an illustrative example, consider the attribute "reduction in gas consumption" of a novel fuel. This is not an attribute that was typically considered by buyers of conventional fuel, but it may become important as soon as fuels with varying characteristics in this respect (such as premium fuels or our application case biofuel) are available.

Need events arise whenever a consumer agent requires the supply of a product. Since our diffusion model deals with products that are consumed on a regular basis, these events are generated repeatedly according to an arrival process that is parameterized individually for each consumer agent. In our application case, we use information on tank size, fuel consumption, and mobility behavior obtained from real consumers for parameterization of the agents' arrival processes.

The purchasing process triggered by a need event can be divided into four distinct phases, namely (i) point of sale selection, (ii) evoked set construction (cf. Narayana and Markin, 1975), (iii) expected utility calculation, and (iv) final purchase decision.

In the first phase, the consumer agent decides at which POS it will make the purchase. For frequently purchased consumable products like fuel, it is reasonable to assume that the POS is selected first and the decision which product to purchase is made at the POS. It is also assumed that consumers tend to visit only a small number of POS on a regular basis. Accordingly, a history list is kept for each agent. Each time a need event arises, it is determined whether the POS is chosen from the history or, alternatively, from the set of all POS, using a fixed probability parameter. In the latter case, the probability of being selected is inversely proportional to the distance between the POS and the agent's home location to which an exponential weighting factor is applied.

In the biofuel application case, it is assumed that once a need for fuel arises, it is always satisfied by purchasing one of the fuels available at the POS. Given the relative inelasticity of short-term fuel demand, it is also reasonable to assume that minor variations in price do not cause immediate changes in an agent's mobility behavior (for a discussion cf. Goodwin et al., 2004).

Next, the consumer agent selects a subset of products that are considered for purchase based on the agent's limited information. For a product to be considered by an agent, it is necessary but not sufficient that the agent is aware of it at the time of purchase and that it is furthermore available at the POS. Additionally, the expected utility of the new product may be required to surpass that of the highest-valued existing alternative product by an agent-specific threshold. This mechanism integrates the concept of innovativeness, which is defined as the "degree to which an individual or other unit of adoption is relatively earlier in adopting new ideas than other members of a social system" (cf. Rogers, 2003, p. 22). In our model, the expected increase in utility required for a new product to be considered for purchase is heterogeneously distributed among the population of agents following Roger's partitioning into five discrete categories: innovators, early adopters, early majority, late majority, and laggards. For innovators, who are characterized as venturesome, the new product consideration threshold parameter will be low or even negative (which implies that innovative products are considered by them even if they provide a lower utility than existing products). For laggards, by contrast, who are characterized as traditional and suspicious of innovations, this parameter is set to a high value.

In the third phase, a utility value is calculated for each product in the evoked set. Based on the agent's current estimation of each product attribute and the individual preferences regarding the respective attribute, partial utilities are obtained for each attribute. While in principle various types of utility functions may be modeled, in our experiments we used a simple additive form and relied on preference data obtained by means of a conjoint analysis. The partial utility values from the conjoint analysis are interpolated to form piecewise linear utility functions for each attribute. Thus, the total utility can be calculated by summing up partial utilities over all of the attributes an agent is aware of.

Finally, a random error is added to the utility of each alternative to model possible mistakes by the consumer, and the final purchase decision is made by selecting the alternative that provides the higher total utility.

Experience events reflect that consumers may obtain information about the characteristics of a product while consuming it, which may alter their notions of the product. In our simulation experiments, a single experience event is scheduled each time a purchase takes place in

the interval between the time of purchase and the next refueling stop.

Personal experience yields new information regarding the estimated attribute values of the purchased product. The "amount" of information obtained depends on the observability of the respective attribute, since some product characteristics can be estimated more easily and directly than others. Analogously to the credibility factor used for weighting information obtained through word-of-mouth communication, observability is used as a weighting factor for new information obtained through personal experience. The ratio of these two factors therefore determines the proportion of influence of word-of-mouth and personal experience. However, while credibility is defined for each communication endpoint irrespective of attributes, observability is defined on a per-attribute basis.

Implementation issues

A formal definition of the simulation model described in the previous section was set down in a detailed specification document, which served as a basis for implementation in Java. Since a complete description of the simulation's software architecture goes beyond the scope of this paper, we will focus on how we addressed specific challenges that arised in the implementation of our agent-based discrete event model described in the previous section.

First, the modeling approach necessitates efficient means for maintaining and processing a list of scheduled events in order for the simulation to be computationally feasible. We decided to rely on the mechanisms provided by MASON (Luke et al., 2004), a fast discrete-event multiagent simulation core written in Java that also provides a fast Mersenne Twister (Matsumoto and Nishimura, 1998) implementation for pseudo-random number generation.

Second, probability distributions are used extensively in our model, e.g., for specifying interarrival times of events. We relied upon the Cern Jet library (`http://acs.lbl.gov/~hoschek/colt/`) to incorporate various types of distributions in our implementation.

Third, recording all simulation events results in the generation of a considerable amount of data that needs to be discretized and analyzed across simulation runs. To this end, we used a flexible logging facility (Apache log4j, `http://logging.apache.org/log4j/`) to produce both comma separated output and optional human readable textual log files. We then used Gnu R to analyze and plot the resulting data.

Fourth, verification of micro-level mechanisms is crucial in agent-based simulation, since implementation errors cannot be easily detected in the simulation's emergent macro-level output. We therefore conducted extensive unit tests of major mechanisms such as consumer agent class methods using jUnit (`http://www.junit.org/`).

Finally, we used geotools GIS toolkit (`http://geotools.codehaus.org/`) for the geographic model implementation and the Java Universal Network/Graph Framework (`http://jung.sourceforge.net`) for representing, visualizing, and analyzing the social network.

SIMULATION EXPERIMENTS

Data acquisition

Potentially relevant product attributes were identified during a thorough discussion with our project partner from the Vienna University of Technology. This technically oriented perspective was complemented with a consumer-oriented view obtained from participants of a focus group. The relevant attributes identified in the focus group were then tested by means of a pre-study with a non-representative convenience sample of 1,000 subjects.

Finally, we commissioned a market research firm to conduct an online survey with 1,000 consumers who were representative for the Austrian market with respect to demographic characteristics such as age, gender, and geographical dispersion. In order to elicit individual consumer preferences, we performed an adaptive conjoint analysis which involved ten paired comparisons of fuels composed from the following attributes and their respective levels: (i) quality (standard or premium), (ii) environmental friendliness (standard or high), (iii) price (€ 1.0,1.1, 1.2, 1.3 or 1.4 per liter, with a reference price for the conventional fuel of € 1.0), (iv) fuel brand (no brand or branded fuel), (v) fuel consumption (equal, 5 % less, or 10 % less than the conventional fuel), and (vi) raw material (crude oil or biomass).

Utility functions for each of the six product attributes were constructed for each individual respondent using partial utility values obtained via a linear programming formulation of the choices provided in the conjoint analysis and interpolating them where needed. In the survey, we also collected individual-level data on tank size, mileage, and average range per filling, which was used to determine the respective parameter values for the simulation. Furthermore, subjects were asked about frequency and interlocutor of communication about fuels in order to adequately model interaction in the social network.

The data collected from each respondent was used to initialize ten identical consumer agents; thus, a total number of 10,000 agents were initialized and distributed in geographical space based on Austrian population density data. More specifically, a 13,997 cell population raster (based on 2001 census data) with a cell size of 2.5km was used. Agents were assigned to cells with a probability proportional to the relative share of the total population in the respective cell and positioned randomly within the target cells.

Based on information obtained in the pre-study and thorough discussions with a sociologist involved in the project, we set social network generation parameters $\alpha = -5$, $\beta = 1$, and $n = 3$ for our simulation experiments.

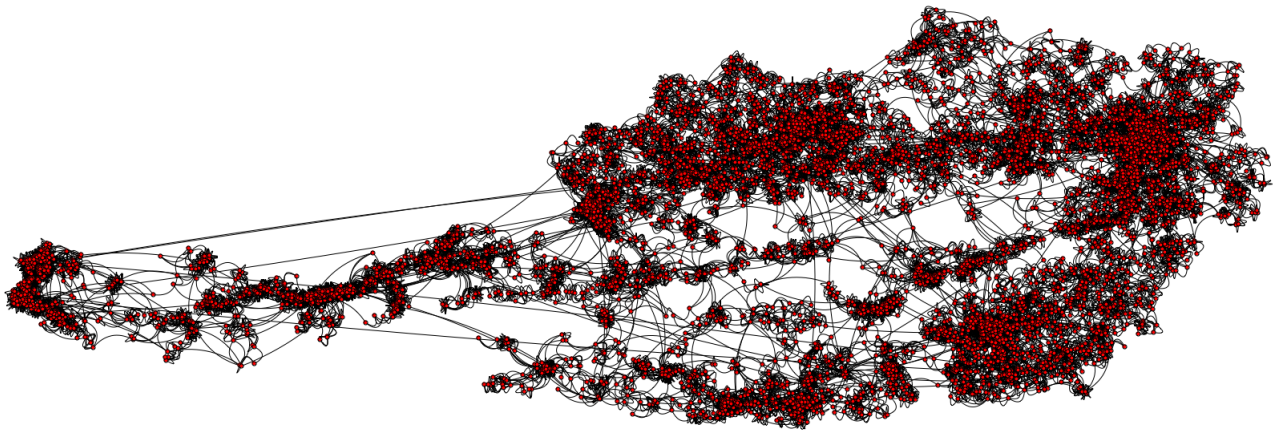


Figure 1: Social network used in the simulation experiments

The resulting network depicted in Figure 1 has an average shortest path of 6.86 and an average clustering coefficient (Watts and Strogatz, 1998) of 0.446.

The geographical location of 1,183 fuel stations was obtained from OpenStreetMap (`http://www.openstreetmap.org`) and used to distribute POS in the simulated geography.

In order to establish a frame of reference and facilitate interpretation of results, all parameters used in the simulation were scaled so that one time unit in the simulation corresponds to one day.

Reference scenario

In the reference scenario, a single type of conventional fuel is available at the beginning of the simulation. At time $t = 10$, the biofuel is introduced at all points of sale and its existence is communicated to ten randomly selected "seed agents" (i.e., 0.01 % of the population); each seed agent receives information about the true attribute values of the product. The time horizon is set to 350 time units, which rougly corresponds to 12 months. Furthermore, we assume a price of € 1.0 per liter for the conventional fuel and conduct experiments with five price levels for the biofuel, namely € 1.0,1.1,1.2,1.3 and 1.4 per liter. We performed 50 simulation runs at each price level.

Results

First, it is interesting to look at adoption, i.e., the development over time of first purchases of the biofuel in the population without considering repeat purchases. The cumulative adoption curves presented in Figure 2 plot the share of consumers that have adopted the innovation over time. The curve for each price level was obtained by discretizing the continuous-time cumulative adoption in intervals of one and calculating average values over all simulation runs in each resulting period. Superimposed box plots indicate the distribution of realizations at $t = \{50, 100, 150, 200, 250, 300, 350\}$. The curves exhibit the typical S-shape commonly observed in empirical diffusion studies (cf. Mahajan et al., 1995). As can be expected, a higher price consistently results in a slower speed of adoption and a higher delay before take-off occurs. Moreover, we see that there is a group of consumers that is largely "immune" to adoption, even at a price equal to the price of the conventional fuel.

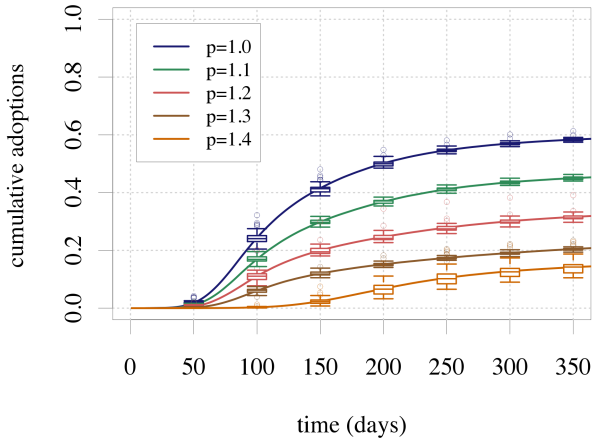


Figure 2: Cumulative adoptions over time at biofuel price levels $p = \{1.0, 1.1, 1.2, 1.3, 1.4\}$ (lines indicate averages, boxplots show results of 50 simulation runs each)

Unlike most diffusion studies, we are not concerned with the diffusion of a consumer durable good, but rather a frequently purchased consumable product. Initial adoption alone is therefore not the only aspect worth considering. Decision makers may be more interested in the question of how large a market share the innovation may obtain when taking repeat purchases into account. To this end, Figure 3 plots revenue market share at each of the five price levels over time. Points are used to mark the results of individual simulation runs. Results indicate that within the first year after market introduction, a considerable share of total revenues could be obtained, even at a price significantly above that of conventional fuel. After $t = 350$, market penetration does not increase significantly any further. For the price levels considered, we find that a higher price leads to a lower revenue market share (which must not necessarily be the case since we

consider revenue, not units). More interestingly, we find that the market share curves are remarkably similar to the adoption curves presented in Figure 2. This implies that most adopters do not purchase the novel product only once, but tend to repurchase it on a continuous basis.

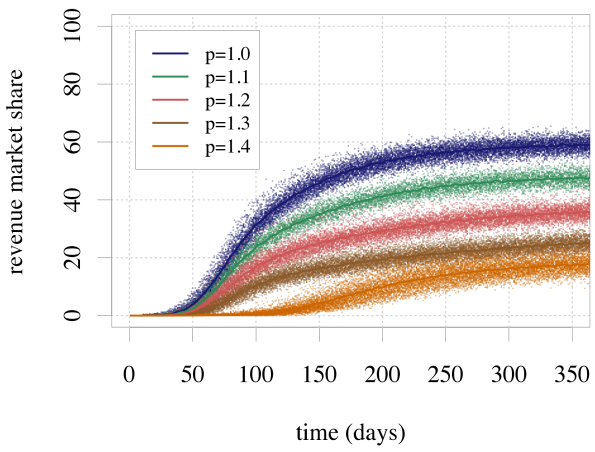


Figure 3: Revenue market share at biofuel price levels $p = \{1.0, 1.1, 1.2, 1.3, 1.4\}$ (lines indicate averages, points mark results of single simulation runs)

While these results do not uncover unexpected or non-intuitive insights, they do enhance confidence in the validity of the model. Furthermore, they are robust despite the complexity and the large number of non-deterministic elements in the simulation. Diffusion curves similar to those presented in Figure 2 could also be obtained by estimating parameters for aggregate models such as the Bass model (Bass, 1969). However, a priori estimation of these parameters is problematic, since they vary widely across applications (Sultan et al., 1990) and at least early data points are needed for estimation. Apart from not relying on single point coefficient estimates, the uniqueness of the "bottom-up" simulation approach results from its ability to permit experimentation. Decision makers can run the simulation with various adaptive pricing strategies while taking potential reaction of competitors into account. Furthermore, they can create scenarios with varying distribution strategies to account for supply limitations and forecast development of market share for each of them, obtaining not only single point estimates, but a distribution of realizations.

CONCLUSIONS

In this work, we have introduced an agent-based model for simulating the diffusion of a second generation biofuel. The ABM is spatially explicit, it embeds consumers and POS in geographic space and accounts for distance in the generation of social networks. It incorporates social interactions that typically play a major role in the propagation of innovations among potential consumers and accounts for individual consumer-related characteristics (e.g., preferences, mobility behavior). Our approach distinguishes itself by accounting for several product attributes and – in addition to the initial purchase decision – also considering repeat purchases. These features are particularly worthwhile for our application case.

Our findings for the Austrian market suggest that while a competitive price is unsurprisingly an important driver for adoption, there exists opportunity even at a price that lies clearly above that of conventional fuel. The simulation results also indicate that a considerable market share could be achieved within the first year following the introduction of such a product. Both results should be of value for investors when planning the market introduction of a second generation biofuel.

Although a discussion of potential pricing strategies and their effects on the diffusion process goes beyond the scope of this work, we would like to point out that the simulation can support a decision maker also in this respect. For instance, the pros and cons of a skimming strategy (i.e., setting a relatively high price first and lower it over time to capture the consumer surplus) versus a penetration strategy (i.e., setting a low initial entry price to accelerate the diffusion process) can be easily assessed by means of the simulation. Furthermore, various energy market scenarios and their impact on the competitiveness of alternative fuels can be simulated. Finally, the simulation enables a decision maker to test the effectiveness of various approaches towards selecting POS for distribution, while accounting for limited production capacity, availability of rich sources of biomass, and the geographic concentration of consumers.

Further research will be conducted in four directions. First, we plan to put additional effort into more thoroughly modeling and validating our social network generation method. To this end, we conducted a sociological survey and are currently in the process of analyzing the data obtained. Second, we intend to model various types of promotional activities, which would enable decision makers to simulate the effects of various communication strategies on the diffusion process. Third, to more accurately reflect both the fact that supply is limited as well as the econmics of fuel distribution, we are currently extending the model with active point of sale agents that discontinue a product if it does not fulfill minimum sales criteria. We expect this extension to result in more interesting diffusion patterns and in a more realistic consideration of the possibility that the diffusion may fail. Fourth, due to the predictive nature of the problem, validation of the entire model is inherently difficult. *A posteriori* validation with real world data obtained after product launch may be possible in the future, but would not be particularly useful after the fact. Alternatively, we can only resort to historic data on the diffusion process of similar products and use that information to validate the model. We plan to investigate feasibility of this approach for our application case using data on premium fuel adoption.

ACKNOWLEDGMENTS

We thank Bernd Brandl and Stefan Fürnsinn for supporting this work with their expertise in sociology and second generation biofuels, respectively. Financial support from the Austrian Science Fund (FWF) by grant No. P20136-G14 is gratefully acknowledged.

REFERENCES

Alkemade, F. and Castaldi, C. (2005). Strategies for the diffusion of innovations on social networks. *Computational Economics*, 25(1-2):3–23.

Barabási, A. L., Albert, R., and Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1-2):173–187.

Bass, F. (1969). A new product growth for model consumer durables. *Management Science*, 15(5):215–227.

Bass, F. M., Krishnan, T. V., and Jain, D. C. (1994). Why the bass model fits without decision variables. *Marketing Science*, 13(3):203–223.

Borshchev, A. and Filippov, A. (2004). From system dynamics and discrete event to practical agent based modeling: Reasons, techniques, tools. In *Proceedings of the 22nd International Conference of the Systems Dynamics Society*, pages 1–22, Oxford.

Deffuant, G., Huet, S., and Amblard, F. (2005). An individual-based model of innovation diffusion mixing social value and individual benefit. *American Journal of Sociology*, 110(4):1041–1069.

Delre, S. A., Jager, W., Bijmolt, T. H. A., and Janssen, M. A. (2007a). Targeting and timing promotional activities: An agent-based model for the takeoff of new products. *Journal of Business Research*, 60(8):826–835.

Delre, S. A., Jager, W., and Janssen, M. A. (2007b). Diffusion dynamics in small-world networks with heterogeneous consumers. *Computational & Mathematical Organization Theory*, 13(2):185–202.

Dorogovtsev, S. N., Goltsev, A. V., and Mendes, J. F. F. (2002). Pseudofractal scale-free web. *Physical Review E*, 65(6):066122.

Easingwood, C. J., Mahajan, V., and Muller, E. (1983). A nonuniform influence innovation diffusion model of new product acceptance. *Marketing Science*, 2(3):273–295.

European Environment Agency (2008). Greenhouse gas emission trends and projections in Europe 2008. Technical Report No 5/2008, Copenhagen, Denmark.

Fargione, J., Hill, J., Tilman, D., Polasky, S., and Hawthorne, P. (2008). Land clearing and the biofuel carbon debt. *Science*, 319:1235–1238.

Fürnsinn, S. (2007). *Outwitting the dilemma of scale: Cost and energy efficient scale-down of the Fischer-Tropsch fuel production from biomass*. Phd thesis, Vienna University of Technology.

Garcia, R. (2005). Uses of agent-based modeling in innovation/new product development research. *Journal of Product Innovation Management*, 22(5):380–398.

Goodwin, P., Dargay, J., and Hanly, M. (2004). Elasticities of road traffic and fuel consumption with respect to price and income: A review. *Transport Reviews*, 24(3):275–292.

Hofbauer, H., Rauch, R., Fürnsinn, S., and Aichernig, C. (2005). Energiezentrale Güssing. Technical report, Bundesministerium für Verkehr, Innovation und Technologie, Vienna, Austria.

Inderwildi, O. R. and King, D. A. (2009). Quo vadis biofuels? *Energy & Environmental Science*, 2:343–346.

Jain, D. C. and Rao, R. C. (1990). Effect of price on the demand for durables: Modeling, estimation, and findings. *Journal of Business & Economic Statistics*, 8(2):163–170.

Kiesling, E., Günther, M., Stummer, C., and Wakolbinger, L. M. (2009). An agent-based simulation model for the market diffusion of a second generation biofuel. In Rossetti, M., Hill, R., Johansson, B., Dunkin, A., and Ingalls, R., editors, *Proceedings of the 2009 Winter Simulation Conference*, pages 1474–1481, Austin, TX. Omnipress.

Latane, B., Liu, J. H., Nowak, A., Bonevento, M., and Zheng, L. (1995). Distance matters: Physical space and social impact. *Personality and Social Psychology Bulletin*, 21(8):795–805.

Logan, W. S. (2008). *Water Implications of Biofuels Production in the United States*. The National Academy Press, Washington, D.C.

Luke, S., Cioffi-Revilla, C., Panait, L., and Sullivan, K. (2004). MASON: A new multi-agent simulation toolkit. In *Proceedings of the 2004 SwarmFest Workshop*.

Mahajan, V., Muller, E., and Bass, F. M. (1995). Diffusion of new products: Empirical generalizations and managerial uses. *Marketing Science*, 14(3):G79–G88.

Manna, S. S. and Sen, P. (2002). Modulated scale-free network in Euclidean space. *Physical Review E*, 66(6):066114.

Matsumoto, M. and Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, 8(1):3–30.

Narayana, C. L. and Markin, R. J. (1975). Consumer behavior and product performance: An alternative conceptualization. *Journal of Marketing*, 39(4):1–6.

Norton, J. A. and Bass, F. M. (1987). A diffusion theory model of adoption and substitution for successive generations of high-technology products. *Management Science*, 33(9):1069–1086.

O'Neill, B. C., Riahi, K., and Keppo, I. (2010). Mitigation implications of midcentury targets that preserve long-term climate policy options. *Proceedings of the National Academy of Sciences*, 107(3):1011–1016.

Pimentel, D. and Patzek, T. (2005). Ethanol production using corn, switchgrass and wood; biodiesel production using soybean. *Natural Resources Research*, 14(1):65–76.

Robinson, B. and Lakhani, C. (1975). Dynamic price models for new-product planning. *Management Science*, 21(10):1113–1122.

Rogers, E. M. (2003). *Diffusion of Innovations*. Free Press, New York, NY.

Schenk, T. A., Löffler, G., and Rauh, J. (2007). Agent-based simulation of consumer behavior in grocery shopping on a regional level. *Journal of Business Research*, 60(8):894–903.

Schwoon, M. (2006). Simulating the adoption of fuel cell vehicles. *Journal of Evolutionary Economics*, 16(4):435–472.

Sen, P. (2006). Complexities of social networks: A physicist's perspective. In Chakrabarti, B. K., Chakraborti, A., and Chatterjee, A., editors, *Econophysics and Sociophysics: Trends and Perspectives*, pages 473–506, Weinheim. Wiley.

Sen, P. and Manna, S. S. (2003). Clustering properties of a generalized critical euclidean network. *Physical Review E*, 68(2):026104.

Sultan, F., Farley, J., and Lehmann, D. (1990). A meta-analysis of applications of diffusion models. *Journal of Marketing Research*, 27(1):70–77.

Travers, J. and Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, 32(4):425–443.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, 393(6684):440–442.

Yook, S.-H., Jeong, H., and Barabási, A.-L. (2002). Modeling the Internet's large-scale topology. *Proceedings of the National Academy of Sciences of the United States of America*, 99(21):13382–13386.

AUTHOR BIOGRAPHIES

ELMAR KIESLING is a doctoral student at the University of Vienna, Austria, from which he holds a master's degree in international business administration. Since 2007, he holds the position of a junior researcher and lecturer at this institution. His research interests include multi-criteria decision support systems, agent-based modeling of diffusion processes, and information security management. His email is: elmar.kiesling@univie.ac.at.

MARKUS GÜNTHER earned a master's degree in international business administration as well as a doctoral degree in economics and social sciences from the University of Vienna. After professional experiences as an IT freelancer in the banking-sector and the automotive industry (corporate R&D and process management), he currently holds a position as researcher at the School of Business, Economics, and Statistics in the working group on Innovation and Technology Management. His research lies in quantitative modeling on topics in innovation and technology management like the role of knowledge and information-sharing or new product diffusion. Previous work has been published in several edited volumes as well as in the European Journal of Operational Research, Technovation, Journal of the Operational Research Society, and in Die Betriebswirtschaft. His email is markus.guenther@univie.ac.at

CHRISTIAN STUMMER is an associate professor of innovation and technology management at the Department of Business Administration, University of Vienna, Austria. He has earned a master's degree in business informatics as well as a doctoral degree in economics and the social sciences. Prior working experiences include positions as a visiting professor at the University of Texas and as the head of a research group at the Electronic Commerce Competence Center in Vienna. His research interest lies in quantitative modeling and providing proper decision support in innovation and technology management particularly with respect to new product diffusion, project portfolio selection, and resource allocation. Christian Stummer has (co-) authored two books, more than two dozen papers in peer-reviewed scientific journals, and numerous reviewed contributions to books and proceedings. He is a member of INFORMS, GOR, ÖGOR, OCG, and the MCDM Society. His email is: christian.stummer@univie.ac.at.

RUDOLF VETSCHERA is a professor of organization and planning at the School of Business, Economics and Statistics, University of Vienna, Austria. He holds a PhD in economics and social sciences from the University of Vienna, Austria. Before his current position, he was full professor of Business Administration at the University of Konstanz, Germany. He has published three books and about seventy papers in reviewed journals and collective volumes. His main research area is in the intersection of organization, decision theory, and information systems, in particular negotiations and decisions under incomplete information. His email is: rudolf.vetschera@univie.ac.at

LEA M. WAKOLBINGER studied at Johann Wolfgang Goethe University Frankfurt, Germany, Case Western Reserve University, Cleveland, United States, and University of Vienna, Austria. She holds a master's degree in international business administration as well as a doctoral degree in economics and the social sciencies from the University of Vienna. After her graduation in 2006, Lea M. Wakolbinger worked as a researcher at the Electronic Commerce Competence Center in Vienna, where she conducted theoretical research in the area of multichannel management and cross-media advertising. Since 2008 she holds the position as a researcher and lecturer at the University of Vienna. Previous work has been published in several peer reviewed scientific journals. Her email is: lea.wakolbinger@univie.ac.at.

TOWARDS ONTOLOGY-BASED MULTIAGENT SIMULATIONS: THE PLASMA APPROACH*

Tobias Warden, Robert Porzel, Jan D. Gehrke, Otthein Herzog, Hagen Langer, Rainer Malaka
Center for Computing and Communication Technologies (TZI)
Universität Bremen
Bremen, Germany
Email: {warden,porzel,jgehrke,herzog,hlanger,malaka}@tzi.de

KEYWORDS

Multiagent-based Simulation, Simulation Modelling, Ontology Engineering.

ABSTRACT

In multiagent-based simulation systems the agent programming paradigm is adopted for simulation. This simulation approach offers the promise to facilitate the design and development of complex simulations, both regarding the distinct simulation actors and the simulation environment itself. We introduce the simulation middleware PlaSMA which extends the JADE agent framework with a simulation control that ensures synchronization and provides a world model based on a formal ontological description of the respective application domain. We illustrate the benefits of an ontology grounding for simulation design and discuss further gains to be expected from recent advances in ontology engineering, namely the adaption of foundational ontologies and modelling-patterns.

INTRODUCTION

Multiagent-based simulation (MABS) has been employed successfully for system analysis and evaluation in a variety of domains ranging from simulating models of bee recruitment to simulating complex business processes, such as supply chain management. The approach lends itself in particular to the simulation of complex systems on the micro-level where individual decision makers are modeled explicitly as autonomous agents embedded in dynamic environments. In contrast to alternatives, such as equation-based modelling, these modelling approaches facilitate the design of complex systems due to task decomposition, a natural mapping from real-world actors or entities to agents, the focus on modelling of individual behaviour (Parunak et al., 1998).

Still, the concrete decision to adopt multiagent-based simulation for evaluation purposes when starting from a blank slate, is often perceived as a mixed blessing as the effort required to design particular multiagent-based simulations, especially when scaling the number of involved agents and environmental complexity, often exceeds that of comparable simulation approaches, such as Petri-nets or queuing networks (Klügl et al., 2002). Off-the-shelf agent frameworks are typically not designed to consider simulation-specific issues, such as synchronization, for which solutions exist in standard simulation approaches (Bobeanu et al., 2004). In addition, the design of the simulation environment itself in which the MAS can be placed, requires significant development resources. Thus, there is still an engineering challenge for multiagent-based modelling and simulation to be addressed.

In this work, we introduce the multiagent-based simulation system PlaSMA and focus on a description of its ontology-based simulation world model. We thereby address how the design challenge to create scalable simulations for complex domains is approached in the PlaSMA system. To that end, the system provides reusable and extendable ontological models based on an explicit logical calculus. For evaluating our approach, we show how the system has been used successfully for simulations in the domain of autonomous logistic processes, where different approaches for self-organized decision making have been examined in terms of their robustness and adaptivity, e.g. in transport processes in dynamic environments (Hülsmann and Windt, 2007). We also describe ongoing work to generalize PlaSMA and use recent advances in ontology engineering, which further facilitates simulation design, interaction and analysis.

PLASMA SIMULATION PLATFORM

The PlaSMA system provides a distributed multiagent-based simulation and demonstration system based on the FIPA-compliant Java Agent Development Framework JADE (Bellifemine et al., 2007)[1]. Although the primary application domain is logistics, PlaSMA is, in principle, applicable for other simulation domains.

The simulation system consists of the basic components simulation control, world model, simulation agents, analysis, and user interface. Simulation control handles world model initialization, provides an interface for world model access, and performs agent lifecycle and simulation time management. Simulation control is jointly executed by a

*This research has been funded by the German Research Foundation (DFG) within the Collaborative Research Centre (CRC) 637 "Autonomous Cooperating Logistic Processes – A Paradigm Shift and its Limitations" at the Universität Bremen, Germany.

[1] *PlaSMA*, the **Pla**tform for **S**imulations with **M**ultiple **A**gents, is developed at the CRC 637 "Autonomous Cooperating Logistic Processes". PlaSMA is available at: `plasma.informatik.uni-bremen.de`

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

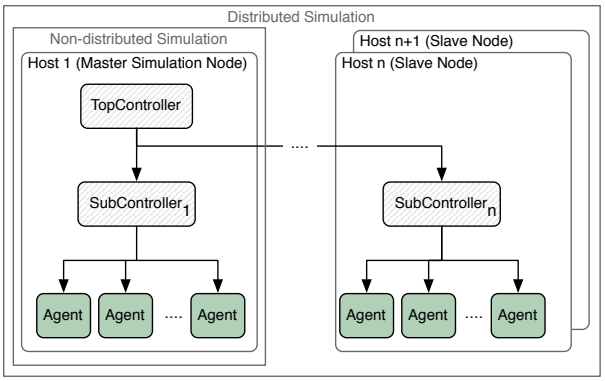


Figure 1: PlaSMA simulation control model

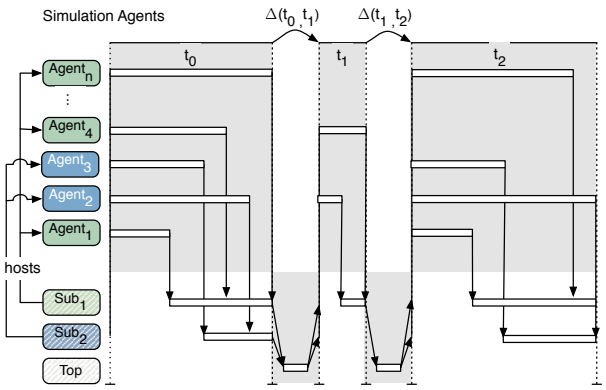


Figure 2: Course of a PlaSMA simulation.

single top-level controller and an additional sub-controller for each processor or computer in distributed settings as depicted in Figure 1. Sub-controllers handle those software agents (called *simulation agents*) that are the actors in a simulation scenario. The interaction between top-controller and sub-controllers concerns agent lifecycle management, runtime control, and time events.

Simulation Time Management

In general, MABS can combine distributed discrete event or time-stepped simulation with decision-making encapsulated in agents as separate and concurrent logical processes (Parunak et al., 1998). In classical simulation systems, the logical processes involved as well as interaction links have to be known in advance and must not change during simulation (Fujimoto, 2000). This is not the case in MABS as each agent may interact with all other agents (Lees et al., 2004). Agents may join or leave simulation during execution, e.g. depending on a stochastic simulation model or human intervention. This flexibility, however, complicates simulation time management.

Initially, it is necessary to distinguish different notions of time related to multiagent-based simulation. Generally, *physical time* refers to the time at which simulated events happen in the real world. *Simulation time* (or *virtual time*) models physical time in simulation. In scenarios where the advancement of *local virtual (simulation) time* or in short LVT (Jefferson, 1990) is directly coupled with the progress of each individual agent in executing its behaviours, heterogeneities in the computational demands of agents in the simulation and the distribution of agents across hosts platforms with varying computational power provoke a problematic divergence of LVTs, leading to the so-called *causality problem* (Fujimoto, 2000). In order to guarantee correct and reproducible simulations (Gehrke et al., 2008), the simulation system has thus to ensure that agents process events in accordance to their time-stamp order. This requirement is addressed by synchronization, which can be either optimistic or conservative in character. In optimistic synchronization, the progression of local virtual time for each agent is in general not restricted which allows executing simulations efficiently since fast processes do not have to wait for slower ones. Optimistic synchronization is demanding in implementation and has high requirements regarding memory. Since cascading rollbacks might require rewinding many steps back in time, preceding execution states must be retained for each agent. Conservative synchronization, by contrast, prevents causality problems by ensuring the correct order of event processing at each time by means of explicit coordination. The price to be paid is thereby a lesser speedup that achievable by parallelism.

The choice of synchronization a mechanism is an important design decision for the implementation of a MABS system. Based on the identification of important quality criteria for MABS systems presented in (Gehrke et al., 2008), a coordinated conservative synchronization approach has been adopted in PlaSMA which is handled concertedly by the simulation controller hierarchy. As shown in Figure 2, the top-controller sends time events to each sub-controller to indicate progression of simulation time. Based on this information, the sub-controllers identify the respective subset of locally managed simulation agents due to act at the next time-stamp, advance the LVT of these agents to the communicated time stamp, and finally send wake-up notifications. The agents then take over from the controllers to perform a single walk through their respective behaviour structure. Computation time will vary depending on the complexity and diversity of agent tasks and the situation at hand. At the end of each action cycle each simulation agent needs to inform its local sub-controller of its requirements with regard to continuative operation. Thereby, the agent may choose to declare explicitly a particular time point in the future. Alternatively, it can choose not to declare such a particular date and rather wait until operation is necessitated due to the reception of messages communicated by other agents or the completion of a previously commissioned action. The activity requests are consolidated by the sub-controllers such that only a single result is propagated further to the top-level controller. The latter then computes the next simulation time-stamp based on these requests. Having done so, it also computes the progress in the physical simulation world model, thus setting the stage for the next action cycle of the simulation agents.

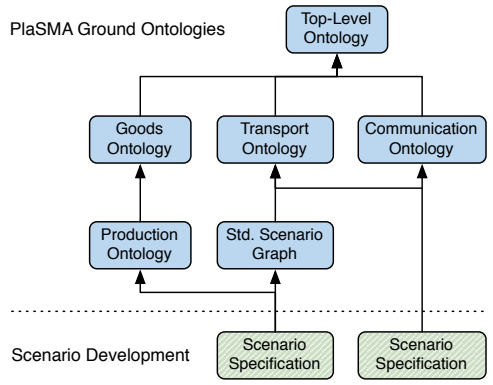


Figure 3: Modular structure of PlaSMA ontologies.

Ontology-based Simulation World Model

The PlaSMA implementation provides a model of the physical world which is based on a declarative, formal, and explicit model expressed as an OWL-DL ontology (Bechhofer et al., 2004). In this way, the initial process of scenario design is turned into an ontology engineering task for which standard tools and modeling principles exist. One of the main benefits of this ontology-based approach can, therefore, be regarded to lie in this *standardization* of the scenario design process.

Our implementation of the PlaSMA simulation system provides a modular set of five ontologies which specify terminological knowledge relevant in the logistics domain. These ontology modules, whose link-up is depicted schematically in Figure 3, constitute the formal basis for scenario modelling at design time and are briefly described below.

- TLO The top-level domain ontology for logistic scenarios specifies general types of physical objects, the basics of traffic infrastructure, organizational membership, ownership of physical objects, and software agents responsible for a set of physical objects or providing abstract services. All other ontologies import and extend this ontology.
- TRANS As ontology for multimodal transportation, this ontology defines location and site types, transport relations in the traffic infrastructure as well as means of transport, handling and loading equipment.
- PROD As ontology for intra- and shop floor logistics, this ontology specifies production resources, and orders with their respective properties.
- COM This communication ontology defines communication and computation devices with properties such as radio coverage or power supply type.
- GOODS The goods ontology provides a general schema to classify goods. For instance, classification may distinguish physical or non-physical good (e.g., property rights), packaged or bulk good, perishable or hazardous goods.

These ontologies are neither considered complete with regard to the diversity of logistic sub fields (i.e. the current focus is on distribution and, to a lesser degree, shop floor logistics) nor mandatory for simulations of other application domains. Therefore, the simulation designer is free to use only a custom subset of core modules, may adapt or extend these with additional concept and property assertions, or create substitute ontologies suitable to model a new domain of interest. How this can be achieved in a systematic fashion is discussed later on.

Scenario Modelling: Infrastructure

In most application scenarios which have been implemented in PlaSMA so far, i.e. in transport logistics (Gehrke and Wojtusiak, 2008) and spontaneous ride sharing (Xing et al., 2009), a basic traffic network is specified as an annotated directed graph. Vertices therein are non-physical locations in the modeled world. In all but synthetic scenarios, these bear a geographic grounding in the form of geo-coordinates and host stationary logistic infrastructure such as production facilities, harbors, cross-docking stations and warehouses. The transport ontology module provides several types of edges which can be classified as specializations of the base concepts LandRoute, WaterRoute and AirRoute, thus allowing for the modelling of multimodal transport relations between locations. Besides the stationary logistic resources, the scenario ontology also specifies non-stationary logistic resources, in particular means of transport and various flavors of freight objects. The modularity provided by OWL thereby allows for a separate modelling of a basic scenario graph and scenario-specific entities populating that graph, thus rendering the former reusable across scenarios.

Scenario Modelling: Agents

Next to this basic model of the physical environment, the PlaSMA ontologies allow for the specification of non-physical entities, such as organizations, and individual agents in their role as decision makers. Modelling of organizational contexts allows for an adequate representation of owner-, responsibility- and membership relations.

Of particular importance with regard to simulation execution is the classification of agent types adopted in PlaSMA and the association of these software agents with entities within the simulation environment. The adopted modelling approach conceptually introduces a partition of all software agents, which constitute a simulation, into distinct agent communities, namely simulation actors and environmental agents.

The former community, made up of object- and service agents, constitutes the MAS which has been deployed in the simulation environment in order to evaluate global performance, patterns of interaction among or the design of particular agents. Object agents in PlaSMA act as artificial autonomous decision makers on behalf of particular physical entities. They may either assume the role of an authoritative digital representative or conduct secondary functions. Service agents offer abstract services to fellow simulation actors such as traffic information, weather prediction or electronic market places. PlaSMA allows for an intersection of both actor groups such that a particular

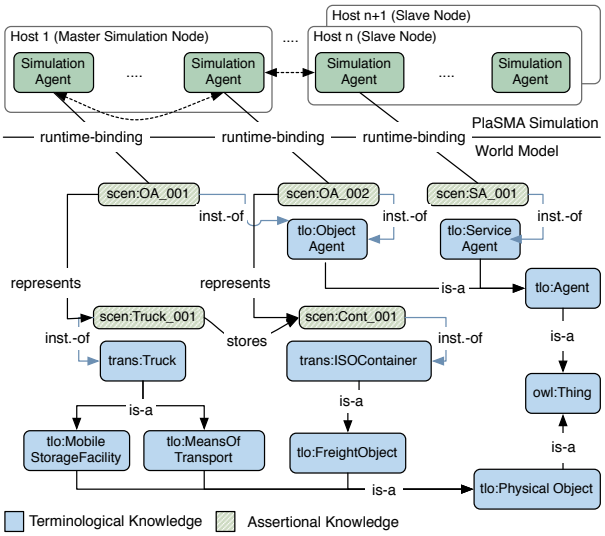


Figure 4: Interrelation of simulation agent, ontological object/service counterpart and physical objects.

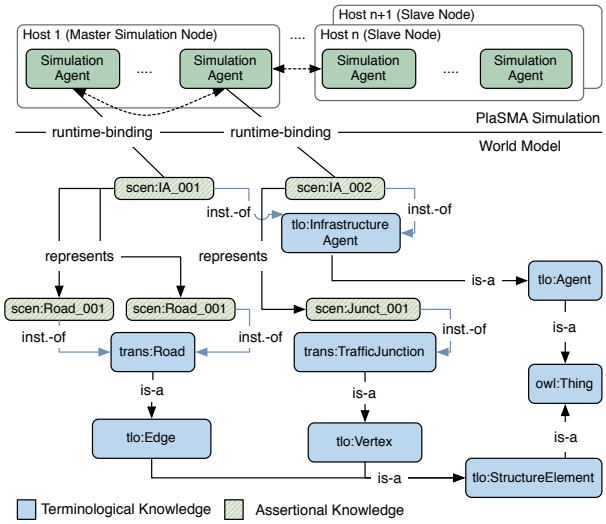


Figure 5: Interrelation of simulation agent, ontological infrastructure counterpart and structure elements.

agent may simultaneously manage physical objects and provide additional services.

The association of object agents and physical entities, whose initial state is modeled at design time, is shown in Figure 4. It may be subject to change as simulation runs unfold. Due to the explicit differentiation of agents and entities, agent modelling enables what may be called *dynamic embodiment* with a mutable 1:n relation from agent to managed entities. The modelling also provides for the additional category of environmental agents whose function can be explained by a theatre analogy. If we think of the modeled environment as set for simulation actors, the former agents as stage technicians modify the set over the course of scenario runs. They are responsible for runtime modification of both topology and characteristics of infrastructure elements within the simulation environment as shown in Figure 5. These agents are also responsible for online creation and destruction of physical entities. In contrast to object and service agents which are modeled explicitly in the PlaSMA top-level ontology, environmental agents are modeled indirectly to retain flexibility as they combine traits of object agent, e.g. when a new container is created and immediately included in a storage facility inventory, and infrastructure agents.

Scenario Modeling: Validation

Besides advantages such as extensibility and a clear-cut ontological grounding of all simulation constituents, the modelling approach on the basis of OWL-DL adopted for PlaSMA allows to leverage inferential capabilities of dedicated ontology reasoning systems in order to ensure the logical consistency and validity of simulation models at design time. This holds both with respect to terminological knowledge which describes the simulation domain on the schema level and assertional knowledge, i.e. the particular scenario specification. Modelling flaws can thus be rectified early in development, thereby reducing the number of modelling iterations due to shortcomings only spuriously discovered at run time.

World Model Interaction

The ontological modelling of a simulation scenario for PlaSMA specifies the initial state of the environment and the agent to object associations. However, for the simulation to proceed, the ontological agent instances need to be bound to applicable PlaSMA agents written in Java.

Once instantiated, the simulation agents have access to their ontological specification using world model queries. In the example presented in Figure 4, the truck agent can retrieve its managed truck objects and their respective status such as information about loaded freight objects. Besides retrieval, the simulation world model also provides means for its manipulation via two related mechanisms, namely physical actions for use by simulation agents and environment events for use by environmental agents. Depending on the modelling granularity chosen for a particular scenario, the aforementioned actions may thereby correspond immediately to actions associated with physical entities, for instance the *drive* action of a truck, but also to complex actions such as *cargo transport between storage facilities*.

PlaSMA provides a growing library of reusable logistic standard actions and an API to create additional actions. Custom actions are programmatically specified in terms of a) preconditions to be met for their execution, b) concluding effects in the world model, and c), for protracted rather than point-wise actions, transitional action effects setting in at the begin of action execution. Due to the ontological foundation of the PlaSMA world model it is possible at run time to exploit inferential capabilities exposed through the world model query interface. For instance, asserted and derived classes of a particular graph edge passed to a drive action can be ascertained and matched with the action specification. Actions in PlaSMA can be conceived as expansion of the world model whose scope is respectively tailored to the domain of simulation. Thus,

they are granted unconstrained read access, allowing for context-sensitive computation of action effects. For instance, environmental events such as traffic jams or severe weather conditions can significantly prolong the execution of drive actions on affected transport relations. PlaSMA ensures that active actions are notified of world model changes. They then need to interpret these changes with regard to their own execution and, if need be, internally compute action effects incrementally.

The world model allows simulation agents to dynamically create additional agents over the whole course of a simulation run. In autonomous logistics, the runtime creation of new agents is often motivated by dynamic production or transport order inflow. For instance, an agent which manages incoming transport orders for a forwarding agency, might want to delegate the supervision of particular orders to specialized agents. These can be instantiated on demand rather than creating a fixed pool of handling agents upon simulation start.

For further development of the PlaSMA simulation world model, the following extensions are considered. Firstly, there is a need for an automated detection and subsequent resolution of conflicts that arise due to simultaneous or temporally overlapping computation mutually exclusive actions. Secondly, the world model query interface so far does not constrain the retrieval of world model information such that it is up to the agent programmer to implement an adequate scope visibility. In order to disburden the programmer, object agents require an explicit perception which accommodates their dynamic embodiment as they assume control of one or multiple logistic entities at a time. In this case, these agents should be constrained to perceive the environment based on sensors attached to represented objects. Additional challenges to be met by object agents then comprise the locality and incompleteness of perception as well as noise which can be introduced by the sensor models.

Scenario Visualization and Control

PlaSMA comprises a visualization client which allows runtime control and progress monitoring of simulation experiments. To initiate a simulation run, the client is connected to the main PlaSMA simulation node. The experimenter may choose from previously deployed scenarios whose runtime parameters such as simulation start and duration, number of successive runs, or logging granularity can be customized. Once a scenario is loaded, the client leverages the NASA World Wind[2] mapping engine to render the annotated directed graph which represents the multimodal transport network as well as the physical logistic entities thereon onto a virtual globe. It also provides a tree representation of both agents as simulation actors and physical objects. In particular, the tree view continuously depicts the associations between software agents in their roles as autonomous decision makers and managed logistic entities, either from the agent perspective (i.e. objects managed by respective agents) or the entity perspective (i.e. agents acting either authoritatively or in secondary roles for respective objects). The selection of entities or agents in the simulation delivers insight to their associated world model state which is useful when tracing entities during simulation runs. The PlaSMA client offers support for a distributed monitoring of simulation runs. Multiple client instances may connect to the same PlaSMA server. This allows for monitoring of joint experiments from different locations via multiple viewports.

Work on extending the client's scope of operation is currently underway. Therein, we focus on supporting intuitive collaborative interaction with larger, and therefore more complex, simulations. Within this approach we investigate multi-touch surface computing environments as a means to provide an intuitive interaction framework which allows human agents a) to manipulate the physical simulation environment directly via environment events and b) to allow for online human-agent interaction. The latter thereby renders possible the involvement of human actors as an additional category of decision makers in what could then be considered a *participative* simulation.

APPLICATION AND EVALUATION

Although actively used as a joint experiment platform by several research groups within the Collaborative Research Centre (CRC) 637 for four years, PlaSMA is still in a prototype stage of development. It is applied for comparison and evaluation of algorithms for logistics planning and special sub-processes therein, such as coordination mechanisms of logistics objects, information distribution, environment adaptation and prediction (Gehrke and Wojtusiak, 2008) as well as routing and cargo clustering algorithms (Schuldt and Werner, 2007). Furthermore, PlaSMA is part of the *Intelligent Container* demonstration platform[3] integrating simulation with real-world hardware in perishable food transport scenarios. In the context of adaptive route planning, PlaSMA has been integrated with the AQ21 machine learning system for prediction of expected traffic and speed on potential routes (Gehrke and Wojtusiak, 2008). PlaSMA is currently integrated with the learnable evolution model (LEM), a library for non-Darwinian evolutionary computation that employs machine learning to guide evolutionary processes (Wojtusiak, 2009). Complexity of simulation surveys ranges from very few agents to large agent communities (20,000).

EXTENDING PLASMA SCOPE AND USABILITY

While the successful application and evaluation so far has adduced initial evidence that the PlaSMA system facilitates the compilation of simulations for multiple scenarios in our application domains, there is still potential for development which we seek to explore in ongoing research.

[2]Web Site: `http://worldwind.arc.nasa.gov/java/` (visited: 2010/02/25)

[3]Web Site: `http://www.intelligentcontainer.com` (visited: 2010/02/25)

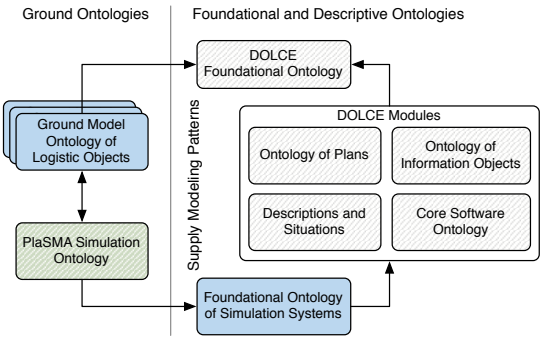


Figure 6: Framework for modelling multiagent-based simulation systems, grounded on DOLCE.

Reaching beyond Simulations With regard to the medium-term goal to propagate multiagent-based implementation of logistic decision and control systems from the lab into real-world production systems, MABS is considered a suitable means to test multiagent applications for compliance with specifications at hand. Although multiagent-based applications are initially deployed in a simulation test bed in early stages of their product lifecycle, agent developers should be put in a position where they can focus on the production use case. We propose to augment multiagent-based simulation environments such that simulation-specific portions of the agent code bases are no longer required (Gehrke and Schuldt, 2009). This renders possible a *uniform agent design* suitable for both simulation and operation. The characteristics of the agents' target environment, either real or simulated, is kept transparent from the point of view of the agents. Working towards the uniform agent design ideal, the PlaSMA system has been extended to handle implicit simulation time synchronization (Schuldt et al., 2008).

Ontology Support for Design, Interaction & Analysis The field of ontology engineering was motivated in general by the promise of yielding scalable, portable and reusable domain models. Initial results reached by the research community, however, fell short of achieving these promises and turned out to yield more pain than gain. As a consequence, the ontology engineering *process* has been revised and put on more rigorous modelling principles such as the employment of foundational models, such as DOLCE (Masolo et al., 2003) or SUMO (Niles and Pease, 2001), the explication of the corresponding ontological commitments as well as the application of design- and content patterns (Gangemi and Presutti, 2008).

As discussed above, we employ simulations as a central method for examining the capabilities and limits of autonomous logistic processes. In order to enhance the PlaSMA ontologies described above, efforts are underway to put the corresponding models on firmer ontological grounds. For this purpose, we employ the DOLCE foundational ontology with several additional modules, namely *Descriptions and Situations* (Gangemi and Mika, 2003), the *Ontology of Information Objects*, the *Ontology of Plans* and the *Core Software Ontology* (Oberle et al., 2005). We currently develop a foundational ontology that models simulations themselves. It will facilitate the import of other ontological models that employ the same foundational top level and, consequently, further the adoption of PlaSMA for simulations in new application domains. Also, the subsequent use of domain models, initially developed specifically for use in PlaSMA, is rendered possible in new contexts. Next to enabling portability and re-usability, one can employ the resulting framework to explicate context-dependent reifications of the simulated entities and their actions, e.g. a logistic object, such as a truck, can be ontologically described as a *MeansOfTransportation*, a *FreightObject*, or a *TrafficObstacle* depending on the context at hand. In the end the resulting framework will provide a foundational domain-independent ontology of simulations - modelling the constituents of simulations *per se* - together with their respective docking stations for domain ontologies, which model the environment and entities, simulated or real.

CONCLUSION AND OUTLOOK

In this paper, we have introduced the multiagent-based simulation system PlaSMA with a particular focus on its ontology-based world model. We have described the benefits of the ontological grounding for simulation design and execution. We have shown, how the adoption of the PlaSMA system to new domains requires - in principle - only the creation and integration of a set of new domain ontologies. We hope to ease this engineering effort by employing a standardized and well-used foundational ontology and dedicated simulation-specific design patterns. Encouraged by successful adoption of PlaSMA for experimentation and demonstration in autonomous logistics, in the future, we will explore the potential advantages of using a foundational simulation ontology beyond increasing the scalability, portability and re-usability of the domain simulation models. We anticipate this generalization to affect also the interaction with agent-based simulations for the human stakeholders. Specifically, in the design phase scenario engineering will be facilitated by exploitation of reusable ontological patterns. At runtime, complex simulations could be rendered (more) accessible for human-computer interaction. Finally, in the analysis phase, an easier and well-grounded evaluation of recorded simulation runs will become feasible. We see usability as a central challenges for MABS of domains where we find a mixture of human and artificial decision makers, as one needs to guarantee that the human stakeholders involved will be able both to understand what is happening within such systems and control a running system.

REFERENCES

Bechhofer, S., van Hamelen, F., Hendler, J., et al. (2004). OWL Web Ontology Language Reference, W3C Recommendation. Technical report, W3C.

Bellifemine, F., Caire, G., and Greenwood, D. (2007). *Develop-*

ing Multi-agent Systems with JADE. Wiley Series in Agent Technologies. Wiley Inter-Science.

Bobeanu, C.-V., Kerckhoffs, E. J. H., and Landeghem, H. V. (2004). Modeling of Discrete Event Systems: A Holistic and Incremental Approach using Petri Nets. *ACM Trans. Model. Comput. Simul.*, 14(4):389–423.

Fujimoto, R. (2000). *Parallel and Distributed Simulation Systems*. Wiley & Sons, New York, NY, USA.

Gangemi, A. and Mika, P. (2003). Understanding the Semantic Web through Descriptions and Situations. In *Proceedings of the ODBASE Conference*, pages 689–706. Springer.

Gangemi, A. and Presutti, V. (2008). *Handbook of Ontologies (2nd edition)*, chapter Ontology Design Patterns, pages 221–244. Springer.

Gehrke, J. D. and Schuldt, A. (2009). Incorporating Knowledge about Interaction for Uniform Agent Design for Simulation and Operation. In *8th Int. Conference on Autonomous Agents and Multiagent Systems*, pages 1175–1176.

Gehrke, J. D., Schuldt, A., and Werner, S. (2008). Quality Criteria for Multiagent-Based Simulations with Conservative Synchronisation. In *13th ASIM Dedicated Conference on Simulation in Production and Logistics*, pages 545–554, Berlin, Germany. Fraunhofer IRB Verlag.

Gehrke, J. D. and Wojtusiak, J. (2008). Traffic Prediction for Agent Route Planning. In *International Conference on Computational Science 2008 (vol. 3)*, volume 5103 of *LNCS*, pages 692–701, Poland, Kraków. Springer.

Hülsmann, M. and Windt, K., editors (2007). *Understanding Autonomous Cooperation & Control in Logistics: The Impact on Management, Information and Communication and Material Flow*. Springer, Berlin, Germany.

Jefferson, D. R. (1990). Virtual Time II: Storage Management in Conservative and Optimistic Systems. In *Proceedings of the Ninth Annual ACM Symposium on Principles of Distributed Computing*, pages 75–89, Quebec, Canada. ACM Press.

Klügl, F., Oechslein, C., Puppe, F., and Dornhaus, A. (2002). Multi-Agent Modelling in Comparison to Standard Modelling. In *Artificial Intelligence, Simulation and Planning in High Autonomy Systems (AIS 2002)*, pages 105–110. SCS Publishing House.

Lees, M., Logan, R., Minson, T., Oguara, T., and Theodoropolus, G. (2004). Distributed Simulation of MAS. In *Multi-Agent Based Simulation, Joint Workshop*, volume 3415 of *LNCS*, pages 25–36. Springer.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., and Otramari, A. (2003). D18: Ontology Library (final). Project deliverable, WonderWeb: Ontology Infrastructure fo the Semantic Web.

Niles, I. and Pease, A. (2001). Towards a Standard Upper Ontology. In *Proceedings of the Int. Conference on Formal Ontology in Information Systems*, pages 2–9.

Oberle, D., Lamparter, S., Eberhart, A., and Staab, S. (2005). Semantic Management of Web Services. In *Service-Oriented Computing - ICSOC 2005*, volume 3826 of *LNCS*, pages 514–519. Springer.

Parunak, H. V. D., Savit, R., and Riolo, R. L. (1998). Agent-Based Modeling vs. Equation-Based Modeling: A Case Study and Users' Guide. In *Multi-Agent Systems and Agent-Based Simulation, First International Workshop*, volume 1534 of *LNCS*, pages 10–25, Paris, France. Springer.

Schuldt, A., Gehrke, J. D., and Werner, S. (2008). Designing a Simulation Middleware for FIPA Multiagent Systems. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 109–113, Sydney, Australia. IEEE Computer Society Press.

Schuldt, A. and Werner, S. (2007). Distributed Clustering of Autonomous Shipping Containers by Concept, Location, and Time. In *5th German Conference on Multiagent System Technologies*, volume 4687 of *LNAI*, pages 121–132, Leipzig, Germany. Springer.

Wojtusiak, J. (2009). The LEM3 System for Multitype Evolutionary Optimization. *Computing and Informatics (28)*, pages 225–236.

Xing, X., Warden, T., Nicolai, T., et al. (2009). SMIZE: A Spontaneous Ride-Sharing System for Individual Urban Transit. In *Multiagent System Technologies: 7th German Conference*, pages 165–176, Hamburg, Germany. Springer.

AUTHOR BIOGRAPHIES

Tobias Warden is a computer scientist and joined the artificial intelligence group at the University of Bremen as a research assistant in 2008. His research interests span distributed knowledge management and collaborative multi-agent learning.

Robert Porzel is a senior researcher at the Digital Media Research Group at the University of Bremen. His research encompasses knowledge representation, contextual computing, and natural language processing.

Jan D. Gehrke is a computer scientist and joined the artificial intelligence group at the University of Bremen as a research assistant in 2005. His research focuses on intelligent agents in logistics as well as knowledge representation and management in MAS.

Otthein Herzog is a professor emeritus. From 1993 to 2009, he held the chair on Artificial Intelligence in the Department of Mathematics and Computer Science at the University of Bremen. Dr. Herzog continues to contribute to the interdisciplinary activities of the CRC 637 which he represented as speaker till 2009.

Hagen Langer is a senior researcher at the Artificial Intelligence Research Group of the University of Bremen. Besides knowledge representation and reasoning, his research focus is on natural language processing.

Rainer Malaka holds the chair on Digital Media in the Department of Mathematics and Computer Science at the University of Bremen. He directs the TZI – Center for Computing and Communication Technologies.

INVESTIGATING ENTREPRENEURIAL STRATEGIES VIA SIMULATION

Martin Ihrig
The Sol C. Snider Entrepreneurial Research Center
The Wharton School of the University of Pennsylvania
424 Vance Hall, 3733 Spruce Street, Philadelphia, PA 19104

KEYWORDS
Agent-based simulation, strategic knowledge management, entrepreneurship, opportunity recognition.

ABSTRACT
This paper presents an agent-based simulation tool that enables researchers to study entrepreneurial strategies, their associated financial payoffs, and their knowledge creation potentials under different environmental conditions. Opportunity recognition processes can be analyzed in detail both on a micro- and macro-level.

INTRODUCTION

The academic field of entrepreneurship aims to develop theories that help us understand entrepreneurial *opportunities* and their formation (Alvarez & Barney, 2008, Alvarez & Barney, 2008). The concept of opportunity is central to the discussion of the entrepreneurial phenomenon (Shane & Venkataraman, 2000). Opportunities are emergent and one of a kind. Opportunity recognition – and the knowledge appropriation and development that underlie it – are complex and non-repeatable processes. This makes simulation modeling an appropriate methodological approach for researching opportunity recognition (Davis, Eisenhardt, & Bingham, 2007, Harrison, Lin, Carroll, & Carley, 2007). The unique simulation tool presented in this paper will help researchers model different entrepreneurial actions and the contexts in which they take place. It thereby allows to conduct innovative research that results in theory driven frameworks and hypotheses that are difficult to obtain from empirical analyses alone. Being able to explore distinct opportunity recognition strategies is the basis for advising entrepreneurs on their paths to success and governments on the right policy actions. Following a short description of the phenomenon under study, we describe the details of the simulation model we have built. We then conduct two virtual experiments that allow us to highlight the distinct simulation and modeling capabilities and to suggest theoretical insights for the field of entrepreneurship.

OPPORTUNITY RECOGNITION STRATEGIES

Reading the literature on opportunity recognition, one might get the impression that we are dealing with only one player: the classical hero of entrepreneurship, the innovator. However, we have good grounds for believing that there are also other valuable strategies for recognizing and realizing opportunities, especially those that involve imitative behavior (Aldrich & Martinez, 2001, Bygrave & Zacharakis, 2004). As the extant literature seems to conflate categories, we are in need of distinctions with a sound theoretical grounding.

Ihrig and zu Knyphausen-Aufseß (2009) put forward a model that differentiates between the *origination* of a new venture idea and its *development*. The Philosophy of Science literature encourages us that this distinction is meaningful. Hans Reichenbach (1938) differentiates between the *context of discovery* and the *context of justification*. The first is close to origination and the second to development. Karl Popper (1968) distinguishes "sharply between the process of conceiving a new idea, and the methods and results of examining it logically". Similarly, we argue that in a first step, entrepreneurs obtain their new venture ideas, and then, in a second step, further develop and refine them. Our theoretical distinction between origination and development lets us build the following two by two matrix (Figure1).

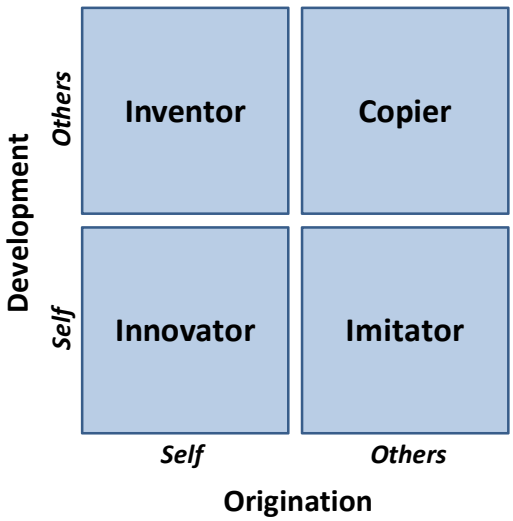


Figure 1: Four different entrepreneurial strategies

Entrepreneurs can either conceive a new venture idea by themselves or obtain the insight from somewhere else. Similarly, the development of a new venture idea can be done either independently or by drawing on others. Our model results in four different entrepreneurial roles or strategies: innovating, inventing, imitating, and copying. By means of simulation, we will be able to explore those strategies, to show that they can be meaningfully distinguished, and to point to theoretical and practical implications. Simulation modeling allows us to operationalize our theoretical concepts and the

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

processes behind it, and then to dynamically analyze micro and macro effects. In particular, we are interested in the comparative payoffs of each of the four entrepreneurial strategies in different environments. We expect to see distinctive knowledge progression and financial performance profiles. In addition, we will be able to study societal level effects that arise from competitive agent behavior.

USING *SIMISPACE2* TO MODEL THE OPPORTUNITY RECOGNITION PROCESS

We take a knowledge-based approach when studying how entrepreneurs obtain their new venture ideas and develop them (Ihrig, zu Knyphausen-Aufseß, & O'Gorman, 2006). Based on Austrian economics (Kirzner, 1997), we consider *knowledge*, and in particular its appropriation, development and exploitation, as the basis for new venture creation. *SimISpace2* is an agent-based graphical simulation environment designed to simulate strategic knowledge management processes, in particular knowledge flows and knowledge-based agent interactions. The simulation engine's conceptual foundation is provided by Boisot's (1995, 1998) work on the Information Space or *I-Space*. In what follows, we describe how we have used *SimISpace2* to build an application-specific model – *SimOpp* – that enables us to study opportunity recognition strategies and their outcomes under different environmental conditions.

Basic workings of the *SimISpace2* environment

Ihrig and Abrahams (2007) offer a comprehensive description of the entire *SimISpace2* environment and its technical details. For this particular research project, we only use a limited set of the features the program offers. Below, we briefly review some of the underlying principles of the simulation and explain how knowledge is represented.

Two major forms of entities can be modeled with *SimISpace2*: agents and knowledge items/assets. When setting up the simulation, the user defines agent groups and knowledge groups with distinct properties. Individually definable distributions can be assigned to each property of each group (uniform, normal, triangular, exponential, or constant distribution). The simulation then assigns the individual group members (agents and knowledge items) characteristics in accordance with the distribution specified for the corresponding property for the group of which they are a member.

Knowledge in the simulation environment is defined as a 'global proposition'. The basic entities are *knowledge items*. Based on the *knowledge group* they belong to, those knowledge items have certain characteristics. All knowledge items together make up the *knowledge ocean* – a global pool of knowledge. Agents can access the knowledge ocean, pick up knowledge items, and deposit them in knowledge stores through the *scanning* action. A knowledge store is an agent's personal storage place for a knowledge item. Each knowledge store is local to an agent, i.e. possessed by a single agent. As containers, knowledge stores *hold* knowledge items as their contents. Stores and their items together constitute *knowledge assets*. Examples of knowledge stores include books, files, tools, diskettes, and sections of a person's brain. There is only one knowledge item per knowledge store, i.e. each knowledge item that an agent possesses has its own knowledge store. If an agent gets a new knowledge item (whether directly from the knowledge ocean or from other agents' knowledge stores), a new knowledge store for that item is generated to hold it.

The concept of a knowledge item has been separated from the concept of a knowledge store to render knowledge traceable. If knowledge items are drawn from a common pool and stored in the knowledge stores of different agents, it becomes possible to see when two (or more) agents possess the same knowledge, a useful property for tracking the diffusion of knowledge.

The separation between a global pool of knowledge items and local knowledge stores is particularly important when it comes to codification and abstraction (these only apply to knowledge stores, not to knowledge items). Knowledge items are held by multiple agents, and one agent's investment in codification or abstraction does not influence the codification and abstraction level of the same knowledge item held by another agent. Agents possess knowledge stores at a particular level of codification and abstraction. If the agent codifies its knowledge and makes it more abstract, the properties of the knowledge item itself – i.e., its *content* – are not changed, but it gets moved to a new knowledge store with higher degrees of codification and abstraction – i.e., its *form* changes.

SimISpace2 also features a special kind of knowledge. A DTI (knowledge *Discovered Through Investment*) is a composite knowledge item that is discovered by integrating the knowledge items that make it up into a coherent pattern. DTIs cannot be discovered through scanning from the global pool of knowledge items. The user determines knowledge items to act as the constituent components of a DTI. The only way for an agent to discover a DTI is to successfully scan and appropriate its constituent components and then to codify and/or abstract them beyond user-specified threshold values in order to achieve the required level of integration. Once these values are reached, the agent automatically obtains the DTI (the discover occurrence is triggered in the simulation). Investing in its constituent components – i.e. scanning, codifying and abstracting them – is the primary means of discovering a DTI. By specifying the values of different DTIs, the user can indirectly determine the values of the networks of knowledge items that produce DTIs. Such networks represent more complex forms of knowledge. Once an agent has discovered a DTI item, it is treated like a regular knowledge item, i.e. other agents are then able to scan it from the agent that possesses it.

Important *SimISpace2* processes and occurrence types used for the modeling

To keep our model and the resulting analyses simple, we only use six of twenty actions that the *SimISpace2* environment features, which will be explained below: relocate, scan, codify, discover, learn, exploit. By conducting those actions, our virtual agents try each period to accumulate new knowledge and develop it, and to discover DTIs. Agents increase their financial funds by capitalizing on the knowledge they possess, in particular DTIs. The financial funds property of agents measures entrepreneurial success. The better the knowledge appropriation, development and exploitation strategy, the higher the funds will be. Agents with financial funds of zero die. Based on different agent group behaviors, the increase of the agents' individual financial funds and the increase of their stock of knowledge occur at different rates. Whereas all agents in our simulation will try to *learn* and *exploit* their knowledge (and thereby to grow their financial funds), agents will differ in their approaches to obtaining and developing knowledge. What follows is a concise review of the critical actions we assign for modeling knowledge appropriation and development.

Scanning. An agent can scan for knowledge, randomly picking up knowledge items, either from the knowledge ocean or from other agents' knowledge stores. The probability of picking from the knowledge ocean (vs. from other agents) can be specified on the agent-group-level. While an agent can scan any knowledge item in the knowledge ocean, it can only scan knowledge items in those knowledge stores (of other agents) that fall within its *vision*. *SimISpace2* agents populate a physical, two-dimensional space (called *SimWorld*), and the vision property determines how far the agent can see within a certain spatial radius from its current location. A knowledge item that is successfully scanned is placed in a new knowledge store possessed by the agent. The new knowledge store picks up the level of codification and abstraction either from the knowledge group that the knowledge item belongs to in the knowledge ocean or from the knowledge store where the agent found the item. Agents will only try to scan knowledge items that they do not already possess at that level of codification and abstraction. The ease with which a knowledge item is scanned from another agent's knowledge store is some positive function of the store's degree of codification and abstraction. Knowledge items in knowledge stores with higher codification and abstraction have a higher probability of being scanned. In the case of the knowledge ocean, ease of scanning is determined by the nature of the network a knowledge item is embedded in. Finally, recall that once an agent has scanned all the components that constitute a given DTI and has codified and abstracted them up beyond a certain threshold, it automatically obtains the DTI (*discover* action is triggered).

Relocating. An agent can relocate within a certain distance of its position in the 100 by 100 grid of the *SimWorld*. Relocating implies moving either closer to or further away from other agents or knowledge stores. The distance an agent moves per relocation depends on the *distance* setting for the relocate action of its agent group. Relocation is relevant to scanning as it affects what knowledge stores and other agents an individual can see. As agents can only scan within the radius of their vision, they are only able to pick up knowledge in a different area by moving. When agents relocate, they leave their knowledge stores behind in the original location, but still retain access to them. (N.B.: When a new knowledge store is created, it is always assigned the same location as the agent that possesses it.)

Codifying and Abstracting. An agent can create new knowledge stores at different levels of codification and abstraction with values ranging from 0 to 1. Every new store represents a per period carrying gain or cost for the agent that is added/deducted at the end of each period from the agent's financial funds. Codification and abstraction are separate actions that affect the knowledge stores (form) in which a given knowledge item (content) is held. The agent must first possess a knowledge item in a store before it can perform these actions. The levels of codification or abstraction of a newly created knowledge store are increased incrementally beyond those of existing stores. The knowledge item in the new knowledge store remains always the same; only the level of codification and abstraction of the knowledge store changes. Stores with higher levels of codification and abstraction are more likely to be scanned from and are more valuable when exploited. However, remember that the more diffused knowledge gets the less value the agent can extract from it. In *SimOpp*, we only use the codification action to model the knowledge structuring process.

Learning. Before a knowledge item can be exploited, it has to register with an agent through a learning process. This can only apply to a knowledge item that an agent possesses. Its chances of success increase with the levels of codification of the knowledge store that holds it.

Exploiting. Agents can capitalize on their knowledge, i.e. generate value for themselves. An agent can only exploit knowledge that it has registered and internalized through learning. The financial funds of an exploiting agent are increased by the value of the different exploiting actions that it undertakes. The *exploit amount* is calculated based on the user-set *base value* of the knowledge item involved. This is done according to the level of codification and abstraction of the knowledge store holding the knowledge item, and to the level of diffusion of the knowledge item (percentage of agents that possess the particular piece of knowledge in a period). The user can define an industry-specific table of revenue multipliers based on abstraction and codification levels. In the *I-Space*, the value of knowledge is some positive function of both its utility (the level of codification and abstraction) and its scarcity (the level of diffusion). Therefore, typically, the higher the levels of codification and abstraction, the higher the revenue multiplier, i.e. more codified and

abstract knowledge is worth more. More codified and abstract knowledge, however, is also more likely to be diffused, thus eroding the value of knowledge. Furthermore, the calculations allows for the effects of obsolescence, which, like diffusion, also erodes value: obsolete knowledge is worthless. Whereas revenue multipliers depend on the characteristics of a knowledge *store* (its level of codification and abstraction), obsolescence solely depends on the properties of the knowledge *item* the store contains.

Basic parameterization and set-up: the *SimOpp* model

We can now describe the *SimISpace2* model designed and built for the opportunity recognition context, *SimOpp*, and present the properties of the participating agent and knowledge groups.

Agents

In line with the framework we have developed, we create four *agent groups*. The following matrix (Figure 2) shows the four agent groups and the relevant *SimISpace2* actions that distinguish them from one another.

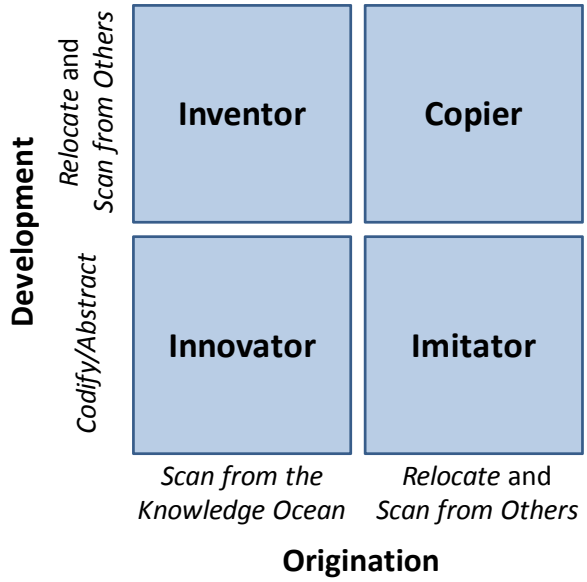


Figure 2: The four entrepreneurial groups in *SimOpp*

The probabilities to chose and perform particular actions vary from group to group based on the conceptual distinctions we have made. In total, agents engage in four activities. There is one activity assigned for implementing *origination* (either *self* or *others*) and one activity assigned for implementing *development* (either *self* or *others*). When it comes to *origination*, the activity of obtaining insights from a third party, as opposed to coming up with it oneself, is implemented with the 'Scan from Others' and the 'Relocate' actions. Agents can move through the *SimWorld* and scan knowledge assets from other agents. The activity of arriving at an insight oneself is implemented with the 'Scan from the Knowledge Ocean' action. When it comes to *development*, the activity of developing by yourself is implemented with the 'Codify' action. The activity of developing by drawing on others is again implemented with the 'Relocate' and 'Scan from Others' actions. To be able to capitalize on their knowledge, all agents can perform the 'Learn' and the 'Exploit' action (activity three and four respectively). Note that the numbering of the activities should not necessarily imply a particular order in which the actions are conducted in the simulation. Knowledge can only be learned once it has been obtained and can only be exploited once it has been learned. However, a random draw each period based on the distributions assigned for the propensities to engage in an action determines which of the possible actions an agent chooses. Looking at each agent group in turn, we can see what actions and properties agents have in common and, based on the description above, what distinguishes them (constant distributions assigned for propensities to engage in a particular action in brackets).

Innovator. Innovators perform four actions; they *scan* (1) and they *codify* (1), and as all the other groups, they *learn* (1) and *exploit* (1). When it comes to scanning though, they can only scan from the knowledge ocean.

Imitator. Imitators can perform five actions; they *scan* (0.5), *relocate* (0.5), *codify* (1), *learn* (1) and *exploit* (1). In contrast to the Innovators, Imitators only scan from the agents surrounding them; they do not scan from the knowledge ocean.

Inventor. Inventors do not codify, they only *scan* (1.5), *relocate* (0.5), *learn* (1) and *exploit* (1). They can scan from both the knowledge ocean and from other agents.

Copier. Copiers also do not codify and only *scan* (1), *relocate* (1), *learn* (1) and *exploit* (1). They only scan from other agents and not from the knowledge ocean.

There are ten agents in each agent group. All agents start with financial funds of 100. The *relocate distance* and *vision* property are the same for all groups, but they change with each scenario (case) under study. Imitators, Inventors, and Copiers are randomly spread out in the *SimWorld* (uniform distribution 0-100 for x and y location); Innovators are clustered together at the center (uniform distribution 45-55 for x and y location) as can be seen in Figure 3 further down.

Knowledge

We use both basic knowledge and the higher-level DTI knowledge in *SimOpp*. We have three distinct basic knowledge groups: *Local*, *Entrepreneurial*, and *New Venture Idea Knowledge*.

Local Knowledge. Local Knowledge represents an understanding of the local market and its culture. It starts at a high level of codification and abstraction (0.7) and has a base value of 5. Remember that the intrinsic *base value* of a knowledge item is the starting point for calculating the *exploit amount*, which represents the increase in financial funds after an exploit action has been performed on a knowledge asset. As in the *I-Space* the value of knowledge is some positive function of its utility and its scarcity, both the level of codification and

abstraction and the level of diffusion are included in the calculation.

Entrepreneurial Knowledge Entrepreneurial Knowledge represents the 'Know-How' (Ryle, 1949). Abilities like to "sell, bargain, lead, plan, make decisions, solve problems, organize a firm and communicate" (Shane, 2003: 94) are examples for knowledge items in this group. To this we would add skills like writing business plans, initiating sales, creating initial products and services, securing initial stakeholders and finances – creating the initial transactions set (Venkataraman & Van de Ven, 1998). Knowledge from this knowledge group starts at a medium level of codification and abstraction (0.5) and has a base value of 10.

New Venture Idea Knowledge. New Venture Idea Knowledge represents the 'Know-What'. Knowledge items in this group are insights about a particular potential service or product offering. Knowledge from this knowledge group starts at a low level of codification and abstraction (0.3) and has a base value of 20.

There are ten knowledge items in each knowledge group. All basic knowledge groups have an *obsolescence rate* of zero, a *codification* and *abstraction increment* of 0.1, and no *per period carrying gain or cost*. All agent groups are endowed with Local Knowledge and Entrepreneurial Knowledge, but they do not possess New Venture Idea Knowledge.

Opportunities

We use DTI knowledge to model opportunities. Once an agent possesses a knowledge item each from the Local Knowledge, Entrepreneurial Knowledge and New Venture Idea Knowledge groups, in knowledge stores with a codification level that is equal or greater than 0.6, it obtains the corresponding DTI, i.e. the agent 'discovers' an opportunity. There are ten DTIs, each being based on a combination of the nth item of each basic knowledge group (e.g., the underlying knowledge items for DTI 1, are knowledge item 1 of Local Knowledge, knowledge item 1 of Entrepreneurial Knowledge, and knowledge item 1 of New Venture Idea Knowledge). DTI knowledge items have a high *starting level of codification* and *abstraction* (0.8), a high (compared to base knowledge) *base value* of 2500, an *obsolescence rate* of zero, a *codification* and *abstraction increment* of 0.1, and no *per period carrying gain or cost*.

Agents obtain opportunities in different ways. Based on the dynamics of the simulation, we can identify the following three.

Opportunity Construction. The classical way is to construct an opportunity. An agent obtains all underlying knowledge items, structures them up to the specified codification threshold, and is then rewarded by obtaining the DTI, i.e. the opportunity (the *discovery* occurrence in the simulation is triggered). As prior stock of knowledge, agents already possess Local Knowledge and Entrepreneurial Knowledge from period one on. Agents can obtain the missing New Venture Idea Knowledge item either directly from the knowledge ocean (*self*) or from a knowledge store of somebody else (*others*). Agents can then reach the required threshold by either codifying the knowledge themselves or by scanning it from another agent (or, in the case of the imitator, by a combination of both).

Opportunity Acquisition. Not only can agents scan from others basic knowledge items, but also can they scan DTIs. This means, in addition to constructing opportunities themselves, agents are able to directly acquire the knowledge about an opportunity by scanning from a knowledge store of another agent that carries a DTI.

Opportunity Amplification. Agents can also further develop and structure their opportunities. They do so, either by codifying their DTIs directly or by scanning from other agents that possess higher codified stores of that DTI.

PUTTING *SIMOPP* TO USE: TWO VIRTUAL EXPERIMENTS

To highlight the distinct modeling capabilities of our simulation tool, we now conduct two virtual experiments. Each scenario is run 60 times, and each run has 1000 periods. All graphs show the average across all runs and some display the standard deviation to indicate the significant difference between the lines. Virtual Experiment 1 models an environment with low access to competitors' knowledge, Virtual Experiment 2 one with moderate access to competitors' knowledge.

Virtual Experiment 1: Access to competitors' knowledge is low

We model low access to competitors' knowledge by setting the vision property and the relocate distance to five (out of 100). Agents can – within a limited radius – see other agents and can move away from their original positions – in little steps – through the 100 by 100 grid of the *SimWorld*. As an example, Figure 3 depicts the area (black) that one (random) agent covers throughout the 1000 periods.

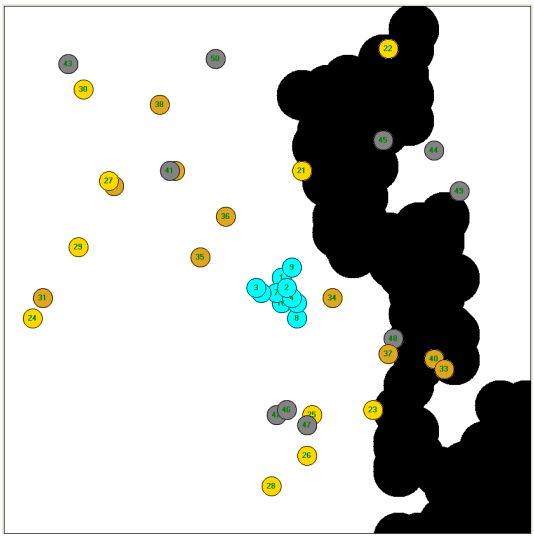


Figure 3: *SimWorld* Report Period 1000

We expect all agent groups to obtain at least some DTIs, because knowledge scanning is possible, even if limited. However, we cannot predict the specific effects this will have on the financial funds. Based on the distributions we have assigned to specify the properties of the four agent groups, we know the different general knowledge appropriation and development behaviors. How those different behaviors or strategies will play out in a population of agents in a knowledge environment, we do not know. We need the simulation to dynamically model the complex relationships among knowledge and agents across time to see how successful or not the different agent groups are in terms of growing their financial funds and knowledge portfolios. Figure 4 shows us the financial profiles for each group.

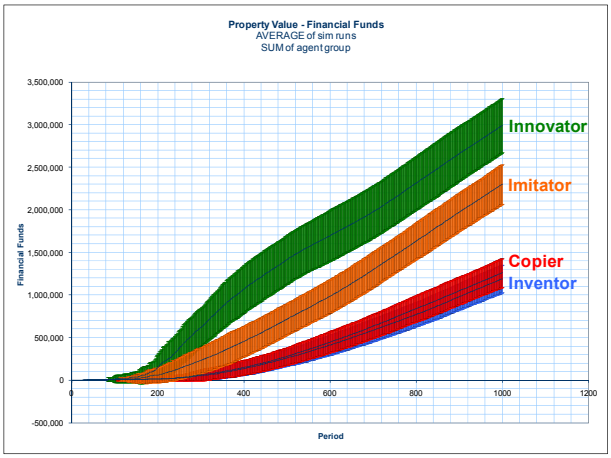


Figure 4: Longitudinal Report Graph Financial Funds

Based on distinct opportunity recognition strategies of the four entrepreneurial types, we can clearly distinguish four different groups. With vision and relocate distance set to five, Innovators perform better than Imitators, and they both outperform Copiers and Inventors, whose financial profiles overlap.

Insight 1a: The financial performance and performance volatility of the four entrepreneurial strategies – innovating, inventing, imitating, copying – will have distinct profiles, as a result of the differences in their knowledge appropriation and knowledge development behaviors.

Compared to a scenario where there is no access to competitors' knowledge (vision and relocate distance set to zero, not illustrated in this paper), we see that the Innovator group loses financial funds, which are picked up by the other three groups we proposed in our conceptual model. This is an example of the inner workings of the simulation. In both scenarios, Innovators follow the same pattern of actions (in particular, the number of exploit actions does not change), but their financial funds in the scenario without access to competitors' knowledge are almost three times higher than in the scenario with low access, demonstrating that rents are competed away. It is a result of what we call the 'diffusion discount effect'. In the former scenario, fewer agents obtain the DTIs and therefore, the ones that do secure them get higher rents. As more diffused knowledge is worth less, innovators in the latter scenario earn less (rent dissipation), because rent is appropriated by other types of entrepreneurs. All of this highlights distinct simulation and modeling capabilities of *SimOpp*, which can be summarized as follows.

Simulation & Modeling Capability 1: *SimOpp* enables researchers to simulate the opportunity recognition process and its respective financial payoffs for different entrepreneurial strategies.

Simulation & Modeling Capability 2: *SimOpp* allows one to simulate competitive agent behavior resulting from different entrepreneurial strategies.

Looking at the relative performance of the four agent groups and their distinctive entrepreneurial strategies, we can sum up the results as follows, giving tentative examples in the context of economic history. These historical stereotypical facts can be seen as a kind of face validation, showing that "the critical characteristics of the process being explored are adequately modeled and that the model is able to generate results that capture known canonical behavior" (Carley, 2002: 262). A pure R&D game (Inventor), originating but not being able to further develop, is not enough. Companies in the UK, for example, have had a long and good track record in invention that goes back to the industrial revolution. However, in recent decades they have had a relatively poor record in turning inventions into commercial products – i.e., in further development. Similarly, a copycat approach (Copier), of not originating and simply copying but not developing, is not sufficient to outperform other strategies. Until recently, many Chinese companies have behaved like this, producing identical copies of foreign products without adding anything to them (copying products without adding ingenuity). Most still do this. Interestingly, both strategies – copying and inventing – result in about the same financial payoffs. Combining pure and applied R&D (Innovator), being able to do both originating and developing, is the most rewarding strategy. Good examples for this are US firms that come up with innovative technologies and know how to develop them to their fullest market potential. Creative imitation (Imitator), by not originating but copying with further development, also represents a highly rewarding path to success. A case in point here is 1970's Japan with companies generating high revenues and profits by the improvement (particularly in production processes) of foreign inventions.

For the development of entrepreneurship theory, an important observation to make is that there is a significant difference between the Copier and the Imitator. Biology makes the distinction between replication (Copier) and reproduction (Imitator). Sexual *reproduction* is the biological process by which new individual organisms are produced through a combination of genetic material contributed from two (usually) different members of a species, thereby giving rise to variety and ultimately to evolution. In contrast,

replication is the biological process resulting in an exact duplicate of the parent organism (Dawkins, 1976).
Extant entrepreneurship theory has neglected this distinction and thereby largely ignored a viable opportunity recognition strategy. For other authors, copying, imitating, emulating are seen as the same behavior. They miss the crucial difference between appropriating a business idea and implementing it one-to-one (copying), and being inspired by certain preexisting ideas and further developing them (creative imitation). The simulation results show that there is a meaningful distinction, based on both financial performance and knowledge acquisition trajectories.

Insight 1b: In terms of their financial profiles, Innovators will differ from Imitators, and they both will be different from Copiers and Inventors.

Insight 2: In particular, imitation and copying are not the same. The different opportunity recognition strategies of Imitators and Copiers will result in fundamentally different performance profiles. Financially, Imitators will consistently outperform Copiers.

Opportunity Construction, Acquisition, and Amplification

The simulation allows us to look behind the financial funds profiles and to explore the accumulation of DTIs (opportunities). Figure 5 shows the paths each group follows to obtain the ten DTIs (as there are ten agents in each group, the maximum is 100). The graph features S-shaped curves that are characteristic for diffusion processes (Mahajan & Peterson, 1985). Note that research on diffusion of innovations (Rogers, 2003) looks at the process of an innovation's adoption and implementation by a *user*. The focus is on the market for and of an innovation. In contrast, we are looking at the knowledge dynamics that help *entrepreneurial agents* construct and attain opportunities. Looking at Figure 5, it is the initial 600 periods that are most interesting to interpret.

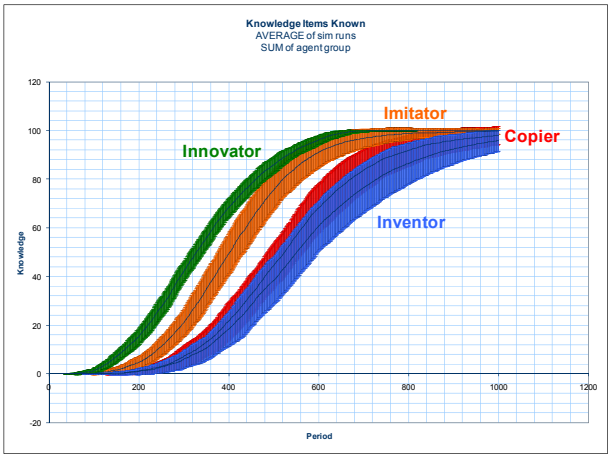


Figure 5: Longitudinal Report Graph DTIs Obtained

The DTIs, with which we model opportunities, have a starting level of codification of 0.8. Remember, that the maximum level of codification is 1 and that the codification increment is set to 0.1. This means that there are up to three knowledge stores per DTI per agent obtainable (with codification levels of 0.8, 0.9, 1.0), or 300 per agent group (Figure 6). As these DTI knowledge stores are the basis for the exploit action, the more stores an agent possesses the better.

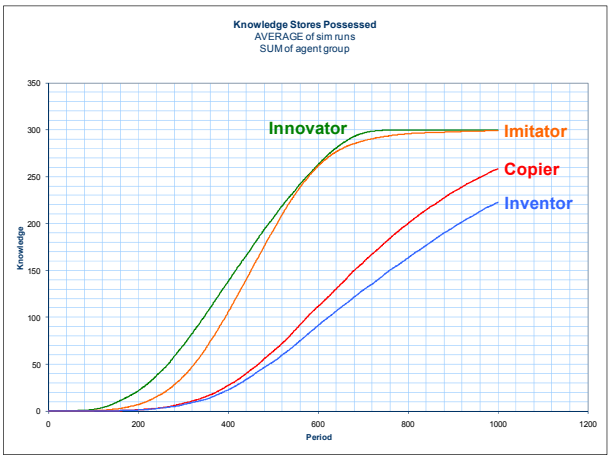


Figure 6: Longitudinal Report Graph DTIs *Knowledge Stores*

Three different occurrences let agents attain DTI knowledge stores. This means, Figure 6 is explained by three distinct actions or ways with which the entrepreneur obtains and develops opportunities. Figure 7 shows the first one, the 'discovery' occurrence.

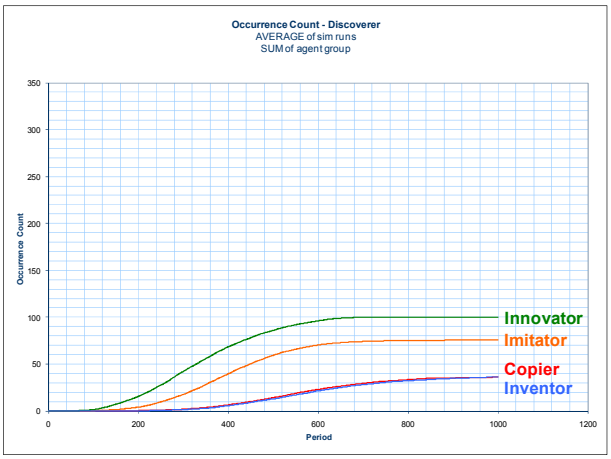


Figure 7: Longitudinal Report Graph DTIs *Discoverer*

The graph depicts the outcome of the *Opportunity Construction* process described earlier. For a 'discovery' to happen, the underlying basic knowledge items have to be assembled and brought to the required threshold. We see that Innovators lead the process with Imitators in second place. The 'discovery' occurrence represents the classical entrepreneurial route of constructing one's opportunity step by step with the help of one's idiosyncratic stock of knowledge (Shane, 2000). The micro processes behind it are as follows. The missing basic knowledge group – new venture idea knowledge – can be obtained directly (innovator, inventor – 'scan from the knowledge ocean') or from someone else (imitator, copier – 'scan from others').

Bringing all the knowledge up to the required threshold can also either be done directly (innovators, imitators – 'codify') or by getting knowledge stores with higher codification levels from someone else (copiers, inventors – 'scan from others').

Doing everything directly, i.e. getting the idea and developing it yourself, the Innovator group comes first at opportunity construction. A combination of obtaining the insights from other agents but developing them oneself helps the Imitator group come second. The other groups lag behind in terms of being able to construct their opportunities. Those groups, however, come first in the next graph that co-explains the total number of DTI knowledge stores: DTI scanning occurrences (Figure 8).

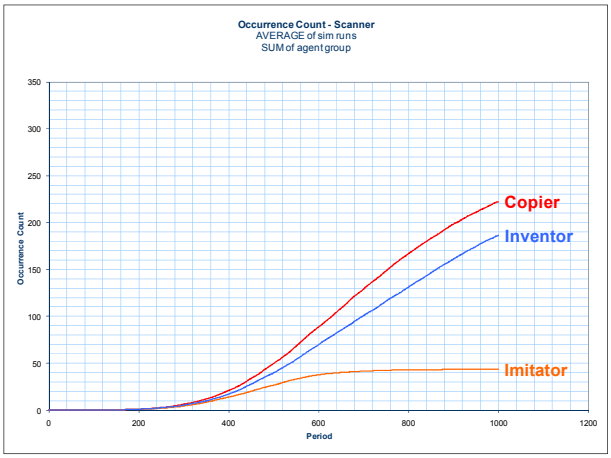


Figure 8: Longitudinal Report Graph DTIs *Scanner*

This graph shows what we described earlier as *Opportunity Acquisition*. The agents scan DTI knowledge stores from other agents. By doing this, they are not only able to obtain the basic knowledge about an opportunity, but also the further developed knowledge items – DTI knowledge stores with higher levels of codification. The 'production' of those can be observed in the last graph that explains the number of DTI knowledge stores: DTI codification occurrences (Figure 9).

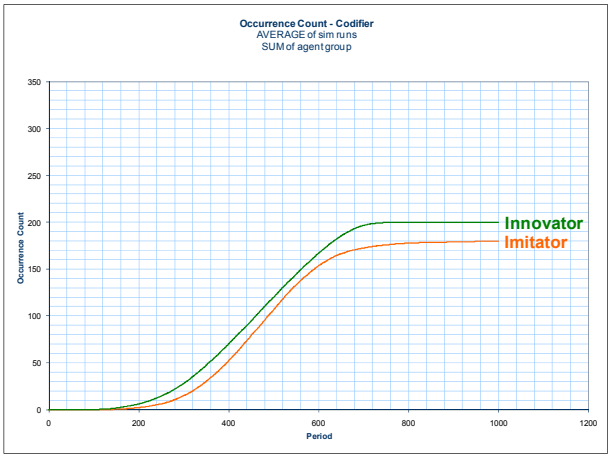


Figure 9: Longitudinal Report Graph DTIs *Codifier*

Figure 9 shows what we have called earlier *Opportunity Amplification*. The Innovator group and the Imitator group are able to further structure their knowledge about opportunities. Again, Innovators lead, building on their first mover advantage of having started to 'discover' DTIs before everybody else; Imitators follow very closely.

Opportunity Construction, *Opportunity Acquisition*, and *Opportunity Amplification* are distinct processes that, taken together, also with the actions behind them, give a more complete and fine-grained picture of what opportunity recognition means. The simulation model enables us to operationalize different entrepreneurial actions and processes, and to later analyze the effects they have on financial performance and knowledge built up. *SimOpp*'s distinct features can be summarized as follows.

Simulation & Modeling Capability 3: *SimOpp* allows one to dissect the opportunity recognition process and to arrive at a more discriminating picture of entrepreneurial strategies. In particular, it lets us distinguish between *Opportunity Construction*, *Opportunity Acquisition*, and *Opportunity Amplification.*

Virtual Experiment 2: Access to competitors' knowledge is moderate

In the previous scenario, we set the vision property and the relocate distance to 5. Five out of 100 is quite a small increment and the question arises what will happen if we increase both vision and relocate distance by another 5. A vision property and relocate distance of 10 is still far from universal sight and hence from total and immediate access to all the knowledge of other agents. Figure 10 shows us the financial profile of all four groups with vision and relocate distance set to 10 – moderate access to competitors' knowledge.

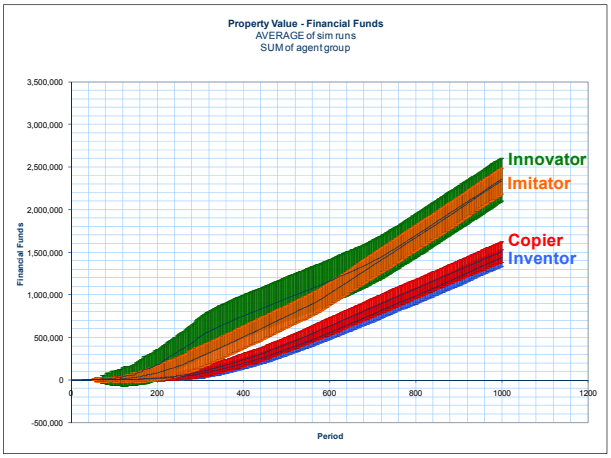


Figure 10: Longitudinal Report Graph Financial Funds

Total output – GDP – is the same as in the previous scenario (total financial funds). Note that as a basic measure of the population's performance, we treat the total financial funds per period (the sum of all agent groups) as a proxy for GDP. Conceptually, it can be viewed as the market value of all final goods and

services made within the *SimWorld* based on the exploitation of the agents' knowledge assets. The financial funds of the Innovator group drop (same amount of exploit actions, but diffusion discount effect), the funds of the Copier and Inventor groups increase, and the funds of the Imitator group stay approximately the same. An intriguing outcome is that now the Innovator group does not do any better than the Imitator group. Imitators can capitalize on their knowledge about opportunities as well as Innovators do, and with a lower standard deviation, meaning with lower instability. This means, not only has our experiment revealed that there is a significant difference between Copiers and Imitators (Insight 1b), but also can it prove that under certain conditions imitating behavior can be as profitable as innovating behavior.

Insight 3: In environments where there is moderate access by other entrepreneurial agents to Innovators and their knowledge, there will be no distinguishable difference in the financial performance of Imitators compared to Innovators. So when access by Imitators to Innovators' knowledge is moderate or greater than moderate, creative imitation will be an equally rewarding alternative to innovation.

Macro effects

In addition to doing analyses on a more micro level – looking at individual agents and agent group behavior – we can also explore effects on a macro level. This is a distinct advantage of simulation methods. Not only are we able to assess the financial performance of different groups of entrepreneurial agents, but also can we see what the societal (population) effects of the agents' actions are.

As we have noted above, GDP stays the same and the financial performance of Innovators is lower after the general increase in vision and relocate distance. Why should society find such a scenario beneficial? One thing society is interested in is the 'outgrowth of entrepreneurial opportunities', products or services that are new to the market. The earlier that entrepreneurs perceive and exploit opportunities, the earlier new products, services, or processes – in short, innovations – reach the economy. So the question is whether – apart from maybe being more realistic – the increase in vision and relocate distance is worth it for society in terms of knowledge diffusion. Our simulation results give evidence that it indeed is valuable! Society gains, because as Figure 11 shows, opportunities and the new products or services that come with it are obtained much earlier than before.

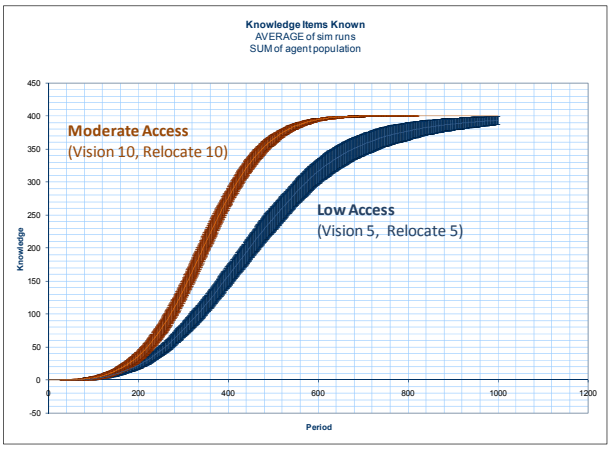


Figure 11: Total Number of DTIs Known (population level) for low access and moderate access scenarios

Looking at the sum of agent groups, this result might have been expected: if the agents have greater access to the knowledge that is out there, the diffusion curve shifts to the left. What is interesting, however, is that the lead group – the group whose agents first obtain *all* DTIs (100%) – changes from Innovators in the low access scenario to Imitators in the moderate access scenario, and they do so *200 periods earlier* as can be seen in Figure 12 (note the two markers, indicating when the respective agent group fully reaches the 100 DTI-knowledge item threshold).

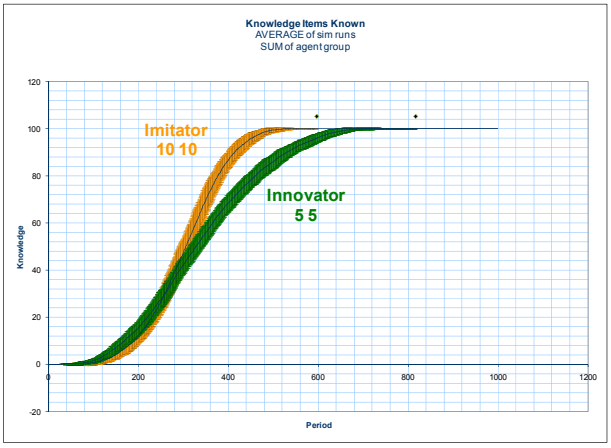


Figure 12: Number of DTIs Known for lead groups in low access and moderate access scenarios

This means that innovations, which are the basis for societal wealth generation, will actually be 'brought to market' earlier, by the creative imitator, not by the classical innovator. This is the foundation of growth and continuing wealth for both a society as a whole and future entrepreneurs. In a virtuous cycle, those entrepreneurs can build on the new insights obtained to discover new sets of opportunities through continuous Schumpeterian-Learning of all participants (Boisot, 1998).

Insight 4: Moderate access to innovators and their knowledge will be accompanied by acceleration in the introduction of innovations, which will thereby also accelerate societal welfare effects.

What does moderate access to innovators mean? It certainly depends on the particular context. Our results can for example be applied to the discussion on property rights (Boisot, MacMillan, & Han, 2007). They do not suggest to abolish property rights, but to establish a system that allows entrepreneurs of different kind to co-evolve opportunities. As we have seen above, society and entrepreneurs will benefit from faster knowledge creation and development based on continuous social learning (Boisot, 1998).

CONCLUSION

In this paper, we have explained how to parameterize and use *SimISpace2* to develop a unique simulation model that lets entrepreneurship researchers study the opportunity recognition process under different environmental conditions. It helps them get a deeper understanding of the nature of opportunities both on a micro and macro level. The practical outcomes are strategy recommendations for both individual entrepreneurs and policy makers. In future research, additional virtual experiments can be conducted that exploit the full range of *SimISpace2*'s parameters.

ACKNOWLEDGEMENTS

This research was supported by the Snider Entrepreneurial Research Center at the Wharton School of the University of Pennsylvania and by I-Space Institute, LLC. The author would like to thank Ian MacMillan and Max Boisot for their extremely helpful and important insights into the paper and Ryan Namdar and Danish Munir for their excellent programming work on *SimISpace2*.

REFERENCES

Aldrich, Howard E. & Martha Argelia Martinez. 2001. Many are called, but few are chosen: an evolutionary perspective for the study of entrepreneurship. *Entrepreneurship Theory and Practice*, 25(4): 41-56.

Alvarez, Sharon A. & Jay B. Barney. 2008. Opportunities, organizations, and entrepreneurship. *Strategic Entrepreneurship Journal*, 2(4): 265-67.

Alvarez, Sharon A. & Jay B. Barney. 2008. Opportunities, organizations, and entrepreneurship *Strategic Entrepreneurship Journal*, 2(3): 171-73.

Boisot, M. H. 1995. *Information space: a framework for learning in organizations, institutions and culture* London: Routledge.

Boisot, M. H. 1998. *Knowledge assets - securing competitive advantage in the information economy*. New York: Oxford University Press.

Boisot, Max, Ian C. MacMillan, & Kyeong Seok Han. 2007. Property rights and information flows: a simulation approach. *Journal of Evolutionary Economics*, 17(1): 63-93.

Bygrave, William D. & Andrew Zacharakis. 2004. *The portable MBA in entrepreneurship* Hoboken, New Jersey: Wiley.

Carley, Kathleen M. 2002. Computational organizational science and organizational engineering. *Simulation Modelling Practice and Theory*, 10(5-7): 253-69.

Davis, Jason P., Kathleen M. Eisenhardt, & Christopher B. Bingham. 2007. Developing theory through simulation methods. *Academy of Management Review*, 32(2): 480-99.

Dawkins, Richard. 1976. *The selfish gene*. Oxford: Oxford University Press.

Harrison, J. Richard, Zhiang Lin, Glenn R. Carroll, & Kathleen M. Carley. 2007. Simulation modeling in organizational and management research. *Academy of Management Review*, 32(4): 1229-45.

Ihrig, Martin & A.S. Abrahams. 2007. Breaking new ground in simulating knowledge management processes: SimISpace2. Paper presented at 21st European Conference on Modelling and Simulation (ECMS 2007), Prague.

Ihrig, Martin & Dodo zu Knyphausen-Aufseß. 2009. Discovering international imitative entrepreneurship: Towards a new model of international opportunity recognition and realization. *Zeitschrift für Betriebswirtschaft*, (Special Issue 1/2009).

Ihrig, Martin, Dodo zu Knyphausen-Aufseß, & Colm O'Gorman. 2006. The knowledge-based approach to entrepreneurship: linking the entrepreneurial process to the dynamic evolution of knowledge. *Int. J. Knowledge Management Studies*, 1(1/2): 38-58.

Kirzner, Israel M. 1997. Entrepreneurial discovery and the competitive market process: an Austrian approach. *Journal of Economic Literature*, 35(1): 60-85.

Mahajan, Vijay & Robert A. Peterson. 1985. *Models for innovation diffusion*. Beverly Hills, CA: Sage Publications.

Popper, Karl. 1968. *The logic of scientific discovery*. London: Hutchinson & Co

Reichenbach, Hans. 1938. *Experience and prediction*. Chicago: University of Chicago Press.

Rogers, Everett M. 2003. *Diffusion of innovations*. 5th ed. New York, NY: Free Press.

Ryle, Gilbert. 1949. *The concept of mind* Chicago: The University of Chicago Press.

Shane, Scott. 2003. *A general theory of entrepreneurship: the individual-opportunity nexus*. Cheltenham: Edward Elgar.

Shane, Scott. 2000. Prior knowledge and the discovery of entrepreneurial opportunities. *Organization Science*, 11(4): 448-69.

Shane, Scott & S. Venkataraman. 2000. The promise of entrepreneurship as a field of research. *Academy of Management Review*, 25(1): 217-26.

Venkataraman, S. & Andrew H. Van de Ven. 1998. Hostile environmental jolts, transaction set, and new business. *Journal of Business Venturing*, 13(3): 231-55.

AUTHOR BIOGRAPHY

MARTIN IHRIG is President of *I-Space Institute, LLC*, Research Fellow at the *Snider Entrepreneurial Research Center* at the *Wharton School* (*University of Pennsylvania*), and Research Fellow at *Birmingham Business School* (*University of Birmingham*, UK). He holds a Master of Business Studies from *UCD Michael Smurfit School of Business* and a Doctor of Business Administration from *Technische Universität Berlin*.

Martin is interested in the strategic and entrepreneurial management of knowledge. He has just completed an assignment collaborating with a large international defense contractor in the US and is currently managing a project investigating the evolution and mapping of knowledge in the ATLAS experiment at CERN.

His e-mail address is ihrig@wharton.upenn.edu.

ENHANCING AGENTS WITH NORMATIVE CAPABILITIES

Ulf Lotzmann
Institute of Information Systems Research
University of Koblenz
Universitätsstraße 1, Koblenz 56070, Germany
E-mail: ulf@uni-koblenz.de

KEYWORDS
Norm innovation, Social simulation, Normative agents, Emergence, Immergence.

ABSTRACT

This paper describes the derivation of a software architecture (and its implementation called EMIL-S) from a logical normative agent architecture (called EMIL-A). After a short introduction into the theoretical background of agent-based normative social simulation, the paper focuses on intra-agent structures and processes. The pivotal element in this regard is a rule-based agent design with a corresponding "generalised intra-agent process" that involves decision making and learning capabilities. The resulting simulation dynamics are illustrated afterwards by means of an application sample where agents contribute to a Wikipedia community by writing, editing and discussing articles. Findings and material presented in the paper are part of the results achieved in the FP6 project EMIL (EMergence In the Loop: Simulating the two-way dynamics of norm innovation).

INTRODUCTION

This paper gives an introduction into several major outcomes of the FP6 project EMIL (EMergence In the Loop: Simulating the two-way dynamics of norm innovation) as an example of simulating emergent properties in complex social systems, funded by the EU initiative "Simulating Emergent properties in Complex Systems" (no. 033841). After an overview on the logical normative architecture EMIL-A (which is based on a scientific theory of norm innovation (Andrighetto et al. 2007)), with EMIL-S a software component dedicated to introduce normative capabilities in existing or newly designed multi agent systems is presented.

The EMIL project especially focuses on understanding and analysing norm innovation processes in social systems, which can be seen here as a special case of complex systems, composed of many different interacting intelligent autonomous agents. In general, including norms in multi-agent models seems to be a promising concept to understand human (and artificial) agent cooperation and co-ordination. Therefore, the design of agents with normative behaviour (i.e. normative agents) is of increasing interest in the multi-agent systems research (Boella et al. 2007).

Because of the fact that norms can be seen as a societal regulation of individual behaviour without sheer pressure, special attention in modelling and analysing norm innovation processes should be given not only to the inter-agent behaviour but also to the internal (mental) states and processes of the modelled agents (intra-agent) (Neumann 2008). Following this, the dynamics of norm innovation can be described mainly by two processes:

- immergence: intra-agent process by means of which a normative belief is formed into the agents' minds (Andrighetto et al. 2008). If this happens often enough, the resulting behaviour becomes a "sociological phenomenon" (Durkheim 1895/1982):
- emergence: inter-agent process by means of which a norm not deliberately issued spreads through a society.

These two processes are the basis for the logical architecture. Together with the process of its transformation into a software architecture it will be sketched in the following section. In the subsequent section software architecture is presented in more detail, followed by an application example.

FROM LOGICAL TO SOFTWARE ARCHITECTURE

To derive a software architecture from the logical and cognitive architecture EMIL-A introduced in (EMIL 2009), several steps are necessary. First of all, it is inevitable to formalize and implement several sequential core procedures:

- "norm recognition", i. e. the discrimination between norms and other social phenomena,
- "norm adoption", i. e. the generation of normative goals,
- "decision making", i. e. checking "against potential obstacles to the goal's pursuit" (EMIL 2009), and
- "normative action planning".

Related to the procedures, several data structures for the individual agents have to be defined for representation of:

- "normative beliefs",
- "normative goals" and
- "normative intentions".

Additionally, a "normative board" as a central inventory of norms is necessary as a shared data storage.

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

Furthermore we have to start with the idea that for norms to emerge and to undergo innovation it will be necessary that agent societies must not consist of agents that are entirely lenient with respect to the behaviour of their fellow agents. Thus agents will have to be endowed with a set of goals which they do not necessarily share with all of their fellow agents.

Goals (see (EMIL 2009) and (Conte 2009)) "are internal representations triggering and guiding action at once: they represent the state of the world that agents want to reach by means of action and that they monitor while executing the action." Thus the process of norm emergence or innovation in an artificial society of agents will have to start with actions arising from individual agents' goals.

The process going on in what one could call a primordial artificial society can be illustrated by an everyday example: *A* does not want to be exposed to the smoke of cigarettes (her goal is a state of her environment which does not compel her to inhale smoke and which makes her cough). At this moment this is not yet a normative goal (but it has a similar consequence): to achieve the goal of living in a smoke-free world when the current environment contains a smoker, say *B*, a decision has to be taken which leads to one of several possible intentions which in turn lead to respective actions. One of the possible decisions *A* might take will be to demand from *B*, the smoker, to stop smoking at once and to abstain from smoking in *A*'s presence in all future. When *B* receives this message as a social input he will have to evaluate this message in the norm recognition procedure. If this event (*A* asks *B* not to smoke in her presence) is the first of this kind, *B* will not recognise a norm but store this message and the situation in which he received it as an event in his "event board". When an event like this is more often observed by *B* (but also by observers *C*, *D*, …) this kind of messages might be interpreted (or recognised in terms of EMIL-A, "inferred or induced by the agent on the grounds of given indicators", (EMIL 2009)) as a norm invocation, and a normative belief – "the belief that a given behaviour in a given context for a given set of agents is forbidden, obligatory, permitted, etc." (see (EMIL 2009)) – is stored in all the recipients of the repeated message.

As soon as a social input (such as a message from another agent in a certain situation) is recognised as a norm invocation a normative belief is generated which may (or may not) be adopted, i.e. transferred to the individual normative long term and working memory which consists mainly of the individual normative board for storing normative beliefs and normative goals). If it turns out that the current state of the world does not conform to the normative goal derived from the adopted norm, it is the turn of the decision maker to select from a repertoire of action plans – which in the case of our artificial primordial society must be predefined. The decision maker generates a normative intention which in turn ends up in an action. EMIL-A foresees that these actions can be

- either of the norm compliance or violation type: actions which influence some physical environment
- or of the norm defence type: actions which lead to norm invocations, direct or indirect punishment or just norm spreading through communicative or non-communicative behaviour.

And as a matter of course, an initial repertoire of action plans must be available in each of the agents of the artificial agent society.

The EMIL-S architecture uses very similar concepts to the ones specified in EMIL-A. The rule concept of EMIL-S is somewhat more complex as rules are also responsible for action planning, not only normative action planning. Scenarios run under EMIL-S must also reflect non-normative behaviour of agents.

For a simulation to be run it is necessary to endow software agents with at least some of the capabilities that human actors have by nature: perceiving at least part of the state of the simulated world and acting, i.e. changing the state of the simulated world. Therefore, EMIL-S provides means to include an interface between agents and their environment. Although EMIL-S restricts itself to model the mind of human actors whereas modelling the body of human actors is the task of a simulation system below (Repast (North et al. 2006), TRASS (Lotzmann, 2009) etc.), agent design in EMIL-S has to include modelling that goes beyond modelling the normative processes.

ARCHITECTURE OF EMIL-S

Any multi-agent simulation system will have to be able to simulate the processes which go on within agents (recognition, memory, decision making), among agents (communication) and between agents and their environment (action). Mental processes within agents are thus separated from actions that agents take with respect to other agents (including communication) and their environment (here we have the usual restricted meaning of environment which does not include other agents proper; in this meaning the environment provides means for communication and other resources). Thus one of the central requirements for this kind of simulation is that agents do not communicate by mind-reading but by messages which have to be interpreted by the recipients of messages before they can have any effect on the recipient agent's behaviour.

The strict separation between mental processes and actions, or to put it in another way, between the "mind" of an agent and its "body", allows also to define a simulation tool which can be used for a very wide variety of simulations of human behaviour – as it incorporates some relevant features of mental processes – without trying to be appropriate for simulating a wide variety of settings in which humans would act. This means that the simulation tool created for designing mental processes – mainly decision making processes –

can be rather general and can be reused for a wide variety of situations, no matter what the concrete actions and their consequences for other agents and the environment are. Decision making does not only concern observable actions, but also internal actions, such as one to form or not to form a given mental state.
What agents must be able to do depends on the scenario one wants to simulate but how agents decide which actions they take can be modelled independently of the concrete scenarios.
Consequently, the general structure of the simulation system mainly consists of the module EMIL-S, in which agents with norm formation capabilities and normative behaviour can be defined ("mind"), and the Simulation Tool module, which contains the physical world of a concrete scenario ("body"). On basis of a common interface specification between these modules, different simulation scenarios – realized by using different simulation tools – can be enriched with normative agents.
The EMIL-S module is the core of the simulation system which represents the "minds" of normative agents and realises and implements the logical architecture of EMIL-A. It also provides a user interface (Agent Designer, shown in Figure 1), which allow modellers to design the mental processes of agents which they believe to be relevant in his or her scenario. More precisely, each agent must be equipped with a set of initial rules, which allows him to act in the simulation environment. Rules in EMIL-S are represented as so-called event-action trees, which is a kind of decision tree that represents the dependencies between events and actions. For each event an arbitrary number of action groups are defined (in the figure G1, GNI1-A and GNI1-O). Each action group represents a number of mutually exclusive actions (A10 to A12 or ANI10 and ANI11, respectively). The edges of the tree are attached to selection probabilities for the respective action groups or actions.

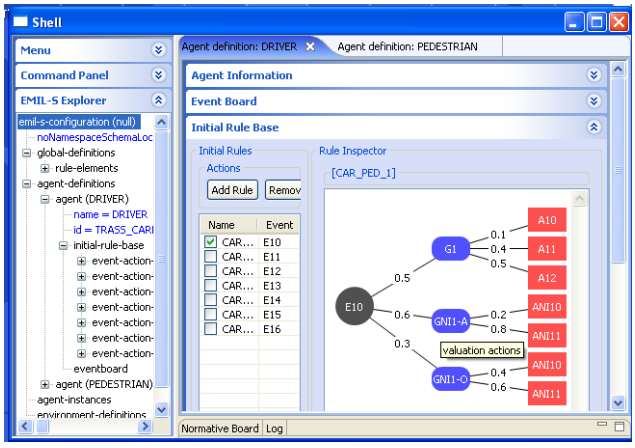


Figure 1: EMIL-S Agent Designer, displaying an event-action tree from the initial rule base

Different scenarios can recommend different simulation tools for the physical layer. In general different simulation tools play the role of the "bodies" of the agents and allow them to act on each other and their environment (which, of course, is also represented in the simulation tool). In addition to the definition of normative agents equipped with an initial rule base in the EMIL-S module, the completion of an executable simulation scenario requires the development of an interface (by using available templates) between EMIL-S and the selected simulation tool. Basically, this interface realises a link between the normative agent parts and their "physical" counterparts in the simulation tool (Figure 2).

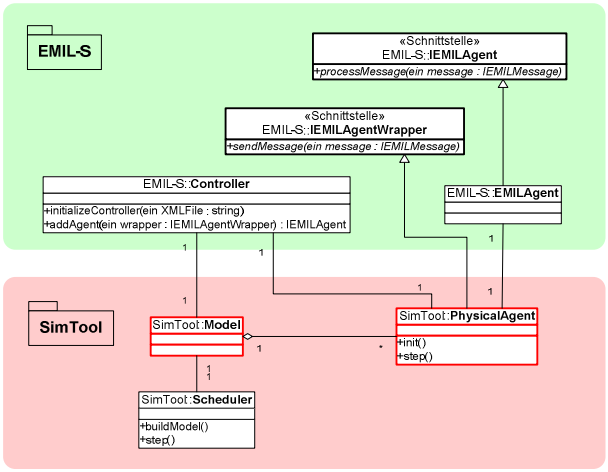


Figure 2: Components of the simulation system: example of an interface between EMIL-S and Repast

EMIL-S AGENT DESIGN

Beyond the requirements derived from the logical architecture, other – in the sense of computer science more technical – aspects have to be regarded. Firstly, the intra-agent process design must allow the handling of complex and adaptive rules. Secondly, the software design should be modularized in a way that general parts of the process are separated from scenarios dependent parts. These two kinds of requirements are essential for all architectural aspects of EMIL-S agents.
Thus, EMIL-S also draws from achievements of several disciplines, in particular from:

- adaptive rules and learning algorithms (cf. Lorscheid and Troitzsch, 2009);
- simulation models (e.g. modelling human needs for market simulations and Wikipedia model, cf. Norris and Jager, 2004; Troitzsch, 2008).

It seems to be generally accepted that each classical intra-agent process consists of three basic steps (as with every other data handling, covering input, processing and output):

- At some point of time an agent as an autonomous entity must check the state of the environment in which it is situated. This is usually done within a perception process. This process changes the agent-internal model of the environment.

- Due to the changed environmental state and due to the agent's internal state, a decision about measures to achieve some individual goals of the agent must be drawn. This is done within a decision process, in many cases based on some sort of rule engines.
- The decision leads to actions directed to the environment with the result of environmental changes.

These steps are shown in Figure 3, together with two additional steps which are essential for adaptive behaviour:

- The action that was performed can change the environment with certain intensity in either a positive or a negative way. Thus, the impact of the action must be evaluated in order to show modified (and preferably better in respect of goal achievement) behaviour at the next appearance of a similar environmental state. This process step is called valuation.
- The result of the evaluation from the previous step must be translated into an appropriate rule change, i.e. the actual rules must be adapted in some way.

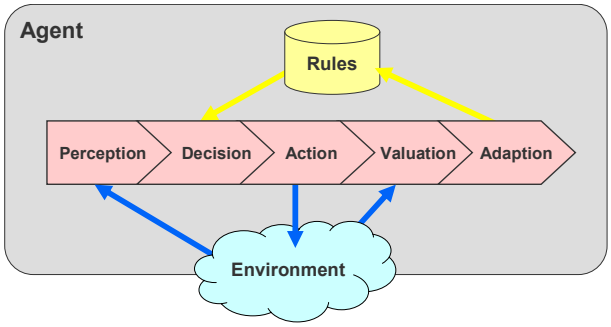


Figure 3: Generalized intra-agent process

The separation into these five process steps gives way to another separation according to the foci of the steps. While the three steps "perception", "action" and "valuation" are connected to environmental affairs, the two other steps "decision" and "adaption" are mainly dedicated to rule base access. This allows the definition of two categories of steps, expressed by two layers:

- A "physical layer", encapsulating all environmental properties, can be seen as a model counterpart to a (human) body. The representation of the body differs from simulation model to simulation model. For example, the physical layer of a traffic participant includes attributes like shape, dimension, orientation, velocity and, furthermore, must feature a special perception sensor able to perceive other agents representing traffic participants as well as topographic elements from the static environment (e.g. a road network). For other models no such complex physical representation is necessary. E.g. for the simulation of agents contributing to a Wikipedia the environment consists of a shared workspace, the physical abilities can be reduced to writing, searching, reading of articles and commenting on them – no aspects of a "real" physical body are relevant.
- On the other hand, the rule decision and adaption process can be abstracted from the environment and coalesced in a "strategic layer". This induces that within the strategic layer a common rule definition language must be established which is used for any kind of simulation scenario.

The advantage of an approach like this is evident: a simulation tool can be realized that completely covers the strategic layer and can be attached (via some sort of interface) as an add-on to other existing simulation tools or programs. Furthermore, due to the rule engine functionality of the strategic layer, the specification of the strategic model aspects can be done at a higher level of abstraction – no "programming" in computer science style is necessary in this respect.

To allow a reasonable level of abstraction, a way must be found to raise the interactions between agents also on an abstract level. On the physical layer a lot of interaction may happen which is not relevant for strategic decisions. On the other hand, all relevant happenings that may occur within the environment also must find their abstract representations. For this purpose a concept based on events and actions is introduced. While events describe all incidents that require attention on the strategic layer, actions express all possible outcomes from the strategic layer. Two different types of events and actions must be distinguished:

- so-called environmental events and actions, directed from or to the physical layer, respectively, and
- so-called valuation events and actions. Origin of a valuation is the "measurement" of the "success" a performed action has had achieved within the environment. For normative simulations according to EMIL-A another source of valuations comes into play: an observing agent valuates (by the act of sending a valuation) another acting agent (which receives this valuation as an event) for an environmental act. This type of valuation is called norm invocation.

The involvement of agents capable to observe and valuate other agents is not only one of the key properties of the EMIL-A framework but also a crucial design element of the agent architecture and, moreover, one of the major innovations of simulator design. For this purpose an agent must be able to cover both the (classical) "actor" as well as the (novel) "observer" roles. This new observer role has some concrete implications, both for intra-agent and for inter-agent processes:

- for inter-agent matters a communication infrastructure must allow to observe ("listen" to) perceptions and actions of observed agents;
- the agent must be equipped with capabilities to generate a model of an observed agent;
- a suitable rule set of the observed agent must be available also for the observer.

While all environmental (and partly valuation) interactions between agents are by definition based on the physical layer, the norm-invocation interactions are situated only within the strategic layer. This kind of interaction, together with a shared "statute book" holding all regular norms that have been either predefined in the scenario or have emerged during the simulation ensure the compatibility to the EMIL-A framework. The two-layer architecture allows to fully integrating these elements within the strategic layer, hence all aspects of the normative process can be made independent from the concrete scenario realization.

In the following the intra-agent process for the actor role is demonstrated for the first cycle of a simulation run.

Figure 4 shows the steps of the decision process that is initiated by an event (E2), which had occurred at the environment and was perceived and by the physical layer. At step 2, only the initial rule base is inspected because the normative frame (as preferred source of rules) is empty at simulation start (time t_0).

During the following adaption process, triggered by a norm invocation, the first normative frame entry is created (Figure 5).

USE CASE: WIKIPEDIA

In this subsection a simulation model which makes use of EMIL-S in combination with the Repast simulation toolkit will be presented. In this use case, findings of the empirical analysis of the behaviour of contributors to and discussants of Wikipedia articles are used to build a simulation model of collaborative writing (Troitzsch 2008). As software agents are still not able to use natural language to produce texts, an artificial language whose symbols do not refer to anything in the real world was invented, and the software agents are endowed with the capability to produce text in this language and to evaluate something like the "style of writing", thus being able to take offence at certain features of texts and blaming the authors of such text. From this kind of

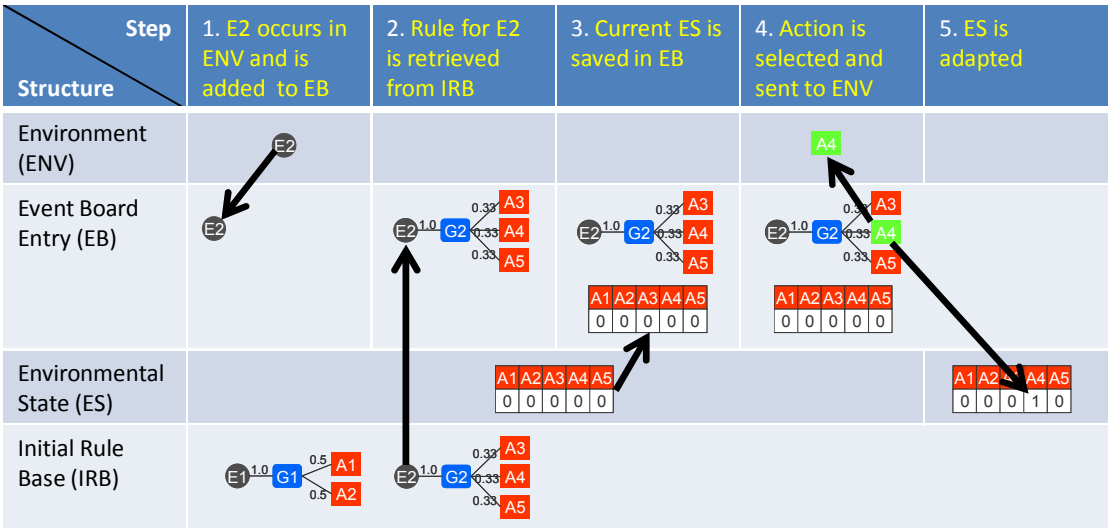


Figure 4: Intra-agent decision process for time t_0

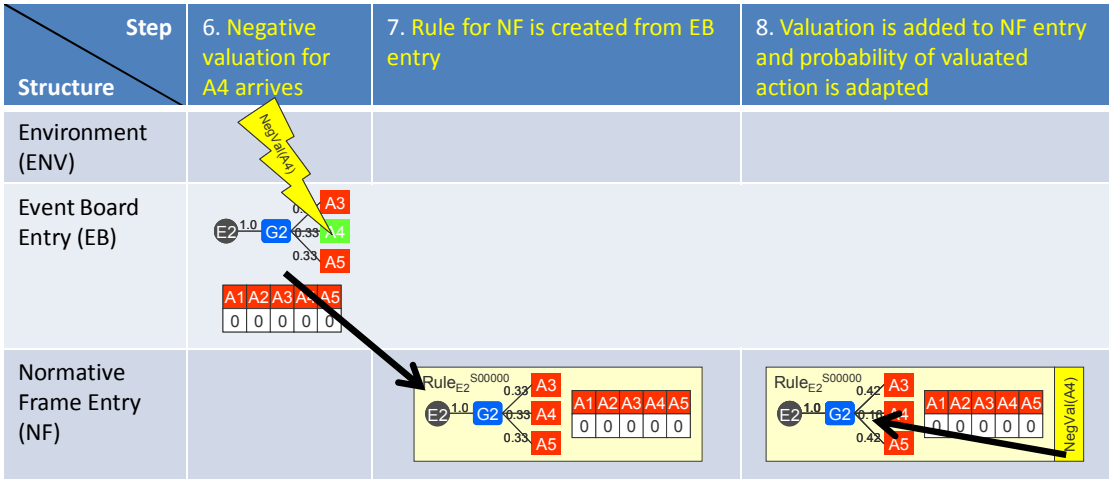


Figure 5: Intra-agent adaption process for time t_0

communication, norms emerge in the artificial society of software agents.

At the beginning of a simulation run agents have a repertoire of messages (or, more precisely, an algorithm that enables them to formulate new messages) that they want to add to the Wikipedia, and at the beginning all agents are authorised to do so. After some time the Wikipedia contains a list of articles which are not yet connected. These articles contain, among others, the name of the author and the content in terms of a string of arbitrary length, the first few elements of which have the special meaning of a keyword. Words consist of alternating vowels and consonants, forming a very primitive language which conveys no meaning, but agents can mull over similarities and differences among Wikipedia articles.

Besides writing, reading and editing articles, agents scan the articles already present for several criteria and comment on them (in the sense of EMIL-S valuations). These commenting actions are of the following types:

- vowel harmony: obeying or violating a special word building rule is assessed, see below;
- duplicate keywords: if more than one article with the same keyword exist then a mild blame (a deontic of the proscription type) is sent to the author(s) of the younger article(s);
- plagiarism: suspected authors are blamed by sanctions (up to the loss of authorisation to write/edit articles).

There are three groups of agents interacting in the scenario:

- "Normal" agents who obey a special vowel harmony (Troitzsch 2008; e.g. the word "aseka" does not conform to the vowel harmony so applying the phonetic process the word would become either "asaka" or "eseke").
- "Rebel" agents who interpret the vowel harmony in the inverse sense, i.e. they prefer words which contain both front and back vowels (e.g. the word "asaka" would be changed to "aseka").
- "Anarchist" agents who have their own word formation rules (e.g. they change every occurrence of the letters: "be" either to "ab" or to "eb", i.e. in addition to a possible vowel change they practice metathesis, such as from Middle English hros to Modern English horse).

The agents were enabled to change their group membership. Every time an agent searches the database for words not obeying its philosophy it changes all occurrences of these words to the "correct" word and blames the author of these words for obeying the latter's rules and/or for making the wrong decision.

Figures 6 to 8 show several output graphs of one of the simulation runs. The initial group affiliation of the agents is set randomly. This assignment has a significant influence for the simulation results, as other runs of exactly the same model reveal (EMIL 2009).

The graph in Figure 6 shows the total number of articles, the number of newly created articles and the deleted articles. Articles are deleted if they are the result of plagiarism or have been added to the Wikipedia as double articles (two articles with same keyword). The total number of articles is increasing slowly; the deletion of articles happens only when "bad" (plagiarism, double) articles are found. In the graph one can see a link between time steps 25 and 30, as at this time a lot of double entries and plagiarisms were found and deleted.

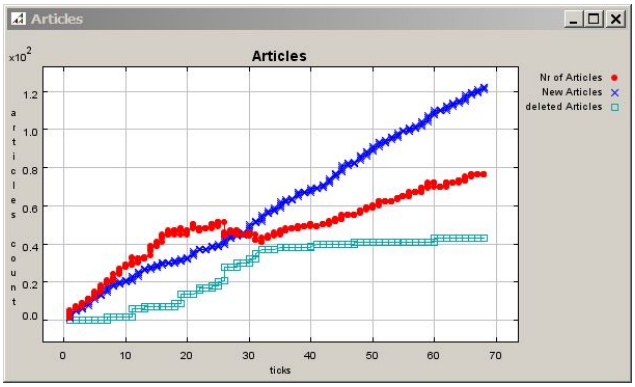


Figure 6: Articles

The graph in Figure 7 shows all blames. Every time an agent blames someone for an action it sends a norm-invocation-message to EMIL-S, and these messages are counted and diagrammed in this graph. The simulation started with two strong groups (the "normal" ones and the "rebels") which blamed each other. After a short while, one of the groups established itself as the dominating one. Every time an agent moves to one of the other groups ("rebels" or "normal" agents), they send a lot of norm-invocation messages for vowel harmony violation. In the graph this curve rises only in the first half of the simulated time (i.e. there are no more vowel harmony blames during the second half of the simulation run).

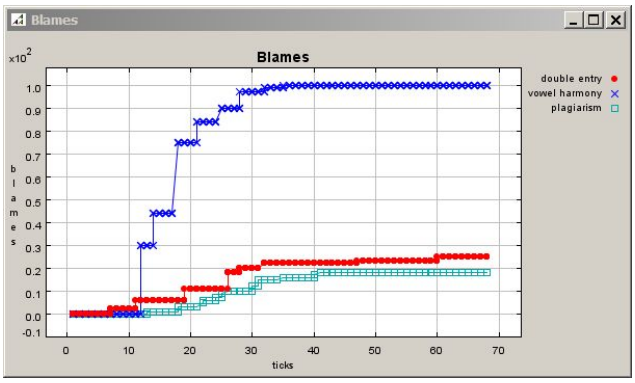


Figure 7: Blames

The graph in Figure 8 shows the group sizes. In this kind of graph, the current number of agents belonging to a group is counted and visualized. Here one can see that in the beginning of the simulation run the agents change quite often between the three groups. At the beginning

of each time tick, every agent decides to which group it wants to belong. After a short period of time the group membership stabilizes. In the actual simulation run this happens between time step 11 and 13; when a particularly high number of norm-invocations for vowel-harmony violation is issued which led to very fast norm learning and consequently to stable group affiliation. As mentioned above, the group membership depends only on the word formation rules the agents comply with.

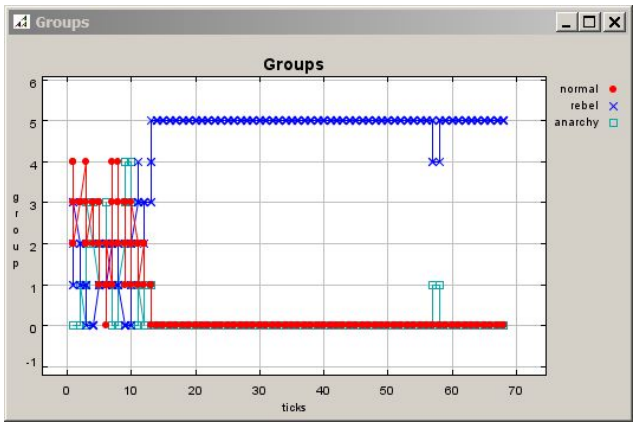


Figure 8: Group sizes

CONCLUSIONS

The paper described architectural aspects of the EMIL-S software for simulating norm formation processes, derived from the logical normative agent architecture EMIL-A. While this paper describes the involved components and processes in general terms (together with a short overview on design and simulation results of the Wikipedia use case), detailed descriptions of application examples in combination with evaluations of the proposed architectures and software would go beyond the scope of this paper. Further information and other application examples can be found in (EMIL 2009), here follows a short use case overview:

- A TRASS-based (Lotzmann 2009) traffic scenario, intended to demonstrate the central features and interfaces of EMIL-S.
- A "multiple contexts" scenario (also TRASS-based), bringing together and into interaction two types of agents, namely one group of agents with a relatively rich cognitive structure and another group of simpler agents which are only social conformers (Andrighetto 2008).
- Another Repast-based scenario of micro-finance groups which goes back to the empirical work of (Lucas dos Anjos et al. 2008).

It is likely that in the future EMIL-S will be used in other and more complex scenarios and applications. Also a further development of EMIL-S towards a powerful platform for an even broader range of rule based normative simulations is one of the most important tasks on the agenda.

REFERENCES

Andrighetto, G., Conte, R., Turrini, P., & Paolucci, M. (2007). Emergence in the loop: simulating the two-way dynamics of norm innovation. In *Proceedings of the Dagstuhl Seminar on Normative Multi-Agent Systems.* Dagstuhl, Germany.

Andrighetto, G., Campennì, M., Conte, R., & Cecconi, F. (2008). Conformity in multiple contexts: imitation vs. norm recognition. In *World Congress on Social Simulation, Fairfax VA July 14-17, 2008.* Fairfax VA.

Boella, G., van der Torre, L., & Verhagen, H. (2007). Normative multi-agent systems. Dagstuhl.

Conte, R. (2009). Rational, goal-oriented agents. In R. A. Meyers, *Encyclopedia of complexity and system science* (pp. 7533-7548). Springer.

Durkheim, É. (1895/1982). *The rules of sociological method and selected texts in sociology and its methods.* London: MacMillan.

EMIL (2009). Emergence in the Loop: Simulating the Two-Way Dynamics of Norm Innovation. *EMIL-T,* Deliverable 5.1.

Lorscheid, I., & Troitzsch, K. G. (2009). How do agents learn to behave normatively? Machine learning concepts for norm learning in the EMIL project. In *Proceedings of the 6th Annual Conference of the European Social Simulation Association.* Guildford, UK.

Lotzmann, U. (2008). TRASS - a multi-purpose agent-based simulation framework for complex traffic simulation applications. In A. Bazzan, & F. Klügl, *Multi-agent systems for traffic and transportation.* IGI Global.

Lucas dos Anjos, P., Morales, F., & Garcia, I. (2008a). Towards analysing social norms in microfinance groups. In *8th International Conference of the International Society for Third Sector Research (ISTR).* Barcelona.

Neumann, M. (2008). A classification of normative architectures. In *Proceedings of the 2nd WCSS.* Fairfax, VA.

Norris, G. A., & Jager, W. (2004). Household-level modeling for sustainable consumption. In *Third International Workshop on Sustainable Consumption.* Tokyo.

North, M., Collier, N., & Vos, J. (2006). Experiences creating three implementations of the Repast agent modeling toolkit. *ACM Tranbsactions on Modeling and Computer Simulation , 16* (1), pp. 1-25.

Troitzsch, K. G. (2008). Simulating collaborative writing: Software agents produce a Wikipedia. In *Fifth Conference of the European Social Simulation Association (ESSA), Brescia September 1-5, 2008.* Brescia.

AUTHOR BIOGRAPHY

ULF LOTZMANN obtained his diploma degree in Computer Science from the University of Koblenz-Landau in 2006. Already during his studies he has participated in development of several simulation tools. Since 2005 he has specialized in agent-based systems in the context of social simulations and is developer of TRASS and EMIL-S. He is also involved in the FP7 project OCOPOMO (Open Collaboration for Policy Modelling) and several other recent projects of the research group. Currently he is doctoral student at the University of Koblenz-Landau. His e-mail address is ulf@uni-koblenz.de.

TYPES OF ANTICIPATORY BEHAVING AGENTS IN ARTIFICIAL LIFE

Pavel Nahodil and Karel Kohout
Department of Cybernetics
Czech Technical University in Prague
Technická 2, 166 27 Prague, Czech Republic
E-mail: nahodil@fel.cvut.cz

KEYWORDS
Anticipation, Agents, Simulation, Behaviour, Artificial Life, Strong and Weak ALife, AI, Artificial Creatures.

ABSTRACT

Anticipation is broad multidisciplinary topic but there are little thoughts on its relation with the reactive behaviour, the similarities and where the boundary is. Reactive behaviour is still considered as the exact opposite for the anticipatory one. It was shown that reactive and anticipatory behaviour can be combined. We will present own viewpoint. Our multi-level anticipatory behaviour approach is based on the current understanding of anticipation from both the artificial intelligence and the biology point of view. The terms as weak and strong artificial life will be discussed here. Original thought is that we use not one but multiple levels of anticipation in a creature design. We focus on the first 2 levels within the 8-factor anticipation framework here. These first two levels may seem trivial, but growing interest in implicit anticipation and reactive anticipation suggest that this is a worthwhile topic. The simplicity of them is intentional to demonstrate our way of thinking and understanding of the topic. The comparison with other results is possible only when these levels are integrated into the complex architecture that can also act as standalone. Examples of experiments follow.

INTRODUCTION AND STATE OF THE ART

The field of Artificial Life (ALife) simulators is very diverse. Most of freely available simulators are concerned with only one phenomenon of ALife.

All beings of nature use anticipation whether they are aware of it or not. Our approach is to our knowledge one of the first that looks at the conscious and unconscious part of anticipation separately. What is not so obvious, that anticipation is not matter of a single mechanism or a control block in a living organism. Anticipation happens on many different levels even in very simple creatures. The works studying anticipation in the past overlooked this fact, focusing on the anticipatory principle or mechanism only. And that exactly allowed us to use not one, but several levels of anticipation in a creature design and to observe the results.

There were two motivational examples to us:

The first example is taken from work (Nadin 2003). Change in posture (standing up from a seated position for example) would cause changes in blood pressure. This is the physics of the body consisting from a liquid (blood), pipes (the various blood vessels), and a pump (the heart). We can understand the mechanism by taking a rubber tube with liquid inside and manipulating it. The pressure of the fluid varies as we change its position from horizontal to vertical according the physical laws. If human beings were subjected to such blood pressure variations as often as we change posture, we all would have a terrible time. So what did we miss? It appears that a certain reflex, called the baroreceptor reflex, affects the heart rate via the nervous system in order to maintain the blood pressure within relatively constant and safe limits. But how does the heart "know" about a change in position? Obviously, the analogy to the simple physical system of a machine with liquid pipes and pumps ends here. The question asked "how does the heart, or whatever controls the heart, know what the body will do" goes beyond the physics of pumping a liquid through pipes. Please note here that this process is not consciously controlled.

As another example showing different level of anticipation, to explain our postulate above about different levels, we can use something very common a tennis player. While playing the opponent, a player is trying to estimate (based on measured observations) the opponent's action and also the result of his action (which is a trajectory of the ball). This estimate is based on the "model" the player has. This model contains among other things the human body physiology, the laws of gravity and physics. These are not expressed as equations but as patterns and rules for reaction. These patterns and responses can be improved by training.

It may seem that these two examples bear little resemblance of them, however they have something in common. In both the *prediction of future state* plays important role in closing the loop of regulation (control) namely. The prediction influences the decision about current state anticipation and the behaviour influenced by the future state anticipatory behaviour. This clearly suggests that anticipation is not just another function of the organism; it is *built in* and applied *on different levels*.

Our goal is to implement anticipation in the agent architecture. While analysing the state-of-the-art, we have felt a gap in the theories shown on the previous example. So initially we focused on revision of the

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

current theories and in the end we tried to formulate own theory. We have received a proof that anticipation, implemented even on these multiple levels can be beneficial with much better results. For this as mentioned, it is necessary to integrate at least the 6 out of our 8 levels in an architecture that can be compared to the other works. Even though our research is not directly tight to an industrial application so far, the anticipation can be applied in variety of areas, because it goes beyond the future state estimation. It provides this as an additional *input* to decision or regulation mechanism. Primary industrial application of our approach is intelligent robotics and humanoids.

WEAK AND STRONG ARTIFICIAL INTELLIGENCE, LIFE AND ANTICIPATION

Before we start describing our view on the reactive and anticipatory behaviour in artificial creatures, we are to discuss the differences between weak and strong approaches in area of *Artificial Intelligence* (AI), *Artificial Life* (ALife) and *Anticipation*. The fundamental difference between strong and weak approach is that in *strong systems* we try to have the required feature while in the *weak systems* we focus "only" on appearing that the system has the required features. Both approaches are valid and have their pros and cons and also their representatives.

Weak and Strong Artificial Intelligence

We can say that strong AI tries to think (act) like humans, while the weak AI tries to think "intelligently" (approximation of human thinking). Reaching the strong AI is hard and it is subject of many arguments and debates. The classical artificial intelligence is focused on the process of thinking, reasoning, problem solving and planning. These problems are usually addressed separately. On the other hand the modern approaches are focused more around behaviour. To this category fall even ALife and the anticipatory behaviour research. Please note that anticipation as such is a general principle and does not fall in the ALife category while the anticipatory behaviour does!

Weak and Strong Anticipation

In view of explicitly mathematically defining systems with anticipation, Dubois introduced the *concept of incursion, an inclusive or implicit recursion*. An incursive system is a recursive system that takes into account future states for evolving. Some non-linear incursive systems show several potential future states, what are called hyper-incursion. A hyper-incursive anticipatory system generates multiple potential states at each time step and corresponds to one-to-many relations. An endo-anticipation is an anticipation built by a system or embedded in a system about its own behaviour. This is no more a predictive anticipation but a built anticipation. In this case, this is a *strong anticipation*. An ego-anticipation is an anticipation made by a system about external systems. In this case, anticipation is much more related to predictions or expectations. This defines a *weak anticipation*. The anticipation of a system can be based on itself, rather than its environment. In a summary the *strong anticipation* refers to an anticipation of events built by, or embedded in a system, *weak anticipation* refers to an anticipation of events predicted or forecasted from a model of a system. We can see that the AI and anticipation understanding are a bit different. Since ALife is part of Artificial Intelligence studies and also employs anticipation, we need to clarify the differences. We identified 8 factors of anticipation in our new approach.

Weak and Strong Artificial Life

Now we would like to put what was said so far into the context of our ALife domain. From AI point of view the ALife belongs to the weak AI category because it is mostly satisfied with approximating rational behaviour and sometimes (especially in life-as-it-could-be simulations) not even that. It seems that the terms of strong and weak properties are different in AI and anticipation fields. Since our primary field is the ALife with anticipatory behaviour we need to be aware of both aspects. What is understood under a strong and a weak ALife? How the weak and strong anticipation are projected in the weak and strong ALife? Since ALife was historically derived from the AI, it also uses the AI understanding of strong and weak property. Weak ALife is viewed as an attempt to learn more about biological life by creating synthetic models of various processes associated with living organisms (e.g. evolution) on artificial media, but do not claim that any parts of their models are actually living themselves. Strong ALife claim that by instantiating such processes in artificial media, the end product will be just as deserving of the term "living" as are biological organisms. Since the artificial creatures created in ALife domain are usually situated in open environment, from the point of anticipation definition they need to employ both weak (*exo*) anticipation and strong (*endo*) anticipation (Elmahalawy and Nahodil 2007a).

From the ALife point of view, the *strong anticipation* is about deriving *information about future without a model* (the only model available is the creature itself). We can easily imagine that if the creature would be complex enough it could model even the situations in the environment on self. The creature in weak understanding has the capability to build the model and then simulate and derive the future information on that. As an opposite of AI where we are trying to achieve strong property, with the anticipation we are trying to achieve weak property. In the both cases we are usually satisfied with the other property (weak for artificial Intelligence and strong for anticipation).

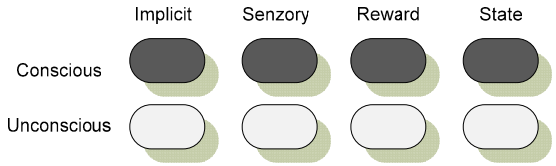
Conclusion of Weak and Strong Properties

Despite the same terminology, the understandings of weak and strong properties are different in the research

areas of ALife and anticipation, making it harder to realize the links and consequences. Our approach is definitely a *weak AI approach* while we are trying to achieve the *weak anticipation*. The reactive anticipation does not create any model to reason hence it uses the strong anticipation. Our goal is to design and build a *multi-level anticipatory creature*, which should definitely exhibit the weak anticipatory property.

MULTI-LEVEL ANTICIPATION

It is obvious that there seem to be several types of anticipation as shown above. Even though we embrace the categorization of Butz anticipation (Butz et al. 2003) we were not entirely satisfied with it. As shown in the examples we struggle to categorize his 4 types from the point of control of the creature's will. We solved this by applying *each category on conscious and unconscious level*, creating thus 8 types of anticipation, which we call multi-level anticipation, or *8-factor anticipation*. This is an original and novel contribution to the theory of anticipation that we came up with and will further develop and research. All the levels are schematically shown on Figures 1.



Figures 1: All Eight Levels of Anticipation.
There are these sorts of Anticipation: *Unconscious Implicit, Conscious Implicit, Unconscious Sensors, Conscious Sensors, Unconscious Reward*, *Conscious Reward, Unconscious State, and Conscious State*

We can say that the complexity grows in the picture from the left to the right and from the bottom to the top. In this article we focus on the first two levels only. One approach or architecture spans through multiple levels usually. First we will describe our understanding of both levels, then we will present an explanatory experiment we conducted to further demonstrate the ideas. To compare the work with other and statistically evaluate, we would need to encompass more than these two levels. (See CONCLUSION.)

Unconscious Implicit Anticipation

Under unconscious implicit anticipation we understand the behaviour that was imprinted in a creature by Nature (evolution) or creator (in artificial case) and that is not voluntary. Reactions and reactive behaviours itself are not anticipatory. It is very often used and wrong understood as an opposite of anticipation. So what exactly are reactions with anticipation (reactive anticipation)? We cannot say that the reaction is associated with prediction of next sensory value, reward or even state because these are not available. We are here at the first and basic level of our multi level anticipatory model and thus there are no means to evaluate a reward or a state. Classical view of implicit anticipation would be satisfied with the fact that it is the prerequisite given to the system either by the long evolution or by the creative mind of architecture designer (Nahodil et al. 2004).
The creature has only the set of *inputs I* and set of possible *outputs O* available (at this very basic level). Let's assume discrete values, which we typically have in a virtual environment. We are not speaking about *agent sensory inputs* or *actions* yet; we intentionally used *inputs* and *outputs* instead. The reason for this is an attempt to generalize this description so it can be used for agent's internal blocks and not only for agent as whole. The reaction base is typically in form of projection $I \Rightarrow O$. The inference mechanism is simple. If any of the input matches, the *output* is executed.
This approach was used by Brooks on MIT round about 1986. There is couple of possibilities for a technical realization from *binary rule base* (Brooks approach) to ACS - *Anticipatory Classifier Systems* (Butz 2002). Let's continue with the formalism. The anticipatory approach would expect another input after the executed action $I \Rightarrow O \; x \; I$. Up till this point this is well known and documented. But on unconscious implicit anticipation there is no mechanism to modify this rule base, other than evolution or redesign. At this level we do not need to solve creation of new or forgetting of obsolete rules here, because the rule base is fixed and it is subject to evolutionary changes only. It seems there are not many options, but *in simplicity is power.* We would like to point out *two possible modifications*, which we also consider *as original ideas*. The first will be described right away, the second will be discussed in the next chapter, and both will be shown in the examples in the *Implicit Anticipation Model* part. One to add to the expectation also expected next action. This is expected to improve the reaction time. It will result in the projection of $I \; x \; O \Rightarrow O^2 \; x \; I$. This describes the expected or better said "prepared" next action. Let's take an example to demonstrate the idea. Imagine the predator evasion situation. One agent is equipped with standard prediction scheme, and second with the modified one as suggested above. Both are in the vicinity of predator (*input I*) the reaction of both is to flee (*output O*). Even if the reaction process is fast it takes some time. Let us analyse what happened in case the original agent needs to take another decision. It has to go through the *action selection process* again to derive new action (even if the new action is the same as previous one). On the contrary if this moment comes to our modified agent it can straight away execute the prepared action.
Except the basic reaction base, another area of involving the implicit and unconscious anticipation is the *inner agent* processes, which imitate the creature physiology, energy management, and other internal

processes. The inner process can either directly or indirectly influence the decision process. If it influences the decision indirectly then it needs some regulation system (once again take the example with the blood pressure). This regulation system is hardcoded within the organism and is not consciously controlled hence it can be classified as implicit unconscious anticipation. The specific implementation of this category can be done via coded *input output base* with an association mechanism. There are of course several possibilities how to code the rule base and the mechanism, but this area is well researched for many years, so we will just build on the existing approaches.

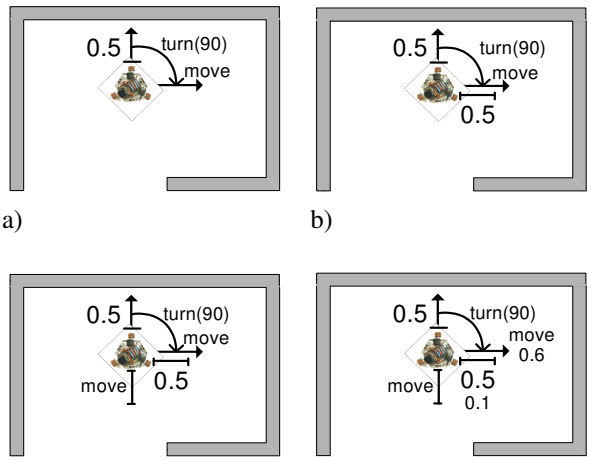
Conscious Implicit Anticipation

The combination may seem illogical at first glance, because of what we have said about implicit anticipation above. It is *something imprinted in the creature* by design. How this can be consciously controlled? The explanation is simple, here still everything depends on the design, but the results are consciously processed or can be chosen to be consciously controlled. This means that inputs of other blocks, such as desired state, are available as inputs for this block. In order to create a new non-atomic action, we can chain the existing actions together, which would have no decision time in between and we can also focus the attention. That is important aspect of behaviour and it can be controlled on this level of our design, even though the higher level can control it too. Our point here is that even if the animal or agent executes the reactive behaviour, it needs to have a focus on something initially.

We will continue here with the formalism we started above. Even here we still have only the set of *inputs I* and set of possible *outputs O*. We can use everything from the previous level up to the next *output prediction* $I \times O \Rightarrow O^2 \times I$. At this point, because it is conscious part, we will introduce the second suggested improvement to the rule base scheme. We do not have the reward yet, but we can have a rate of change for the *input* value. In the ALife, the *output values* have typically discrete values often not expressed by numbers (action for example). In case of the *output* some statistical measure such as probability or likelihood it is appropriate. This is another parameter that can bring value to the decision process and help to choose the optimal action. This describes the typical scenario, but in fact any combination in term of discrete and continuous of the *input* or *output* can occur. So we are adding two new values the *ri* and the *ro*, which we will call *rateability* (combination of words rate and probability). This enriches our rule equation so we get $I \times O \Rightarrow O^2 \times I \times \Re^2$. For continuous or discrete values that are differentiable we can use the rate (first derivation /difference providing the speed of change). For the non-differentiable values we can use probability of their occurrence counted during the agent lifetime to compute the *rateability*. Obtaining these values is not an original idea standalone, but we believe it is new in context of anticipatory behaviour.

Implicit Anticipation Model

We will summarize and illustrate here what was described in the previous two subheadings. We will also compare these to the classical approaches and show that even though what we described seems to be basic and simple the amount of information varies significantly.



Figures 2: The four possible options of available data in the same situation: *a) the classical reactive approach, b) the classical anticipatory approach, c) first suggested improvement - action anticipation, d) second suggested improvement – rateability evaluation*

As mentioned the situations on the Figure 2 are described below with examples reflecting the respective values. The points a. and b. show *current approaches*, the points c. and d. show the *approaches that we suggest use.*

a) The classical reactive approach

IF x THEN y where $x \in I, y \in O$ (1)

Example: IF distance < 0.5 THEN turn(90)

b) The classical anticipatory approach

IF x THEN y EXPECT z where $x, z \in I, y \in O$ (2)

Example: IF distance < 0.5 THEN turn(90) EXPECT distance >= 0.5

c) **First suggested improvement – action anticipation**

IF x AND PREVIOUS_ACTION a
THEN y EXPECT z AND EXPECT_ACTION b
where $x, z \in I, y, a, b \in O$ (3)

Example: IF distance < 0.5 AND PREVIOUS_ACTION move THEN turn(90) EXPECT distance >= 0.5 AND EXPECT_ACTION move

d) **Second suggested improvement – rateability evaluation**

IF x AND PREVIOUS_ACTION a
THEN y EXPECT z AND EXPECT_ACTION b
WITH $< r_i, r_o >$

where $x, z \in I, y, a, b \in O, r_i, r_o \in \Re$ (4)

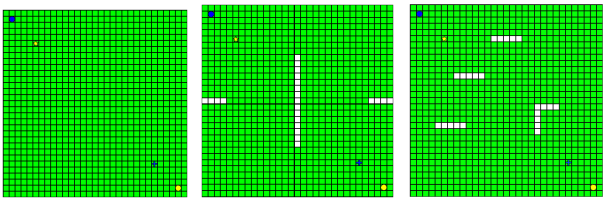
Example IF distance < 0.5 AND PREVIOUS_ACTION move THEN turn(90) EXPECT distance >= 0.5 AND EXPECT_ACTION move WITH <0.1, 0.6>

So far we have described what we understand under these two layers of our design. It is due to describe how to achieve this, what algorithms, approaches and techniques are used. So we need to address how the inference, storing the rules, recalling them, forgetting and matching the input on the sensors to the rules will be done. We have several options here, but we still cannot use any of the successful approaches like Reinforced Learning (RL) or ACS at this level because these consider reward or outcome and these are not available on the presented two levels yet. Two things will be taken under consideration here. One, the above described two blocks are part of a larger architecture and they will certainly interact with other blocks. So we can, on the premise of having other blocks, use ACS and say that it will not only implement the two levels described but others. Second thing is that our goal is to provide a self-sufficient architecture for each level not only for testing and experiments purposes but also for interoperability. This means that we still need a support mechanism for this to make it work. At the second discussed level the animate should be in conscious control and hence the desired value should be available (even when produced by higher levels). We use (for the technical implementation) the *emotivector* introduced by Carlos Martinho (Martinho 2007). We have selected the emotivector for it's simple yet effective way of implementing the attention focus.

EXPERIMENTS AND OBTAINED RESULTS

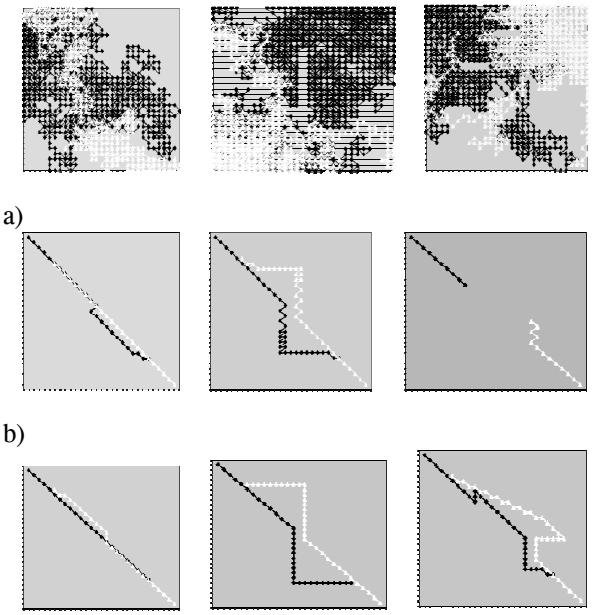
We would like to demonstrate the above suggested approach on experiments conducted namely with the *unconscious implicit anticipation*. Simulations in the virtual environment are the methods (for ALife in most cases) how to test theories and compare effectiveness of results with others. We have conducted our experiments in the REPAST simulation tool (REPAST 2009).

The setup of the experiment is placing an agent (robot) into a simulated world with several objects. There are walls obstructing way here and the beverage, which the agent has to reach. The goal for the agent architecture is to reach the target in effective manner. We focus on showing the *unconscious implicit anticipation effects* to demonstrate our new proposal in more details. There are *3 layouts of the environment* on Figures 3. There are always two agents in each simulation, each trying to reach different goal location.



a) b) c)

Figures 3: The three layouts of the environment *a)* one *without obstacles, b), c) two with different types of obstacles.*



a)

b)

c)

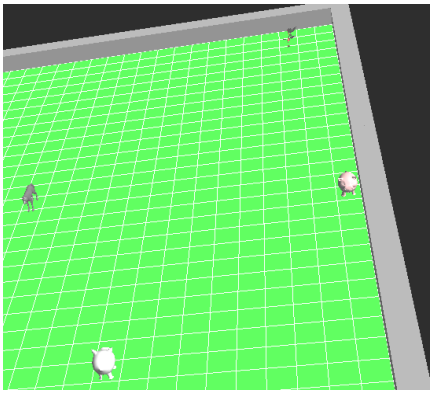
Figures 4: The performed experiments on the top the layouts of the environment. *In the rows results with different decision control mechanism a) random movement b) unconscious implicit anticipation behaviour c) implicit anticipation behaviour based on "wall following".*

The objectives of this experiment are very simple. We are trying to show here the meaning and differences in the implicit anticipation. The setup of the experiment is three different obstacle scenarios with two agents, each trying to reach food. In the first experiment b), the unconscious implicit anticipation of "right angles" was used, which means that when meeting the obstacle rotating by 90 degrees is expected to overcome it. We can see that in this case the 3rd scenario is not achievable (agent get stuck, or oscillate between two positions). In this example the simple imprinted mechanism fails, due to incorrect anticipation. In the third experiment, we used different mechanism "wall following"; hence we anticipate continuing by the wall and not changing direction too often. This architecture can successfully complete the third scenario. What was shown here is the *unconscious implicit anticipation,* as we understand it, the different behaviour patterns could be either designed or evolved.

With the *conscious implicit anticipation* two experiments were conducted. The first experiment was

devoted to the selection of the best predictor for the *emotivector*. Two possibilities were evaluated by *Simple Predictor* (SP) and by *Limited Simple Predictor* (LSP). LSP uses the same equation, but also keeps *history of the input values*, calculates the mean and the deviation and *limits of the prediction* if outside the statistical range (they are described in detail in (Kohout and Nahodil 2007), (Elmahalawy and Nahodil 2007b).

The Simple Predictor convergence speed was slow (convergence to 0.1 ± 0.01 was in 87 steps) while the other predictor was able to converge in 5 steps. These obtained values are then further processed. However the properties of predictor play the key role in the overall design. The other experiment was designed to test the *attention focus mechanism*. In this scenario there were two static agents and two moving agents. One static agent was an observer evaluating the environment input and focusing attention (shown as wolf), the second agent was a non-moving agent (shown as flower). The second two agents (shown as pigs) had different moving patterns. One agent was running in circles and second one was moving randomly the scenario. This simulation is shown on Figures 5 below.



Figures 5. Attention focus mechanism - the 3D model.

The *moving agents are much more interesting for the observing agent than the static* ones. This static agent was expected based on the fact, that emotivector is sensitive to the observed value change in time. This reveals the strong and the weak sides. For the attention focus, the changes in the environment are relevant only on this level. This is acceptable on the *basic "reactive"* level.

CONCLUSION

In this paper we have focused on the *topic of reactive anticipation.* We have shown that the reactive behaviour employs anticipation inherited by the creator (nature, designer) and we demonstrated the *effect of the implicit anticipation on behaviour* on many experiments. These experiments have brought a new light into the anticipation research. These are however basics for our *multi level anticipatory approach.* This approach was not yet proven to be significantly better than other works here due to the limitation of this paper to only the first two levels. The goal was to describe basics of our novelty approach to encompass *the consciousness in the anticipation.* This is a potent topic, and we tried to prove that in behaviour based robotics. This paper also includes a short discussion about weak and strong properties in areas of AI, ALife and Anticipation. The conclusion is that *AI and ALife current definitions are hard to relate to the anticipatory ones*. Authors believe that this paper could be a good starting point for further research.

ACKNOWLEDGMENT

This research has been funded by the Dept. of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague and by the Centre for Applied Cybernetics under Project 1M0567.

AUTHOR BIOGRAPHIES

PAVEL NAHODIL obtained his scientific degree Ph.D. in Technical Cybernetics from the Czech Technical University in Prague, Czech Republic in 1980. Since 1986 he has been a Professor of Mobile Robotics at the Department of Cybernetics at the Faculty of Electrical Engineering in Prague. His present professional interest includes artificial intelligence, multi-agent systems, intelligent robotics (control systems of humanoids) and artificial life approaches in general. He is (co-) author of several books, university lecture notes, hundreds of scientific papers and some collections of scientific studies. He is also a conferences organizer + reviewer (IPC Member) and a member of many Editorial Boards.

KAREL KOHOUT obtained his first - EE MSc. degree in Technical Cybernetics also from the Czech Technical University in Prague, Dept. of Cybernetics in 2006. He is still a PhD student at the same Department of Cybernetics, finishing his PhD theses under supervising of Pavel Nahodil now. He has taken active part on many international conferences. He is also (co-) author of more then 20 published papers in topic of behaviour based control and simulations in Artificial Life Domain mainly.

REFERENCES

Butz, M. V., Sigaud, O., Gérard, P. 2003. Anticipatory Behavior In: *Adaptive Learning Systems: Foundations, Theories, and Systems*, Springer Verlag, Berlin.

Butz, M.V. 2002. An Algorithmic Description of ACS2. In: *Lanzi P-L, Stolzmann W, Wilson, S.W. (Eds). Advances in learning classifier systems,* LNAI vol. 2321, Springer Verlag, Berlin, 211–229.

Elmahalawy, M. A. and Nahodil, P. 2007a. Studying of Anticipation Behavior in ALife Simulator. In: *Research and Development in Intelligent Systems XXIV AI-2007, 27th SGAI IC on AI,* Part 9, Springer, London 369-374.

Elmahalawy, M. A. and Nahodil, P. 2007b. A New Approach Used for Anticipation to Improve the System´s Behaviour. In: *21st European Conference on Modelling and Simulation* ECMS 2007, – *Simulations in United Europe,* Prague, 599-604.

Isla, D. A. and Blumberg, B. M. 2002. Object Persistence for Synthetic Characters. In: *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems.* C. Castelfranchi and W. Johnson (*Eds*). Part 1, ACM Press, New York, 1356–1363.

Kohout, K. and Nahodil P. 2007. Simulation Environment for Anticipatory Behaving Agents from the Artificial Life Domain. In: *Proc. of Intelligent Virtual Agents 2007*, LNCS vol. 4722, Springer Verlag, Berlin, 387-392.

Martinho, C. 2007. Emotivector: Affective Anticipatory Mechanism for Synthetic Characters. PhD Thesis, Instituto Superior Técnico, TU in Lisbon, Lisbon.

Nadin, M. 2003. Not Everything We Know We Learned. In: *Anticipatory Behaviour in Adaptive Learning Systems*. LNCS 2684. Heidelberg: Springer-Verlag.

Nahodil, P., Slavík, P., Řehoř, D., Kadleček, D. 2004. Dynamic Analysis of Agents' Behaviour – Combining ALife, Visualization and AI. In: *Engineering Societies in the Agents World IV*. Springer-Verlag, Berlin, 346-359

REPAST 2009 [online]. Last revision 2009. [cit. 2010-01-10]. http://repast.sourceforge.net/

SIMULATION OF HIGHLY HETEROGENEOUS TRAFFIC FLOW CHARACTERISTICS

V. Thamizh Arasan and G. Dhivya
Transportation Engineering Division, Department of Civil Engineering
Indian Institute of Technology Madras, Chennai – 600 036, India
E-mail: arasan@iitm.ac.in

KEYWORDS
Microsimulation, Traffic Flow, Concentration, Heterogenity, Occupancy

ABSTRACT

Simulation models are mathematical/logical representations of real-world systems. Microscopic traffic simulation models have been playing an important role in traffic engineering and particularly in cases, in which field studies, involving extensive data collection, over a wide range, is difficult or expensive to conduct. As the available simulation models can only replicate homogeneous traffic flow, a model of heterogeneous traffic flow, named, HETEROSIM was developed to simulate heterogeneous traffic flow. In this model, a dynamic stochastic type discrete event simulation is adopted in which the aspects of interest are analysed numerically with the aid of a computer program. The model was validated using field observed data of traffic flow. Then, the model was applied to measure one of the fundamental charactertistics of traffic flow, namely concentration. It is a broader term encompassing both density and occupancy. Occupancy takes into account the traffic composition and speed, and hence, occupancy is more meaningful than density. The concept of occupancy can not be directly applied under heterogeneous traffic conditions, as the traffic has no lane discipline. In this paper, a new concept named, 'area-occupancy' is proposed to measure traffic concentration.

INTRODUCTION

Simulation models are mathematical/logical representations of real-world systems, which take the form of software executed on a digital computer in an experimental fashion. The user of traffic simulation software specifies a "scenario" as model inputs. The simulation-model results describe system operations in two formats: (1) statistical and (2) graphical. The numerical results provide the analyst with detailed quantitative descriptions of what is likely to happen. The graphical and animated representations of the system functions can provide insights so that the trained observer can gain an understanding of why the system is behaving this way. However, it is the responsibility of the analyst to properly interpret the wealth of information provided by the model to gain an understanding of cause-and-effect relationships.

Simulation models may also be classified as being static or dynamic, deterministic or stochastic, and discrete or continuous. A simulation model, which does not require any random values as input, is generally called deterministic, whereas a stochastic simulation model has one or more random variables as inputs. Random inputs lead to random outputs and these can only be considered as estimates of the true characteristics of the system being modeled. Discrete and continuous models are defined in an analogous manner. The choice of whether to use a discrete or continuous simulation model is a function of the characteristics of the system and the objectives of the study (Banks et al. 2004). For this study, a dynamic stochastic type discrete event simulation is adopted in which the aspects of interest are analysed numerically with the aid of a computer program.

As this study pertains to the heterogeneous traffic conditions prevailing in India, the available traffic simulation models, such as CORSIM, MITSIM, VISSIM, etc. which are based on homogeneous traffic conditions, where clear lane and queue discipline exists, are not applicable to study the heterogeneous traffic flow characteristics. Also, the research attempts made to model heterogeneous traffic flow (Katti and Ragavachari 1986; Marwah 2000; Kumar and Rao 1996; Khan and Maini 2000) are limited in scope and do not address all the aspects comprehensively. Hence, there was a need to develop appropriate models to simulate heterogeneous traffic flow. Accordingly, a model of heterogeneous traffic flow, named, HETEROSIM was developed (Arasan and Koshy 2005). The application of the model to study the traffic flow characteristics, at micro level, is dealt with, after a brief description of the important features of the model and its validation, in this paper.

MODELLING HETEROGENEOUS TRAFFIC FLOW

The framework of the heterogeneous traffic flow simulation model, HETEROSIM, is explained briefly here to provide the background for the study. For the purpose of simulation, the entire road space is considered as single unit and the vehicles are

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

represented as rectangular blocks on the road space, the length and breadth of the blocks representing, respectively, the overall length and the overall breadth of the vehicles (Figure 1). The entire road space is considered to be a surface made of small imaginary squares (cells of convenient size - 100 mm square in this case); thus, transforming the entire space into a matrix. The vehicles will occupy a specified number of cells whose co-ordinates would be defined before hand. The front left corner of the rectangular block is taken as the reference point, and the position of vehicles on the road space is identified based on the coordinates of the reference point with respect to an origin, chosen at a convenient location, on the space. This technique will facilitate *identification* of the type and location of vehicles on the road stretch, at any instant of time, during the simulation process.

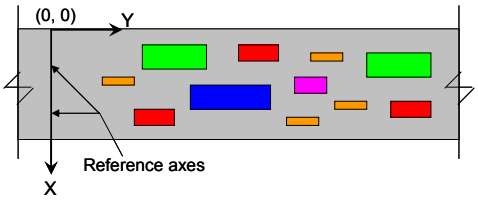


Figure 1: Reference Axes for Representing Vehicle Positions on the Roadway

The simulation model uses the interval scanning technique with fixed increment of time. For the purpose of simulation, the length of road stretch as well as the road width can be varied as per user specification. Due to possible unsteady flow condition at the start of the simulation stretch, a 200m long road stretch, from the start of the simulation stretch, is used as warm up zone. Similarly, to avoid the possible unsteady flow at the end of the simulation stretch due to free exit of vehicles, a 200m long road stretch at the end of the simulation stretch is treated as tail end zone. Thus, the data of the simulated traffic-flow characteristics are collected covering the middle portion between the warm up and tail end zones. Also, to eliminate the initial transient nature of traffic flow, the simulation clock is set to start only after the first 50 vehicles reached the exit end of the road stretch. The model measures the speed maintained by each vehicle when it traverses a given reference length of roadway. The output also includes, the number of each category of vehicles generated, the values of all the associated headways generated, number of vehicles present over a given length of road (concentration), number of overtaking maneuvers made by each vehicle, and speed profile of vehicles.

The logic formulated for the model also permit admission of vehicles in parallel across the road width, since it is common for smaller vehicles such as motorised two-wheelers to move in parallel in the traffic stream without lane discipline. The model was implemented in *C++* programming language with modular software design. The flow diagram illustrating the basic logical aspects involved in the program is shown as Figure 2.

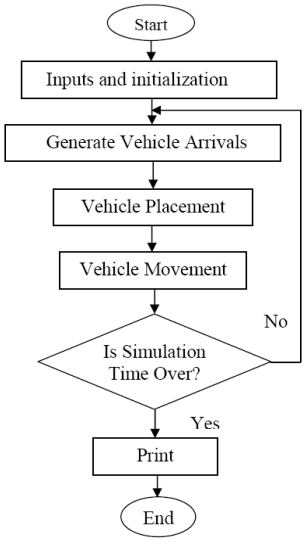


Figure 2: Major Logical Steps of the Simulation Model

Vehicle Generation

In a stochastic traffic simulation process, the vehicles arrive randomly, and they may have varying characteristics (e.g. speed and vehicle type). Traffic-simulation models, therefore, require randomness to be incorporated to take care of the stochasticity. This is easily done by generating a sequence of random numbers. For generation of headways, free speed, etc., the model uses several random number streams, which are generated by specifying separate seed values. Whenever a vehicle is generated, the associated headway is added to the sum of all the previous headways generated to obtain the cumulative headway. The arrival of a generated vehicle occurs at the start of the warm-up road stretch when the cumulative headway equals the simulation clock time. At this point of time, after updating the positions of all the vehicles on the road stretch, the vehicle-placement logic is invoked.

Vehicle Placement

Any generated vehicle is placed at the beginning of the simulation stretch, considering the safe headway (which is based on the free speed assigned to the entering vehicle), the overall width of the vehicle and lateral clearances. If the longitudinal gap in front is less than the minimum required safe gap (space headway), the entering vehicle is assigned the speed of the leading vehicle, and once again the check for safe gap is made. If the gap is still insufficient to match the reduced speed of the entering vehicle, it is kept as backlog, and its entry is shifted to the next scan interval. During every scan interval, the vehicles remaining in the backlog will be admitted first, before allowing the entry of a newly generated vehicle.

Vehicle Movement

This module of the program deals with updating the positions of all the vehicles in the study road stretch sequentially, beginning with the exit end, using the formulated movement logic. Each vehicle is assumed to accelerate to its free speed or to the speed limit specified for the road stretch, whichever is minimum, if there is no slow vehicle immediately ahead. If there is a slow vehicle in front, the possibility for overtaking the slow vehicle is explored. During this phase, the free longitudinal and transverse spacing available for the subject vehicle (fast moving vehicle), on the right and left sides of the vehicle in front (slow vehicle), are calculated. If the spacing is found to be adequate (at least equal to the movable distance of the vehicle intending to overtake plus the corresponding minimum spacing in the longitudinal direction and the minimum required lateral spacing in the transverse direction), an overtaking maneuver is performed. If overtaking is not possible, the fast vehicle decelerates to the speed of the slow vehicle in front and follows it.

The model is also capable of displaying the animation of simulated traffic movements through mid block sections. The animation module of the simulation model displays the model's operational behavior graphically during the simulation runs. The snapshot of animation of traffic flow, obtained using the animation module of HETEROSIM, is shown in Figure 3. The model has been applied for a wide range of traffic conditions (free flow to congested flow conditions) and has been found to replicate the field observed traffic flow to a satisfactory extent through an earlier study (Arasan and Koshy 2005). It may be noted that though the model is primarily intended for simulating heterogeneous traffic flow, it can also be used to simulate homogeneous traffic condition by suitably changing the input data with regard to roadway and traffic conditions.

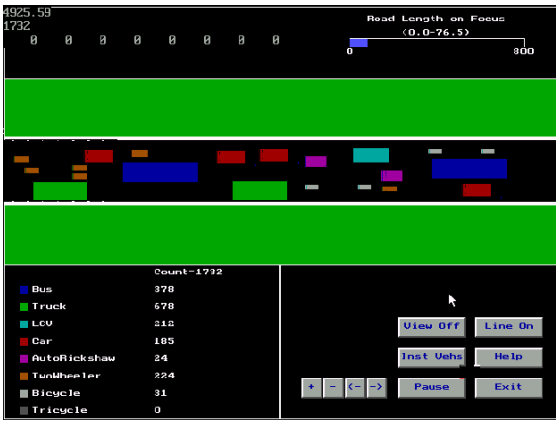


Figure 3: Snapshot of Animation of Simulated Heterogeneous Traffic Flow

Though the model is generally validated, it was decided to check for the appropriateness of the model for the specific requirements of this study by revalidating the model. Field data collected on traffic flow characteristics such as free speed, volume, composition, etc. were used in the validation of the simulation model. The simulation runs were made with different random number seeds and the averages of the values were taken as the final model output.

DATA COLLECTION

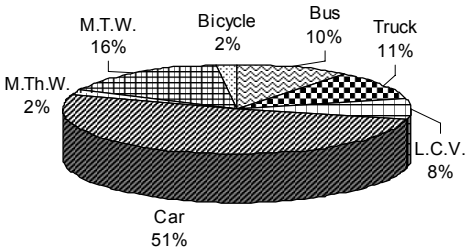
Study Stretch

The stretch of intercity roadway between km 99.4 and km 99.7, of National Highway No. 45 between the cities, Chennai and Chengalpet, in the southern part of India, was selected for collection of traffic data for the study. The study stretch is a four-lane divided road with 7.5 m wide main carriageway and 1.25 m of paved shoulder for each direction of movement. The stretch is straight and level with no side road connections. Also, the traffic flow on the study stretch was unhindered by the road side land uses.

Traffic Characteristics

Collection and analysis of data play a pivotal role in the development of successful simulation models. The field data inputs required for the model were collected at the selected stretch, which had a total carriageway width (including shoulder) of 8.75 m for each direction. A digital video camera was used to capture the traffic flow for a total duration of 1h. The video captured traffic data was then transferred to a Work Station (computer) for detailed analysis. The inputs required for the model to simulate the heterogeneous traffic flow are: road geometry, traffic volume, and composition, vehicle dimensions, minimum and maximum lateral spacing between vehicles, minimum longitudinal spacing between vehicles, free speeds of different types of vehicles, acceleration and deceleration characteristics of vehicles, the type of headway distribution and the simulation period. The required input traffic data for simulation purpose was obtained by running the video of the traffic flow at a slower speed ($\frac{1}{8}^{th}$ of the actual speed) to enable one person to record the data by observing the details displayed on the monitor of the computer. A total of 595 vehicles were observed to pass through the study section in one hour and the composition of the measured traffic volume on the study stretch is as depicted in Figure 4. It may be noted that Animal drawn vehicles and Tricycles, which may be present in small numbers on certain intercity roads, are not present on the study stretch. The free speeds of the different categories of vehicles were also estimated by video capturing the traffic under free-flow conditions. The speeds of the different categories of vehicles were measured by noting the time taken by the vehicles to traverse a trap length of 30 m. The observed mean, minimum and maximum free speeds of various classes of vehicles and their corresponding standard deviations are shown in columns (2), (3) ,(4) and (5) respectively of Table 1. The overall dimensions

of all categories of vehicles, adopted from literature (Arasan and Krishnamurthy 2008), are shown in columns (2) and (3) of Table 2.



L.C.V. - Light Commercial Vehicles, M.Th.W. – Motorised Three-Wheelers, M.T.W. - Motorised Two-Wheelers

Figure 4: Traffic Composition at the Study Stretch

Table 1: Free Speed Parameters of Vehicles

Vehicle type (1)	Free speed parameters in km/h			
	Mean (2)	Min. (3)	Max. (4)	Std. Deviation (5)
Buses	70.0	55.3	84.3	8.9
Trucks	63.1	44.3	80.5	11.0
L.C.V.	66.8	54.9	83.4	8.2
Cars	85.1	58.6	118.0	17.3
M.Th.W	50.2	45.5	61.3	4.9
M.T.W	57.9	38.9	83.8	11.2
Bicycles	14.0	10	20	4.5

L.C.V. - Light Commercial Vehicles, M.Th.W. – Motorised Three-Wheelers, M.T.W. - Motorised Two-Wheelers

Table 2: Observed Vehicle Dimensions

Vehicle type (1)	Average overall dimension (m)	
	Length (2)	Width (3)
Buses	10.3	2.5
Trucks	7.5	2.5
L.C.V.	5.0	1.9
Cars	4.0	1.6
M.Th.W	2.6	1.4
M.T.W	1.8	0.6
Bicycles	1.9	0.5

L.C.V. - Light Commercial Vehicles, M.Th.W. – Motorised Three-Wheelers, M.T.W. - Motorised Two-Wheelers

Any vehicle moving in a traffic stream has to maintain sufficient lateral clearance on the left and right sides with respect to other vehicles/curb/ median to avoid side friction. These lateral clearances depend upon the speed of the vehicle being considered, speeds of the adjacent vehicles in the transverse direction, and their respective types. The minimum and maximum values of lateral-clearance share adopted from an earlier study (Arasan and Krishnamurthy 2008), are given in columns (2) and (3), respectively, of Table 3. The minimum and the maximum clearance-share values correspond to, respectively, zero speed and free speed conditions of respective vehicles. The lateral-clearance share values are used to calculate the actual lateral clearance between vehicles based on the type of the subject vehicle and the vehicle by the side of it. For example, at zero speed, if a motorized two-wheeler is beside a car, then, the clearance between the two vehicles will be 0.2 + 0.3 = 0.5m. The data on, acceleration values of different vehicle categories, at various speed ranges, taken from available literature (Arasan and Krishnamurthy 2008), are shown in Table 4.

Table 3: Minimum and Maximum Lateral Clearances

Vehicle type (1)	Lateral-clearance share (m)	
	At zero speed (2)	At a speed of 60 km/h (3)
Buses	0.3	0.6
Trucks	0.3	0.6
L.C.V.	0.3	0.5
Cars	0.3	0.5
M.Th.W	0.2	0.4
M.T.W	0.1	0.3
Bicycles	0.1	0.3*

*- Maximum speed of these vehicles is 20 km/h

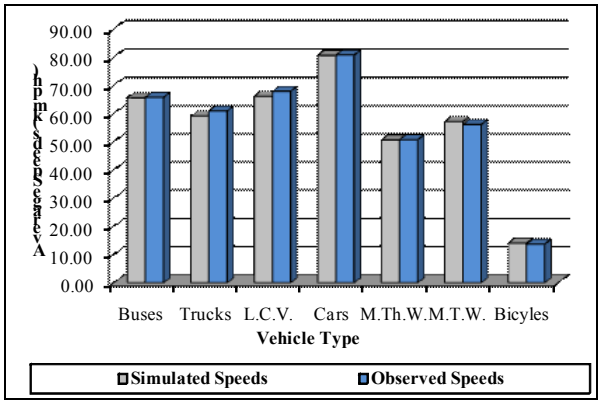
Table 4: Acceleration Rates of Different Categories of Vehicles

Vehicle type (1)	Rate of acceleration at various speed ranges (m/s^2)		
	0-20 km/h (2)	20- 40 km/h (3)	Above 40 km/h (4)
Buses	0.89	0.75	0.67
Trucks	0.79	0.50	0.43
L.C.V.	0.82	0.45	0.35
Cars	1.50	1.10	0.95
M.Th.W	1.01	0.45	0.30
M.T.W	1.35	0.80	0.60
Bicycles	0.10	-	-

L.C.V. - Light Commercial Vehicles, M.Th.W. – Motorised Three-Wheelers, M.T.W. - Motorised Two-Wheelers

MODEL VALIDATION

For the purpose of validation, the simulation model was used to replicate the heterogeneous traffic flow on a stretch of road. The total length of road stretch, for simulation purpose, was taken as 1,400 m. The simulation model was run with three random number seeds, and the average of the three runs was taken as the final output of the model. The observed roadway condition, traffic volume and composition were given as input to the simulation process. The inter arrival time (headway) of vehicles was found to fit into negative exponential distribution and the free speeds of different categories of vehicles, based on the results of an earlier study, (Arasan and Koshy, 2005) was assumed to follow Normal distribution. These distributions, then, formed the basis for input of the two parameters for the purpose of simulation. To check for the validity of the model, it was decided to consider the derived traffic flow characteristics at the micro level so that the validation is satisfactory. Accordingly, the field observed and simulated mean speeds of each of the categories of vehicles were compared to check for the validity of the model. The results of the experiment, for the observed traffic volume of 595 vehicles per hour, are shown in Figure 5. It can be seen that the simulated speed values significantly replicate the field observed speeds for all vehicle types.



L.C.V. - Light Commercial Vehicles, M.Th.W. – Motorised Three-Wheelers, M.T.W. - Motorised Two-Wheelers

Figure 5: Model Validation by Comparison of Observed and Simulated Speeds

A statistical validation of the model, based on observed and simulated speeds of different categories of vehicles, was also done by conducting t-test. The value of t-statistic, calculated based on the observed data (t_0), is 0.89. The critical value of t statistic for level of significance of 0.05 (95% confidence limit), at 6 degrees of freedom, obtained from standard t-distribution table is 2.97. Thus, it can be seen that the value of t statistic, calculated based on the observed data, is less than the corresponding table value. This implies that there is no significant difference between the simulated and observed means speeds.

MODEL APPLICATION

The 'HETEROSIM' model can be applied to measure the heterogeneous traffic flow characteristics on roads for varying traffic and roadway conditions. Here, the application of the model is specific to measure one of the fundamental charactertistics of traffic flow, namely concentration. Concentration is a road traffic measure which explains the extent of usage of road space by vehicles. It is a broader term encompassing both density and occupancy. Density is a measure of concentration over space and occupancy measures concentration over time of the same vehicle stream. Occupancy takes into account the traffic composition and speed, in its measurement and hence, occupancy is more meaningful than density. The concept of occupancy can not be directly applied under heterogeneous traffic conditions, as the traffic has no lane discipline. In this paper, a new concept named, 'area-occupancy' is proposed to measure traffic concentration of any roadway and traffic conditions (Arasan and Dhivya 2008). Considering a stretch of road, area-occupancy is expressed as the proportion of time the set of observed vehicles occupy the detection zone on the chosen stretch of a roadway. Thus, area-occupancy can be expressed as follows;

$$AreaOccupancy = \frac{a_i \sum_i (t_i)_{AO}}{AT} \quad (1)$$

where, $(t_i)_{AO}$ = time during which the detection zone is occupied by vehicle i and the subscript, $_{AO}$ stands for area-occupancy.

a_i = area of the detection zone occupied by vehicle i during time t_i

A = area of the whole of the road stretch

T = total observation period.

Validation of Concept of Area-Occupancy

The concept of area-occupancy can be said to be applicable for any traffic stream under both heterogeneous and homogeneous traffic conditions. To check for the validity of the concept of area-occupancy, as the first step, the density and area-occupancy of a homogeneous traffic stream are related independently to the speed and flow of a stream under homogeneous (cars-only) traffic condition. Since the scope of the experiment is to prove a fundamental relationship, uniform traffic flow on a single traffic lane was considered. Accordingly, the HETEROSIM model was used for simulating the cars-only traffic (100% passenger cars of assumed length 4 m and width 1.6 m) on a 3.5m wide road space - single traffic lane (with no passing). The traffic flow was simulated for one hour (T = 1h) over a stretch of one km. During validation of the model, it was found that three simulation runs (with three different random seeds) were sufficient to get consistent simulation output to replicate the field observed traffic flow. Hence, for model application

also, the simulation runs were made with three random number seeds and the averages of the three values were taken as the final model output. The simulation was run with volumes varying from a low level to the capacity flow condition. Using the features of the simulation model, the times $(t_i)_{AO}$ were recorded for each of the simulated vehicles considering a detection zone length of 3m. The density *(k)* was calculated using the following equation, $q = ku$, where, q = flow of the traffic stream and u_s = space mean speed of the traffic stream. Also, area-occupancy was calculated using equation (1). To depict the validity of area-occupancy, the results of the simulation experiment were used to make plots relating (i) area-occupancy with speed and flow and (ii) density with speed and flow as shown in Figure 6and 7 respectively. It can be seen that area-occupancy and density exhibit similar trends of relationships with speed and flow. The trends of the curves relating area-occupancy with (i) speed and (ii) flow (Figure 6) are the same as those relating density with (i) speed and (ii) flow (Figure 7). Thus it can be concluded that area-occupancy is a valid measure which can be used to represent the concentration of road traffic.

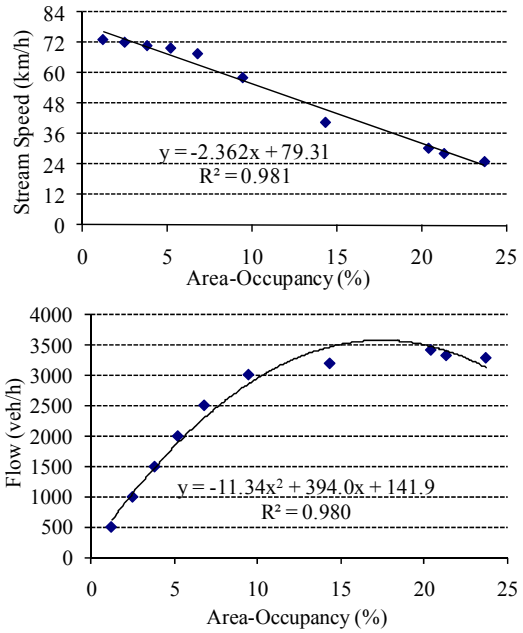
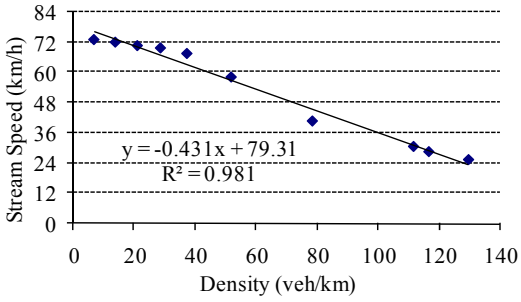


Figure 6: Relationship between Area-Occupancy, Speed and Flow of Homogeneous Traffic



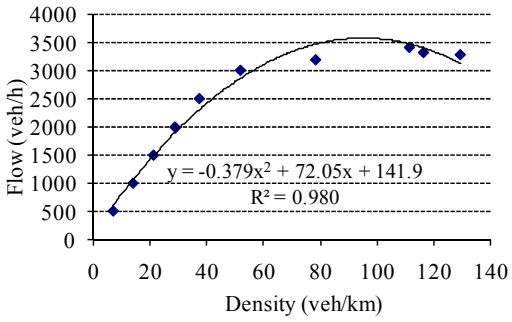


Figure 7: Relationship between Density, Speed and Flow of Homogeneous Traffic

Area-Occupancy of Heterogeneous Traffic

The concept of area-occupancy was applied to heterogeneous traffic condition and relationships were developed between flow, area-occupancy and traffic stream speed. The simulation model was used to simulate one-way flow of heterogeneous traffic on a six lane divided road, with 10.75m wide main carriageway and 1.5m of paved shoulder, for various volume levels with a representative traffic composition prevailing on Intercity roads (Figure 4). The traffic flow was simulated on one km long road stretch for one hour. Using the features of the simulation model, the time $(t_i)_{AO}$ was recorded for each of the simulated vehicles, considering a detection zone of length 3m. The area-occupancy was estimated using equation (1). The average stream speeds and flow of the heterogeneous traffic, for various volume levels were also obtained as simulation output. Then, plots relating the area-occupancy, speed and flow, were made as shown in Figure 8.

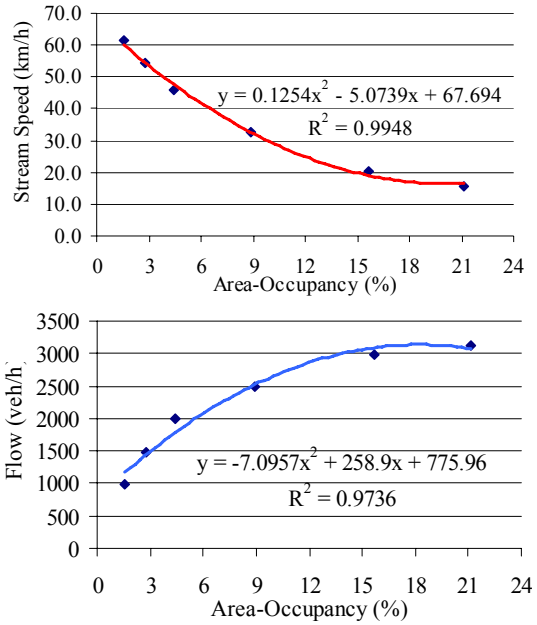


Figure 8: Relationship between Area-Occupancy, Speed and Flow of Heterogeneous Traffic

It may be noted that the decreasing trend of the speed with increase in area-occupancy and increasing trend of the area-occupancy with increase in traffic flow are found to be logical indicating the appropriateness of the area-occupancy concept for heterogeneous traffic.

CONCLUSIONS

The following are the important conclusions of the study:

1. The simulation model of heterogeneous traffic flow, named, HETEROSIM is found to be valid for simulating heterogeneous traffic flow on intercity roads to a satisfactory extent.
2. From the results of model validation, it is found that the simulation model significantly replicate the field observed traffic flow characteristics.
3. It is found, by using simulation model, that the new concept, area-occupancy is a valid measure which can be used to represent the concentration of road traffic under homogeneous traffic condition.
4. From the relationship developed between area-occupancy speed and flow, using the simulation model, it is found that, for the representative traffic composition, the trend of the curves are found to be logical indicating the appropriateness of the area-occupancy concept for heterogeneous traffic conditions.

REFERENCES

Arasan, V.T. and. Koshy, R.Z., 2005. "Methodology for Modeling Highly Heterogeneous Traffic Flow." *ASCE Journal of Transp. Engg.*, Vol. 131, No. 7, 544-551.

Arasan, V.T. and Krishnamurthy, K., 2008. "Effect of Traffic Volume on PCU of Vehicles under Heterogeneous Traffic Conditions." *Road & Transport Research Journal, ARRB*, Vol-17, No. 1, 32-49.

Arasan, V.T. and Dhivya, G., 2008. "Measuring Heterogeneous Traffic Density." In *Proceedings of International Conference on Sustainable Urban Transport and Environment*, World Academy of Science, Engineering and Technology, Bangkok, 342-346.

Banks, J.; Carson, J.S.; Barry, L.N.; and David, M.N., 2004. *Discrete-Event System Simulation.* Pearson Education, Singapore, Third Edition, 12-14.

Katti, V.K. and Raghavachari, S., 1986 "Modeling of Mixed Traffic with Speed Data as Inputs for the Traffic Simulation Models." *Highway Research Bulletin*, No. 28, Indian Roads Congress, 1986, pp. 35-48

Khan, S.I. and Maini, P., 2000. "Modeling Heterogeneous Traffic Flow." *Transportation Research Record*, No. 1678, 234- 241.

Kumar, V.M. and Rao, S.K., 1996. "Simulation Modeling of Traffic Operations on Two Lane Highways." *Highway Research Bulletin*, No. 54, Indian Roads Congress, 211-237.

Marwah, B. R. and Singh, B., 2000. "Level of Service Classification for Urban Heterogeneous Traffic: A Case Study of Kanpur Metropolis." In *Transportation Research Circular E-C018: Proceedings of the 4th International Symposium on Highway Capacity*, Maui, Hawaii, 271–286.

V. THAMIZH ARASAN is currently a full Professor in the Transportation Engineering Division of the Department of Civil Engineering of Indian Institute of Technology Madras, Chennai, India, which is one of the higher technological institutions in the country. He has professional experience of about 30 years in teaching research and consultancy in the area of Transportation Engineering. Travel demand modeling and traffic flow modeling are his areas of research interest. He has guided a number of doctoral degree students and has published more than 100 research papers in international and national journals and conference proceedings. Four of his papers published in journals have received awards for excellence in research. Prof. Arasan has successfully completed several sponsored research projects both at national and international levels. The international projects are: (i) on Development of Transportation Planning Techniques for Indian conditions in collaboration with the Technical University of Braunschweig, Germany and (ii) on Enhancing the Level of Safety at Traffic Signals in collaboration with the Technical University of Darmstadt, Germany. Prof. Arasan is member of several professional bodies and Technical committees. His e-mail address is : arasan@iitm.ac.in

G. DHIVYA is a Ph.D. Scholar in Transportation Engineering Division, Department of Civil Engineering, Indian Institute of Technology Madras, Chennai, India. Her doctoral research work is in the area of 'Microsimulation Study of Heterogeneous Traffic Flow Characteristics'. She obtained her undergraduate degree in the area of Civil Engineering in the year 2003 from University of Madras, Chennai, India and post graduate degree in the area of Urban Engineering in the year 2005 from Anna University, Chennai, India. She received Gold Medal in the Anna University, being first in University Rank in the year 2005. She has published six papers, based on her Ph.D. research work, so far. Her e-mail address is : dhivya.viky@gmail.com

SIMULATION OF TRAFFIC LIGHTS CONTROL

Krzysztof Amborski, Andrzej Dzielinski,
Przemysław Kowalczuk, Witold Zydanowicz
Institute of Control and Industrial Electronics
Warsaw University of Technology
Koszykowa 75, 00-662 Warszawa, Poland
Email: ambor@isep.pw.edu.pl

KEYWORDS
Simulation, traffic control.

ABSTRACT

Simulation of traffic control is very important nowadays for dealing with increasing urban traffic. It helps to construct the strategy of traffic lights switching and determine alternative path in the case of an accident or traffic jam. In the paper we present the application of simulation tool DYNASIM being used by French company DYNALOGIC. It has big number of alternative possibilities in modeling of crossings and conditions. In the paper will be described exemplary simulation of chosen crossings in Warsaw with choice of several different conditions. The simulation package can be free available for universities for educational use.

INTRODUCTION

Control of street traffic in urban conditions is nowadays crucial problem in majority of developed countries. Rapid increase of number of vehicles – before all personal cars and pick-up's – creates in many cities the situation which is difficult to manage. It is important not only to assure fluent movement of cars in chosen directions in specific time periods, but also determination of alternative ways in the situations in which original main communication ways are blocked - by road accident or cars being out of order. Basis for planning of control algorithms is gathering information about the traffic and traffic history on given crossing or/and series of crossings conjugated with a given one. But it is not enough. There is also a need to observe the situation and adapt realized strategy to the changing road situation (Gaca 2008). For this purpose a number of various devices is used – such as induction loops, infrared sensors, radar sensors, photo and video cameras and others. Signals from these sensors are sent mostly through radio transmitters (for short distance) and further by fiber optics (for longer distance).

SIMULATOR DYNASIM

Simulator Dynasim (Citilabs 2005) is a software application that models and simulates the operation of transportation infrastructure. It includes a graphical editor, a simulation engine, and tools for viewing and analyzing simulation results. In fact there are three simulation engines inside:

- microscopic — the simulation considers in details each moving vehicle according to its behavior and immediate environment
- stochastic — the simulation obtains many of the parameter values, like those describing behavior, from statistical distributions
- event-driven — the simulation alters vehicle behavior as a result of simulation events, e.g. change of traffic signal from green to red.

Simulator DYNASIM offers two tools for viewing and analyzing simulation results: the animator and the data viewer. The animator reproduces the vehicle movements calculated by the simulation engine in animations. The data viewer displays the statistical results as graphs according to criteria measured during the simulation, such as the travel time or flow. To support iterative studies, simulator Dynasim organizes data into projects.

A project groups all the related simulation alternatives. A single simulation alternative, called a scenario, is defined by up to four individual components.

Projects reflect an existing condition and then test and compare a number of alternative scenarios. Possible scenarios include:

- change of transportation infrastructure — this can include adding lanes, changing a road's geometric configuration, altering the traffic control, and so on.
- change of transportation infrastructure — this can include evaluating traffic volumes forecast for future years or for changes in the environment.
- change of timing plans or the operation of traffic signals
- change of public transport routes, frequencies, schedules, or vehicle fleets.

A simulation scenario consists of four independently managed components:

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali
David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

1. flow scenario
2. network scenario
3. signal scenario
4. public transport scenario.

Flow scenarios quantify the vehicle demand for the modeled transportation system. Origin-destination matrices specify vehicle demand by vehicle type and time period. You can define flow scenarios that change by time of day, forecast year, or land-use development.

Initial assumptions

Input data for simulation are divided on several categories:
Geografical data:

- movement trajectory – determines the way on which vehicles are moving
- crossings – they have traffic lights, can also have some sensors
- input and output vectors – together with trajectory they determine flow of vehicles

Vehicles data:

- determine dimensions (length), maximal velocity and acceleration.
- Movement of vehicles is modeled by the equation:

$$A2(t+0.25) = a[V1(t) - V2(t)] + b(X1(t) - X2(t) - t\,V2(t) - L] \quad (1)$$

where:

- Ai(t + 0.25) — Acceleration of vehicle i at the moment "t+0.25"
- Vi(t) — Speed of vehicle i at the moment t
- Xi(t) — Position of vehicle i at the moment t
- a, b, t — Parameters that describe three types of acceleration for vehicle 1
- L — Length of vehicle

For every area of simulation we can measure following values:

- Flow of vehicles
- Transition time between crossings (mean and maximal values)
- Number of vehicles, moving between crossings
- Speed of vehicles (mean and maximal values)

Output data are delivered in Excel files. Structure of these files should be carefuly analysed before interpretation.
Simulator DYNASIM can also deliver analysis of various detail degree (scenarios). One can consider only vehicles flow without data from the sensors. One can also take into account privileged public transport, etc.

Example of application

As an example of application let us consider one of typical crossings in Warsaw – the crossing which is next to Warsaw University of Technology, Nowowiejska and Al. Niepodległości (Fig. 1).

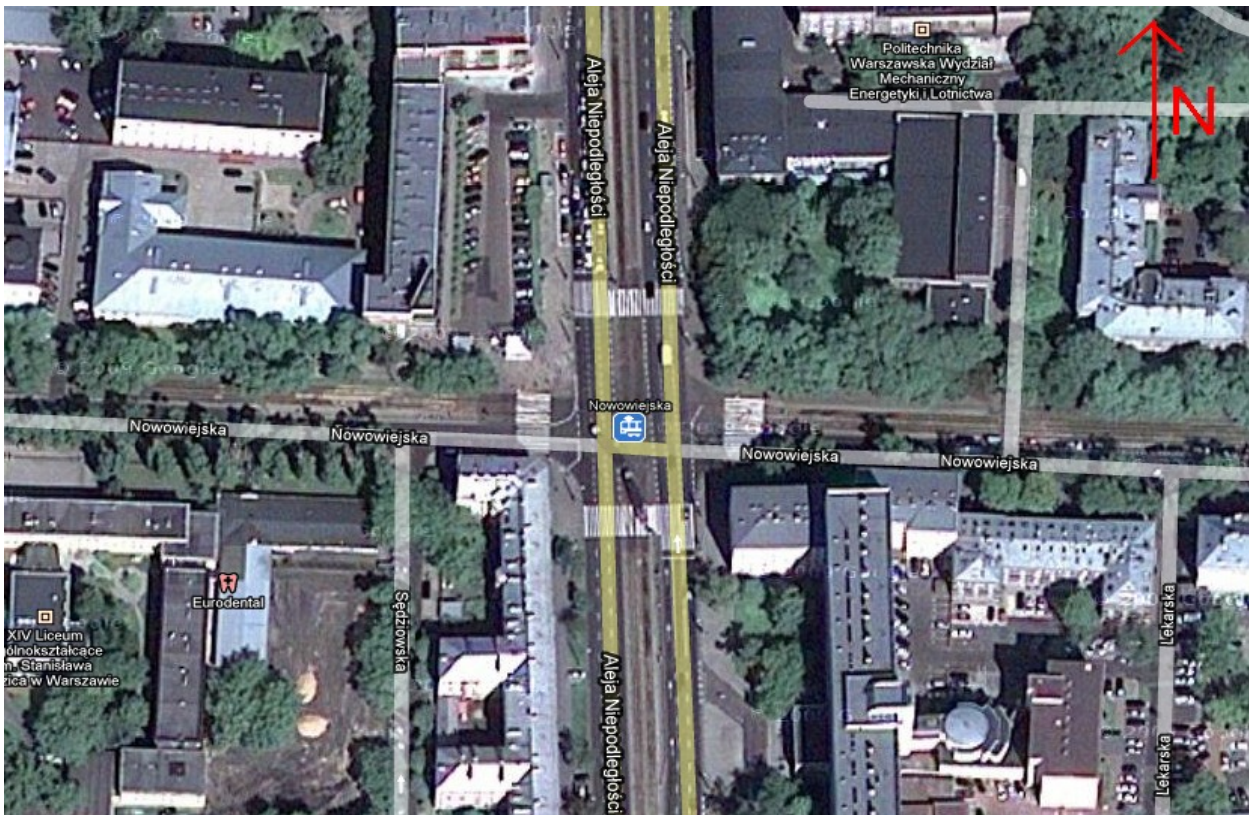


Fig. 1. Air view of the crossing of Al. Niepodległości and Nowowiejska

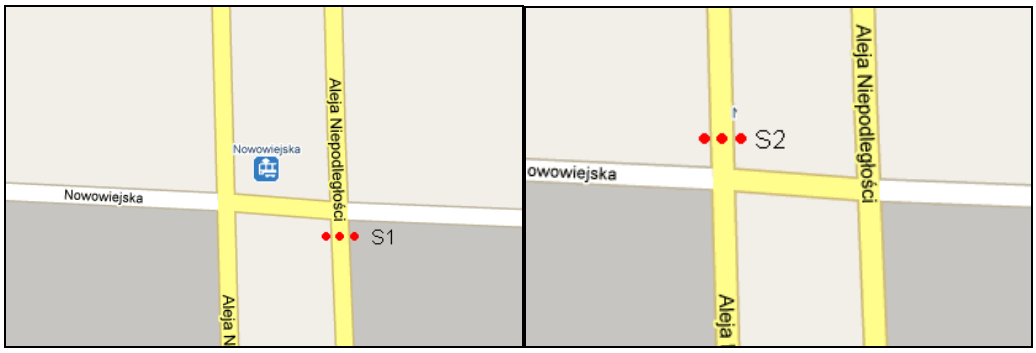


Fig. 2. Simulation of traffic lights on the crossing of Al. Niepodległości and Nowowiejska.

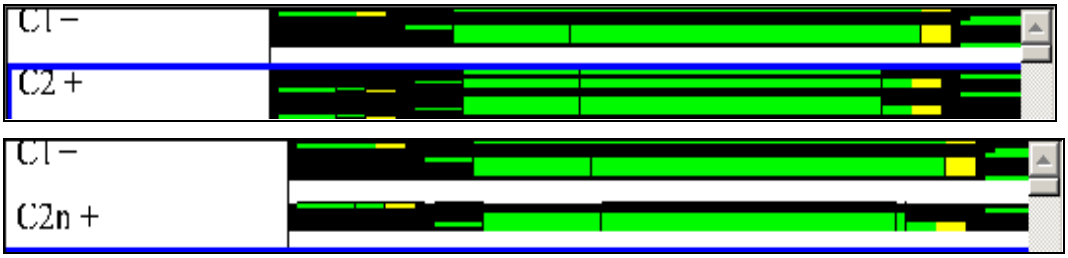


Fig. 3. Scenarios for controllers in simulation of traffic lights on the crossing of Al. Niepodległości and Nowowiejska.

Simulator DYNASIM enables import of maps in various formats – as bitmap of vector presentation. The main street - Al. Niepodległości - has 6 lanes, three on each side. Nowowiejska, the smaller one, has 2 lanes in each direction. It is pretty complicated, although not penultimate. Complication results from the fact that besides cars there are also public transport buses and streetcars. Moreover the streetcars (or tramways) can turn right and left, so separate traffic lights for them can be installed and used. There are also some constraints in turning on this crossing – namely turning left from Al. Niepodległości into Nowowiejska is forbidden for cars, but not for tramways.
In the simulation program the traffic lights are synchronized for all lanes in given direction (Fig. 2). Traffic lights S1 and S2 are controlled by controller C1, traffic lights S3 and S4 (east-west direction) are controlled by controller C2. Controllers C1 and C2 work with different scenarios (Fig. 3).
Simulation package Dynasim enables defining of:

- Traffic lanes for every direction,
- Sources of vehicles and the target of their way
- Trajectories
- Traffic lights with controllers.

Simulation conditions assumed traffic on the level of 20000 vehicles per day (data from General Traffic Measurement Office, forecast on the year 2010) taking into consideration peaks at hours 8:00 i 16:00. Measurement period, to which sequential data are taken, is 1h.

Results of simulation.

Simulation has been performed 100 times for traffic in every possible direction. After averaging of results we got visualisation of the traffic in North-South direction (Al. Niepodległości) presented on Fig. 4.

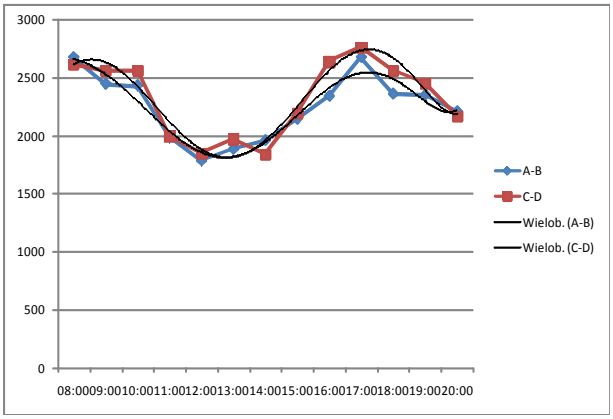


Fig. 4. Results of simulation of traffic in North-South direction (Al. Niepodległości)

It shows the flow (number of cars per hour) passing the crossing in given direction in relation to the time of a day. In the simulation package DYNASIM we can also check what will be changed after introducing some changes in traffic organisation, e.g. by introducing bus lane on every side of main street – Al. Niepodległości.

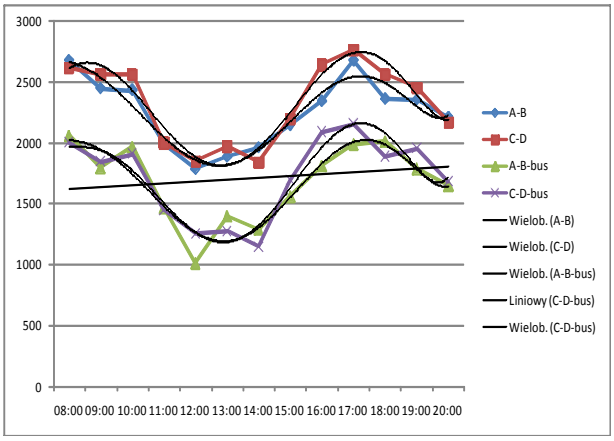


Fig.5. Results of simulation after introduction of bus-lane on Al. Niepodległości.

After changing simulation conditions we get simulation results as presented on Fig. 5.
Output data from each simulation are stored in three kinds of files:

1. RAW files with all simulation data,
2. STAT files with aggregated statistical data (calculated from RAW files)
3. Excel files.

An example of STAT file for one of the simulation runs is presented below on Fig. 6. The stat file lists a value measured at a data collector for each time sample in each iteration. In addition, the file lists statistics—mean, standard deviation, confidence interval, maximum, minimum, 25th percentile, 50th percentile, and 75th percentile—for the measured value during each iteration across all samples, during each time sample period across all iterations, and across all time samples and all iterations.

	Time 1	Time 2	Time 3	Time 4		Mean	Std dev	Conf int	Max	Min	25th pct	50th pct	75th pct
Iteration 1	25.00	26.00	45.00	15.00	:	27.75	12.53	15.42	45.00	15.00	17.50	25.50	40.25
Iteration 2	27.00	38.00	58.00	83.00	:	51.50	24.61	30.29	83.00	27.00	29.75	48.00	76.75
Iteration 3	65.00	57.00	48.00	25.00	:	48.75	17.29	21.28	65.00	25.00	30.75	52.50	63.00
Iteration 4	34.00	38.00	41.00	30.00	:	35.75	4.79	5.89	41.00	30.00	31.00	36.00	40.25
Mean	37.75	39.75	48.00	38.25									
Std dev	18.57	12.82	7.26	30.48									
Conf int	22.86	15.78	8.93	37.52									
Max	65.00	57.00	58.00	83.00									
Min	25.00	26.00	41.00	15.00									
25th pct	25.50	29.00	42.00	17.50									
50th pct	30.50	38.00	46.50	27.50									
75th pct : Overall	57.25	52.25	55.50	69.75	:	40.94	17.79	8.02	83.00	15.00	26.25	38.00	54.75

Sample .stat file

Fig. 6. Example of STAT-file with explanation of data structure.

CONCLUSIONS

Simulation presented in the paper is based on professional simulation program used for commercial purposes. It is well structured, relatively easy to operate and efficient in everyday work. Simulation scenarios record the data specified by the output groups associated with the network scenario's layers. Therefore, one does not need to specify output parameters for each simulation scenario; instead, you simply specify the data collected for each unique network. The combination of output group and layer have to be unique. One can use the same output group name in different layers. This corresponds to two different configurations of the same transportation system.

REFERENCES

Gaca S., Suchorzewski W., Tracz M.: Engineering of road traffic, Warszawa, WKiŁ 2008.
Kawalec P., Firlag K.: „Simulation research of local controllers with dispersed structure for road traffic". International Scientific Conference Transport in XXI century", Warszawa, 2004.
Suda J.: "Public transport fleet management. Warsaw application" Warsaw Conference on Public Transport, 10-11 October 2005

Citilabs 2005: "Description of the simulator Dynasim". Internal report, 2005.

AUTHOR BIOGRAPHIES

KRZYSZTOF AMBORSKI was born in Lublin, Poland. He received his M.Sc. in Telecommunications (1964) and Ph.D. in Control Systems (1972) from Warsaw University of Technology. In 1971 he received also M.A. in Mathematics from the University of Warsaw (Poland). Since then he was with the Institute of Control and Industrial Electronics at Warsaw University of Technology, now as associate professor. In 1994-2002 he was active as an expert in Multimedia Broadband Services at Telekomunikacja Polska S.A. In 1998/2004 he was professor at the University of Applied Sciences in Darmstadt (Germany), in 2003 he was visiting professor at the University of Wisconsin-Platteville (USA).
E-mail: ambor@ee.pw.edu.pl

ANDRZEJ DZIELINSKI is a professor at the Warsaw University of Technology. He received M.Sc. in 1983 in electrical engineering specialising in Control Systems, Ph.D. in 1992 and D.Sc. in 2002. His main professional interests are in computational intelligence, modelling, simulation and automatic control. He is all the time with the Institute of Control and Industrial Electronics, now as Deputy Director for Science.
E-mail: adziel@ee.pw.edu.pl

PRZEMYSŁAW KOWALCZUK is with the Warsaw University of Technology. He is specializing in databases, database management, simulation in logistics and computer science, working towards his Ph.D.
E-mail: pkowalczuk@ee.pw.edu.pl

WITOLD ZYDANOWICZ received his M.Sc. in Telecommunications (1964) and Ph.D. in Control Systems (1972) from Warsaw University of Technology. He is co-author of several textbooks on control systems. He is now adiunct there.
E-mail: witek@isep.pw.edu.pl

MATHEMATICAL SIMULATION OF THE MAGNETIC FIELD OCCURRED BY THE ARMATURE REACTION OF THE SYNCHRONOUS MACHINE

Aleksandrs Mesņajevs
Andrejs Zviedris
Faculty of Power and Electrical Engineering
Riga Technical University
1 Blvd Kronvalda, 1010 Riga, Latvia
E-mail: kbl@inbox.lv, aaazzz@eef.rtu.lv

KEYWORDS
Mathematical simulation, synchronous machine, magnetic field.

ABSTRACT

Using proposed methodology and mathematical simulation parameters of the armature winding, magnetic system geometry is estimated, as well as the saturation's influence on the mentioned parameter is evaluated.
Methods are based on magnetic field mathematical simulation, solving partial derivatives of the differential equation comparatively to the vector magnetic potential A and using one of the most efficient methods – the finite element method (FEM) (Voldek 1978). There is theoretically justified and described method of obtaining the armature winding phase magnetic-flux linkage, i.e., $\Psi(\omega t)$, from the vector potential which, as a function of spatial coordinates $A(x, y)$, is received by mathematical simulation of the stationary magnetic field.
The method is based on solving several stationary magnetic field equations. In equations as a field source are defined the phase currents instantaneous values, which are aligned with the rotor rotation angle $\alpha_i = \omega t_i$.
According to the classical synchronous machine two-reaction theory, synchronous reactances X_d and X_q are determined from the magnetic field's fundamental harmonic in the air gap.

INTRODUCTION

Synchronous machine's inductive reactances and the processes depending on them determine the electromagnetic field, i.e., it's spatial distribution and change in time. This field in operating conditions can be described by Maxwell's equations.
FEM is the most efficient numerical method, which easily allows accurate enough to consider those specific, important factors such as any geometrical shape and size complexity of the magnet system's individual elements, the ferromagnetic material's non-linear characteristic, as well as the field sources (any actual distribution of winding current).

As experience shows while studying the magnetic field, it can be assumed that the machine's magnetic field is plane-parallel.
It is appropriate to substitute the electromagnetic field equations system with one equation which depends on the magnetic vector potential, having only one (axial) component $A = A_z$ in a plane-parallel field. In such equation A is spatial coordinates x, y and time t function, which means that the equation (1) describes space and time alternating magnetic field (Voldek 1978)

$$\frac{\partial^2 A}{\partial x^2} + \frac{\partial^2 A}{\partial y^2} = -\mu j_a , \qquad (1)$$

where μ - magnetic permeability; and j_a – the external field source's current density.
Solving this type of equation is rather difficult, therefore, it is appropriate to reduce it to number of simpler tasks.
In this case, the task is based on the fact that in time varying process ($\partial/\partial t \neq 0$) can be viewed as a single fixed process set of different consecutive time points t_1, t_2, ... ,t_i, ... ,t_n. So, for example, if the field source current density is in time varying sinusoidal function $j_a(t) = j_{am} \sin \omega t$, then equation (1) must be solved n times, each time in the right side of that equation define moment of time t_i that corresponds to current densities moment value $j_a(t_i) = j_{am} \sin \omega t_i$. Solving such a task as the results receive vector potentials, which are essentially a table formed functional dependence $A = f(t)$. A similar approach is used solving non-stationary field equations, when rotor speed $v \neq 0$. In this case, it is possible to solve a number of magnetostatic field equations. Each of them corresponds to different consecutive rotor positions.
For the mathematical simulation of the magnetic field and obtaining results the complex multi-functional program *QuickField* (QuickField 2009) is used.
Software provides opportunities for the following actions:

- to describe the geometric model (or topology) of the object under study;

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

– to assign the medium characteristics, including various ferromagnetic material magnetization curves $B = f(H)$;

– to assign field sources - a current density in windings as a function of spatial coordinates;

– to assign the Dirichle and/or the Neuman boundary conditions;

– to solve tasks with high precision;

– to get a visual picture of the field;

– to calculate various electromagnetic field differential and integral characteristics.

This paper addresses the following main tasks:

– to apply and to use the available modern software for mathematical simulating of the magnetic field using numerical methods;

– to illustrate with examples the practical use of methods, which quantifiably estimate the magnetic constructive parameters of the system and, above all, the saturation effect on parameters of the machine's electromagnetic field.

DETERMINATION OF THE SYNCHRONOUS REACTANCE USING THE SYNCHRONOUS MACHINE'S MATHEMATICAL SIMULATION RESULTS

As known (Voldek 1978), the magnetic asymmetry of the salient pole synchronous machine's rotor ensures reluctances, which are bigger in quadrature axis (q direction) than the reactance in the direct axis (d direction). Therefore, for the synchronous machines with rotor's magnetic asymmetry, it is appropriate to use a two-reaction method based on the superposition principle. According to this principle, the direct axis (Φ_d flux) and quadrature axis (Φ_q flux), where the magnetic flux is operating, are mutually independent. It should be noted, that this assumption is correct only for machines with unsaturated magnetic system. However, making additional adjustments based on the magnetic field mathematical simulation results; two-reaction method can also be used for machines with a saturated magnetic system.

According to the two-reaction method theory, the synchronous inductive reactances X_d and X_q, can be determined:

$$X_d = \frac{E_d}{I_d} \tag{2}$$

$$X_q = \frac{E_q}{I_q} \tag{3}$$

In these equations, EMF E_d and E_q are the EMFs induced by the fundamental harmonic armature winding of the direct field and the quadrature field; I_d and I_q - armature current's direct and quadrature components.

The effective value of the EMFs E_d and E_q can be determined using the following formulas

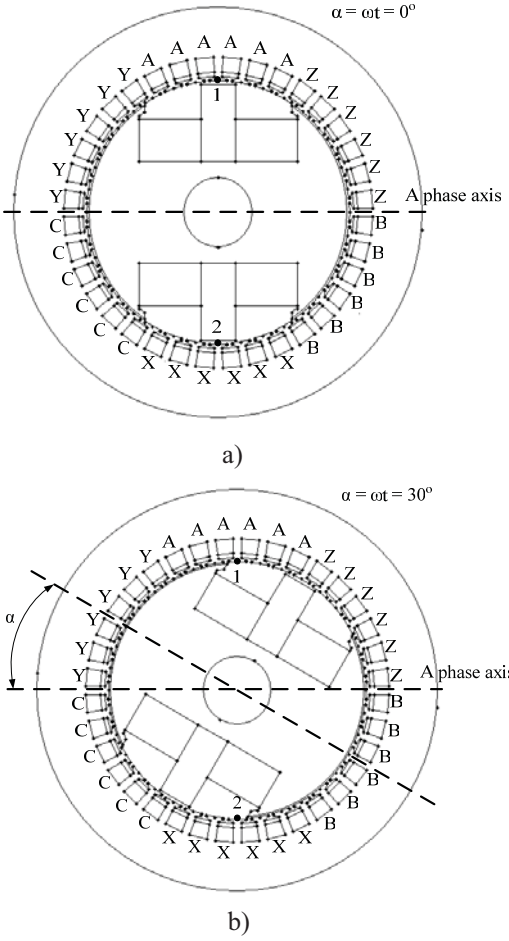
$$E_d = 4{,}44 f \Psi_{d1} k_{w1} \tag{4}$$

$$E_q = 4{,}44 f \Psi_{q1} k_{w1} \tag{5}$$

where Ψ_{d1} and Ψ_{q1} - fundamental harmonic ($\nu = 1$) amplitude values of flux linkages generated by the direct field and the quadrature field , k_{w1} - the fundamental harmonics winding factor.

The necessary sequence of operations is the following: first, determine Ψ_{d1} and Ψ_{q1} (see (4) and (5)), which, in turn, are required for reactance X_d and X_q determination from (2) and (3).

The figure 1 shows the calculation region of the simplified magnetic field model for the three-phase ($m = 3$) two-pole ($2p = 2$) machine with a number of armature slots $Z = 36$.



Figures 1. Distribution of currents in the armature winding phases for determination of the direct axis field's flux linkage $\Psi_{d(\omega t)}$ of the armature in time moments $\omega t_i = 0^\circ$ (a) and $\omega t_i = 30^\circ$ (b), when the rotor rotation angle (electric degrees) is 0° and 30°

The instantaneous value of current in armature windings phases may be determined by the following formulas:

$$\left.\begin{aligned} i_A &= I_m \cdot \cos(\omega t - \alpha) \\ i_B &= I_m \cdot \cos(\omega t - 120^\circ - \alpha) \\ i_C &= I_m \cdot \cos(\omega t - 240^\circ - \alpha) \end{aligned}\right\} \quad (6)$$

where $I_m = \sqrt{2} \cdot I$ - the armature current as the peak value of the field source.

To determine the flux linkage $\Psi_d(\omega t_i)$ with sufficient accuracy, it is preferably that the time step is chosen small enough. As experience of numerical experiments shows, the rotor rotation angle step is appropriate to be chosen so, that $\Delta\alpha = t_z/2$, where t_z - the armature tooth pitch.

The magnetic flux of one pole can be obtained with different armature current instantaneous values according to (7) with the rotor position angle α_i. After a series of magnetic field calculations for time moments t_i

$$\Phi_d(\omega t_i) = \Phi_d(\alpha_i) = (A_{1i} - A_{2i})l \quad (7)$$

and flux linkage with armature winding phase as a function of time

$$\Psi_d(\omega t_i) = (A_{1i})wl \quad (8)$$

where A_{1i} and A_{2i} - the vector potential values on the surface of the armature in points 1 and 2 (fig. 1.), w - number of turns per phase, l - the machine length in axial direction.

Numerical harmonic analysis of the function $\Psi_d(\omega t_i)$ allows to obtain the fundamental harmonic amplitude value Ψ_{d1} of the flux linkage's, which is required according to formula (4) for determination of EMF E_d.

Similarly to the expression (5), the EMF E_q can be obtained if α is replaced with $\alpha - 90^\circ$ (see table 1.) while simulating quadrature axis magnetic field in expression (6).

RESULTS OF THE MAGNETIC FIELDMATHEMATICAL SIMULATION, ITS' ANALYSIS AND USE

The simulation of the synchronous machine magnetic field is made for the experimental three-phase machine with $2p = 2$, in which a single layer full step ($y = \tau$) winding is located in armature 36 slots. Other machine parameters: $f = 50$ Hz, turn number per phase $w = 60$, slots per pole and phase $q = 6$, winding factor $k_w = 0.96$, armature rated current $I_N = 6.56$ A. The ferromagnetic elements of the machine magnetic system are made up from the electrical-sheet steel. The main geometric dimensions are: outer diameter of the armature $D_a = 0.212$ m, the armature's inside diameter $D = 0.136$ m, the air gap $\delta = 0.001$ m; the armature's active length $l_\delta = 0.125$ m; the pole overlapping factor $\alpha_p = 0.604$, the pole pitch $\tau = 0.214$ m m.

Calculations were made for five armature current values $I = (0.4;\ 0.6;\ 0.8;\ 1.0;\ 1.2)I_N$, where the effects of saturation on the magnetic field's character, on parameters X_d and X_q values are evaluated. To obtain the flux linkage with armature winding's phase for a certain current value as a function of time, the instantaneous current value is set according to the rotor rotation angle with step $\Delta\alpha = 5^\circ$ (electric degree), which is matched with the armature current phase's change along the same angle.

Because of the periodicity and symmetry of the function $i(\omega t)$, it is sufficient if the calculations are made in one quarter of the period of current changes, that is $\alpha = \omega t = 90^\circ$ (electric degree).

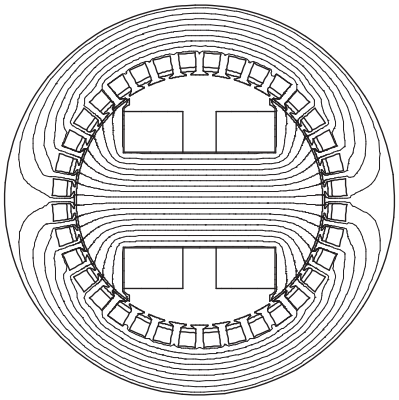
The estimated by the formula (7) armature winding slot's current instantaneous value in relative units are given in table 1. Phase's zone width equals to 6 slots for chosen machine.

Table 1. Phase current's instantaneous value in relative units for different rotor rotation angles (electrical degrees)

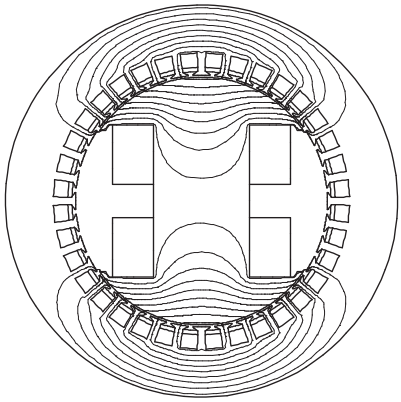
$\alpha = \omega t(^\circ)$	$i^*(A)$	$i^*(B)$	$i^*(C)$
0	1.000	-0.500	-0.500
5	0.966	-0.423	-0.574
10	0.985	-0.342	-0.643
15	0.966	-0.259	-0.707
20	0.940	-0.174	-0.766
25	0.906	-0.087	-0.819
30	0.866	0	-0.866
35	0.819	0.087	-0.906
40	0.766	0.174	-0.940
45	0.707	0.259	-0.966
50	0.643	0.342	-0.985
55	0.574	0.432	-0.996
60	0.500	0.500	-1.000
65	0.423	0.574	-0.996
70	0.342	0.643	-0.985
75	0.259	0.707	-0.966
80	0.174	0.766	-0.940
85	0.087	0.819	-0.906
90	0	0.866	-0.866

In the above mentioned tables angle α values are shown for simulation of direct axis field cases assuming that the *A*-phase axis coincides with the pole direct-axis (see also fig 1.). The angle $\alpha' = \alpha - 90^\circ$ should be used for simulation of the armature quadrature-axis field.

Figure 2 shows the direct axis and quadrature axis magnetic field's picture, acquired for machine, if the armature current components I_d and I_q in relative units $I_d = I_q = 1$, i.e. equal to the rated current.



a)



b)

Figures 2. Armature direct reaction (a) and quadrature-axis reaction (b) magnetic field pictures for angle in moment ωt, when the phase *A* current is at maximum value

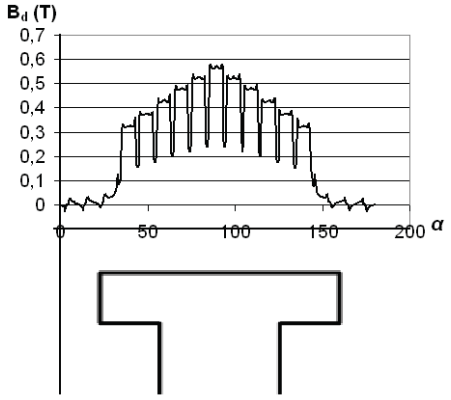
Figures 3 shows distribution curves of the armature direct axis and quadrature axis magnetic flux densities B_d and B_q determined on the teeth and slots middle points on the armature surface.

Flux linkage curves $\Psi_d(\omega t_i)$ and $\Psi_q(\omega t_i)$ are obtained with calculations carried out for different rotor rotation angles, observed with the armature phase currents angle for moments t_i ($\alpha_i = \omega t_i$).Various armature currents I_d^* and I_q^* values for one half of the period are shown on Figure 5.

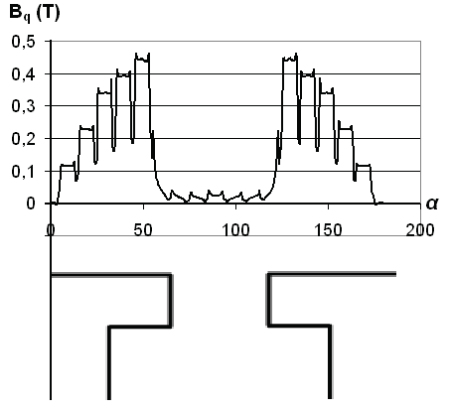
The flux linkage fundamental harmonics Ψ_{d1} and Ψ_{q1} are determined during numerical harmonic analysis of $\Psi_d(\omega t_i)$ and $\Psi_q(\omega t_i)$. The EMFs E_d and E_q values for various armature currents are calculated (see table 2 from formulae (4) and (5)).

Table 2. EMF E_d and E_q fundamental harmonics dependence on the armature currents

I^*	E_d (V)	E_q (V)
0.4	44.44	17.45
0.6	65.09	26.16
0.8	80.94	34.58
1.0	93.29	42.60
1.2	103.49	49.90



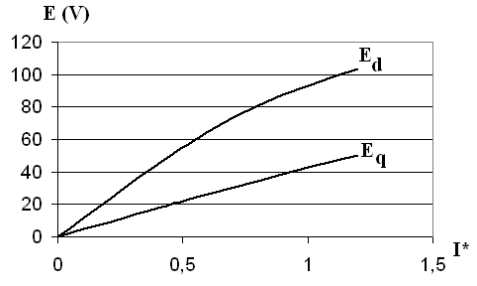
a)



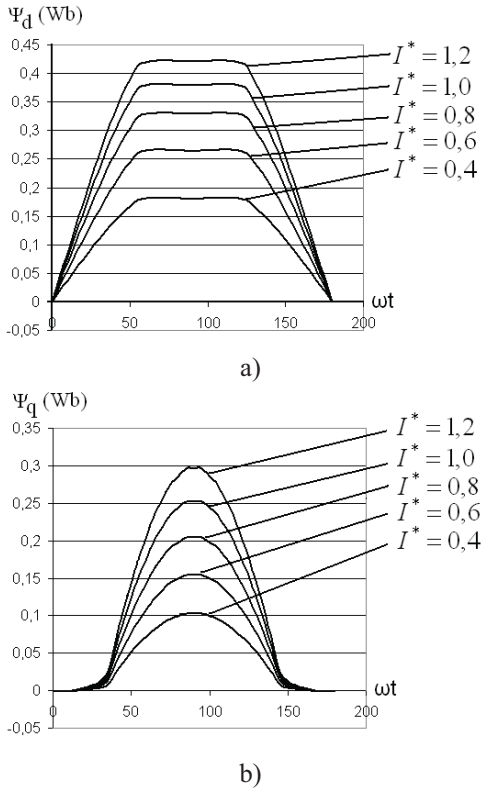
b)

Figures 3. Armatures direct axis (a) and quadrature axis (b) magnetic flux densities distribution in the air gap (on the armature surface); (angle α in electrical degrees)

Magnetic system's saturation can be assessed by curves $E_d = f(I^*)$ and $E_q = f(I^*)$ shown on fig. 4.



Figures 4. Armature winding EMF dependence from armature currents in relative units



Figures 5. By armature direct reaction (a) and quadrature reaction (b) induced flux linkage with armature windings phase as a function of time

Reactances X_d and X_q calculated from formulae (2) and (3) are given in table 3. Table 3 also indicates the saturation coefficients $k_{\mu d}$ and $k_{\mu q}$ values corresponding to the rated current $I^* = 1$.

Table 3. Reactances dependence from saturation level induced by armature currents

I^*	X_d (Ω)	X_q (Ω)	X_d / X_q	$k_{\mu d}$	$k_{\mu q}$
0.4	16.93	6.65	2.55		
0.6	16.53	6.64	2.49		
0.8	15.42	6.59	2.34		
1.0	14.22	6.49	2.19	1.22	1.02
1.2	13.14	6.34	2.07		

CONCLUSIONS

The finite element method offers tremendous opportunities for determining the synchronous machine's magnetic field and from it depending characteristics. Synchronous direct axis reactance X_d and quadrature axis reactance X_q can be correctly determined from the magnetic field's mathematical simulating results applying the principle of superposition used in the classical synchronous machine theory. Saturation effect on parameters X_d and X_q can be estimated by the saturation factors $k_{\mu d}$ and $k_{\mu q}$. Saturation factors values are depending on magnetic system geometric dimensions and on level of saturation ($k_{\mu d} = 1.2 \div 1.5$, $k_{\mu q} = 1.0 \div 1.2$). Estimation results are also confirmed by the experimental results of examined machine.

REFERENCES

Bianchi N. 2005. Electrical machines analysis using finite elements. Boca Raton. CRC Press.

QuickField. 2009. Finite element analysis system. Version 5.7. User's guide. Svendborg. Tera Analysis Ltd.

Voldek A. 1978. Electrical machines. Leningrad. Энергия.

Zviedris A. 1973. Electromagnetic calculations in electrical machines. Riga. РПИ.

AUTHOR BIOGRAPHIES

ALEKSANDRS MESŅAJEVS was born 1985 in Latvia. In 2008 he graduated from Riga Technical University, gaining M.Sc.ing. degree. Presently he is a PhD student.

In March 2008 received certificate "Base and advanced simulation using QuickField software".

From 2006 he is working as a laboratory assistant and scientific assistant in Electrical Machines and Apparatus Department of Riga Technical University.

ANDREJS ZVIEDRIS was born 1938 in Latvia. In 1961 he graduated from the Riga Polytechnical Institute (RPI) Faculty of Electrical and Power Engineering, gaining the qualifications of Engineer in the Electrical Machines and Apparatus speciality. In 1970 defended a thesis and obtained a Candidate degree of Technical Sciences. Dr.Sc.ing. degree was conferred to A.Zviedris in 1992.

After graduating the RPI A.Zviedris has worked as an assistant Professor (1968-1973), and Department of Electrical Machines and Apparatus Head (1973-1984). In 2001 he was elected as Associate Professor in the Department of Electrical Machines and Apparatus, where work to date.

Scientific activities of A. Zviedris are related to Mathematical Simulation of magnetic fields in Electrical Machines. The results of his research are aprobated in more than 50 scientific publications and technical reports in International Scientific conferences.

A. Zviedris is an Expert in Standardization Technical Committee of Electrotechnical Terminology of Electrical Engineering, a member of Latvian Union of Scientists.

AN ESTIMATION OF PASSENGER CAR EQUIVALENT OF MOTORBIKES

Ngoc-Hien Do[†1], Quynh-Lam Ngoc Le[2], and Ki-Chan Nam[3]
[1,3] Department of Logistics Engineering, Korea Maritime University,
#1 Dongsam-dong, Yeongdo-gu, Busan, 606-791, Republic of Korea
[2] Department of Industrial Systems Engineering, Ho Chi Minh City University of Technology
268 Ly Thuong Kiet, District 10, Ho Chi Minh City, Viet Nam
Email Addresses: [1]hienise97@yahoo.com, [2]lequynhlam@yahoo.com and [3]namchan@hhu.ac.kr

ABSTRACT
Much research has been conducted on trying to solve traffic problems and applied successfully in many developed countries where the car is the main transport mode. However, it has not been effective in most developing countries where the motorbike, rather than the car, is the main transport vehicle. This paper estimates the passenger car equivalent of motorbikes under simulation analysis. Simulation scenarios are constructed and run, and the results are analyzed to determine that factor. The research shows that the passenger car equivalent of motorbikes depends on traffic conditions. The factor is expected as a useful parameter when applying traffic research in developing countries.

KEY WORDS
Mixed traffic system, Homogeneous traffic system, Simulation analysis, Passenger car equivalent of motorbikes

1. Introduction

Traffic systems particularly in developing countries are completely different from those in developed countries, where the motorbike is the main transport form rather than the car, so traffic systems' behaviors are distinct. Rather than fully complying with the traffic regulations, vehicles are also guided by nature rules. Together with other characteristics, traffic systems in developing countries are complicated. As in developed countries, traffic issues are interesting research topics and challenges for both the governments and researchers.

Many research models have been efficiently applied. Traffic flow theories including both macroscopic and a microscopic treatment are intended to provide an understanding of the phenomena relative to the movement of individual vehicles along a highway. Other theories about queuing and delays at isolated intersections have been constructed to discuss the effect of isolated intersections on the delay to drivers. For reasons of safety and avoiding the coincidence of two cars occupying the same space at the same time, a traffic control theory was presented. Other traffic theories such as traffic generation, distribution, and assignment have been introduced [1].

An agent-based approach used to design a Transportation Regulation Support System (TRSS) is able to monitor the network activity under normal conditions and to automatically adjust to the environment changes by proposing feasible solutions in order to optimize the traffic flow [2]. To examine the effects of vehicle policy intervention on urban development and population, GDP and environment aspects, a system dynamics approach based on the cause-and-effect analysis and feedback loop structure was proposed [3]. To reduce infrastructure investment, a single lane for traffic in two directions is constructed by using Automatic Guided Vehicles [4].

In addition, many traffic simulations have been developed and efficiently used. The CORSIM, a microscopic traffic simulation model, has been constructed and used mainly in the U.S., which has specific strengths that assist in the modeling of complicated geometry conditions, to simulate different traffic conditions, account for the interactions between different components of networks, interface with external control logic and program, and model time-varying traffic and control conditions. Furthermore, many new traffic simulations are useful tools to support traffic operation analysis such as INTERGATION, VISSIM, MITSIM, WATSIM, PARAMICS, and TRANSIMS [5].

Although the aforementioned models have been successfully applied in developed countries such as France [2], Netherlands [4] and USA [5] and the modern cities as Dalian, China [3], they are unlikely to give reliable results if absolutely applied to the traffic systems in developing countries. The passenger car is usually used as one of the important indexes of the traffic systems. Rarely research has been done on the passenger motorbike, while it is the primary transport mode in most of developing countries. For example in Viet Nam, according to the Vietnamese Traffic and Public Works Service (2007), about 21.72 million individual motorbikes were registered in 2007, in a population of 81 million, compared with only 1.11 million cars. Therefore, for application in developing countries in general further considerations and adjustments need to be investigated.

†: Corresponding author
Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

In this paper, the passenger car equivalent of motorbikes is estimated under simulation analysis, where a simulation model for the mixed traffic system [6] is used to support. For this estimation, suitable simulation scenarios are generated. Analyses are performed based on simulation results. The passenger car equivalent could be used as a referent factor to convert the number of motorbikes to that of car and vice versa. It supports the more convenient determination of alternatives to improve traffic systems in developing countries. Finally, some suggestions and conclusions are proposed.

2. Methodology

The system considers two situations. The first system comprises 100% cars and 100% motorbikes that are analyzed separately. Two simulation model groups are developed, in which one is used to simulate only the cars' behaviors and the other only the motorbikes'. Based on the simulation results, the system's parameters in individual situations, the relations among the system's factors, and the system's saturated conditions are determined. Finally, based on comparisons between two cases, an estimation of passenger motorbikes equivalent of a car is done in the homologous traffic system.

The other situation considers a mixture of traffic (car and motorbike) on the road. The interaction between two transport modes is considered. Actually, the ratio between them is changed according to the simulation scenarios under the system's saturated conditions. The simulation results are recorded and analyzed to determine that factor.

2.1 Input Data

A set of experiments are constructed on a stretch of road, as shown in Figure 1. The 8-meter wide, two-lane road comprises three segments, in which vehicles are generated in the first 100m-long segment as a warm-up segment, the main physical part of the system is the second 500m-long segment, and the final 100m-long segment is used to release vehicles. The physical system is coded and converted into a part of the simulation data file, which is a fixed-data element of simulation scenarios.

Another unchangeable component is vehicles' physical parameters. When a vehicle travels on the road, from the top-down view it occupies an area, so in this simulation program each vehicle is represented by an appropriately sized rectangular. Therefore, the simulation model is modeled as a 2D one. Each vehicle has its own velocity and acceleration. The maximum velocity differs not only between cars and motorbikes, but also among the same vehicles. These characteristics are the main reasons for the conflictions among the entities in the system. All the relative information is shown in Table 1.

Table 1: Vehicles' Parameters

	Physical information (m)		Velocity (km/h)			Acceleration (m/s^2)
	Length	Width	$Vmax_a$	$Vmax_b$	Vmed	
Car	4	1.8	40	80	50	2.5
Motor	2	0.8	30	60	40	3

The maximum speed is modeled as drawn from a uniform distribution in the range $Vmax_a$ to $Vmax_b$. In addition, the initial speed follows a normal distribution, in which Vmed is the mean of population. Furthermore, the Poisson distribution is used to generate vehicles. In this paper, its mean, the number of vehicle entering the system per minute, depends on the scenarios, which is called the expected volume factor.

2.2 An Overview of Simulation Model

Many logic models have been applied to simulate driver-vehicle-units (DVUs) behaviors, such as car-following, free acceleration, lane-changing, direction-changing, stop-run, and intersection-conflict models [6]. Especially, the lane-changing approach usually used to simulate passenger car is modified to simulate the passenger motorbike. It is called as sub-lane changing approach, in which a sub-lane is a virtual lane on the road. Motorbikes can occupy any lateral position across the carriageways instead of traveling within real lanes, so they can move to one or two sub-lane(s) either on the left or the right hand side, while cars have to change to the next real lane on the left hand side. Although moving to the right side or changing more than one sub-lane to overtake another is illegal, it is ubiquitous. Priorities of lane usage are denoted from one to four, in decreasing order of common priority, which is called as flex-passing rules as shown on the Figure 2. In addition, motorbikes usually move in virtual groups. The sharp, speed and quantity of DVUs of the groups are changed usually and depends on the lead vehicle(s) and DVUs behaviors in the groups.

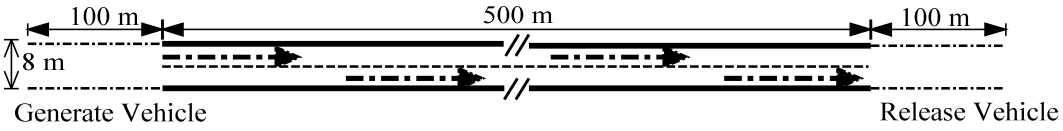


Figure 1: Physical Simulation System

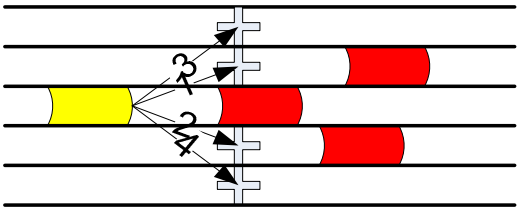


Figure 2: Flex-Passing Rules on Virtual Lanes

Some factors or indexes such as Volume IN, Volume OUT, Average Speed, and Density Index are used to validate the simulation model and evaluate the simulation results. These factors are recorded directly or indirectly through parameters obtained from counter machines set up at all inputs and outputs of roads. Among these factors, Volume IN (VI_{ij} - vehicles/minute) is the average number of vehicles type i^{th} travelling into the system at the road j^{th}, Volume OUT (VO_{ij} - vehicles/minute) is the average number of vehicles type i^{th} exiting the system at road j^{th}, and the Average Speed is determined through vehicle's travelled distance and time. The density of traffic used to evaluate the system's "busy level" is determined by using a formula provided by Gazis (2002) [1] as shown in equation (1).

$$\text{Density} = \frac{\text{Volume OUT}}{\text{Average Speed}} \qquad (1)$$

The other factor is the Utility factor, in percentage, which expresses the system's utility. It is evaluated by comparing with the highest density level or system's capacity for each homogeneous situation. Actually, the traffic states on the road section are determined by the simulation settings such as warm-up time and other conditions.

2.3 Experiments

Simulation scenarios are built up and grouped into three main groups according to the simulation's primary purposes. Group 1 considers the case with 100% cars traveling in the system. The number of vehicles entering the system increases up to the saturated point, at approximately 94 vehicles per minute. The second group considers that case with 100% motorbikes operating in the system. Similarly with the previous case, the number of entities entering the system increases up to the saturated point too, at approximately 540 motorbikes per minute. After determining the saturated points in both cases, the last main group is considered under the saturated conditions with mixed traffic systems.

Actually, the ratio between cars and motorbikes is changed by different simulation scenarios, but two methods are used to ensure that the system operates at saturated conditions. In the first way, the number of cars is generated at high levels above the saturated point and then the number of motorbikes entering the system increases. In the second way, the number of motorbikes entering the system is maintained above the saturated point, and the number of cars entering the system increases.

3. Analysis

In the first situation with 100% cars, the system obtains the highest "busy level", at approximately 281.10 cars per kilometer, at which its average speed is reduced to the low level at about 19.05 kilometers per hour. After reaching the saturated point, the system oscillates around it at high "busy levels". The system's serving capacity is shown as equation (2),

$$y_1 = f_1(x) = -8325.8x^6 + 25826x^5 - 28916x^4 + 14251x^3 - 4075.2x^2 + 1797.1x - 25.235$$
$$\forall x \in [0,1],\ y_1 > 0 \qquad (2)$$

where, y_1: The number of motorbikes entering the system per minute, in vehicles per minute.
x: System's utility, in percentage

In the other one with 100% motorbikes, similarly, after determining the system's saturated conditions and system's utility, the interrelation between the Volume IN factor and system utility factor is fitted as equation (3). As shown on the Figure 3, both systems seem stable when their utility factors reach around 60 percent. Before that point, although two factors increase simultaneously, the slope of motorbike curve is larger than another one.

$$y_2 = f_2(x) = -2351.7x^6 + 7823x^5 - 9727.8x^4 + 5569.1x^3 - 1582.5x^2 + 366.32x - 2.829$$
$$\forall x \in [0,1],\ y_2 > 0 \qquad (3)$$

where, y_2: The number of cars entering the system per minute, in vehicles per minute.

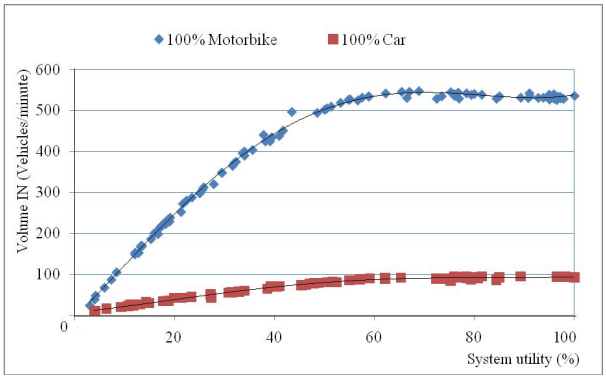


Figure 3: Interrelations of Volume IN Factor and System Utility Factor in Homogeneous Case

Based on the interrelations among the system's factors in both situations, the passenger car equivalent of motorbikes is estimated at each different system utility as in equation (4). The system serves an equivalent quantity between the cars and motorbikes, and the ratio between them at each system utility is concluded from the equations (2), (3) and (4) and shown on Figure 4.

The equivalent factor increases rapidly when the system utility factor is lower than 20 percent. It reaches above 6.5 when the system utility factor ranges from 20 percent to 40 percent. In the other stage, the conversion ratio seems to be slightly reduced,

$$f(x) = \left|\frac{y_1}{y_2}\right| = \left|\frac{f_1(x)}{f_2(x)}\right| \qquad \forall x \in [0,1] \qquad (4)$$

where:

y_1 : The number of motorbikes entering the system per minute, in vehicles per minute.

y_2 : The number of cars entering the system per minute, in vehicles per minute.

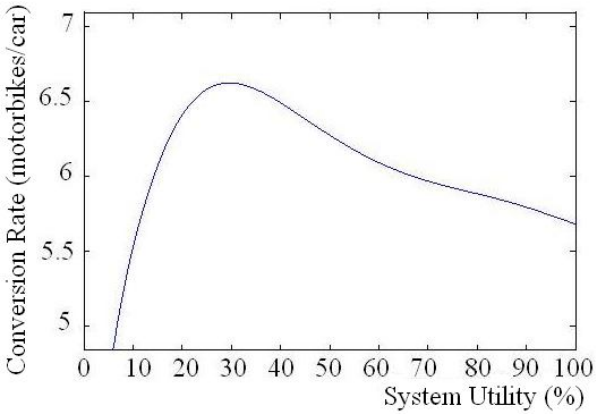


Figure 4: The Passenger Car Equivalent of Motorbikes in Homogeneous Case

In the mixed traffic case, cars and motorbikes travel in the same system, and thus affect together. As mentioned previously, the system is only considered at the high "busy level" and there are two ways to generate the saturated conditions. In the first one, the cars make the system operate at high level and the number of motorbikes is increased. The interrelations between the Volume IN factor and the system utility factor of two populations are determined as shown on Figure 5. Two curves are inversed together because cars and motorbikes travel in the same system. Actually, when the number of motorbikes increases, that of cars decreases. The reduction of cars in quantity and system's utility, simultaneously, follows an exponential distribution as equation (5),

$$y_2 = f_2(x_2) = 1.9894e^{4.0484x_2} \qquad \forall x_2 \in [0,1] \qquad (5)$$

where, x_2 : Car's utility, in percentage

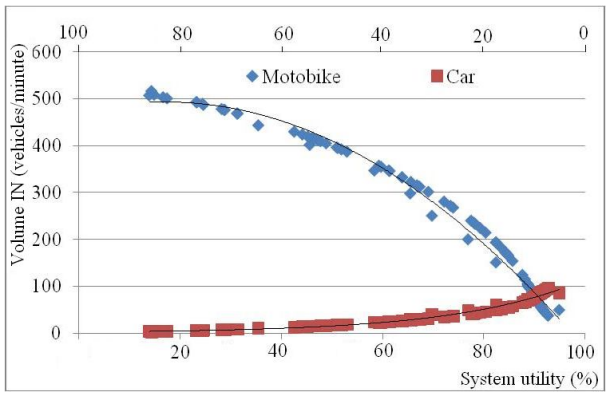


Figure 5: Volume IN –System Utility Relation in a Mixed Traffic System

In addition, the relation of the Volume IN factors between two populations is inverted and fitted as equation (6). It follows a poly-function instead of a linear one as shown on the Figure 6. The changeable ratio between two populations is withdrawn from the equation (6) and shown as equation (7). The passenger car equivalent of motorbikes depends on the system's car utility. It is determined from the equations (5) and (7) and shown on the Figure 7.

$$y_1 = 0.0372y_2^2 - 8.7581y_2 + 533.74 \qquad (6)$$

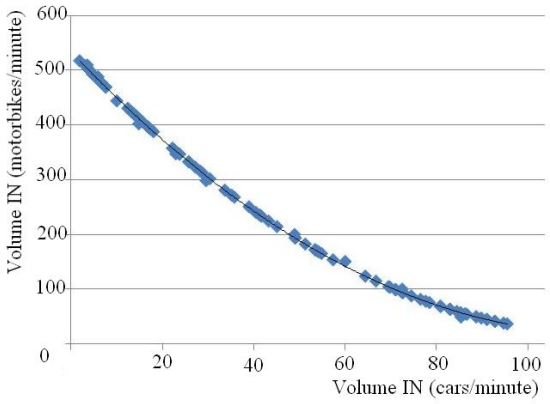


Figure 6: The Volume IN Inversion of Two Populations in the Mixed Traffic System

$$\left|\frac{dy_1}{dy_2}\right| = \left|0.0744y_2 - 8.7581\right| \qquad (7)$$

The conversion rate increases when the car's utility in the system decreases. The curve increases so fast when the car's utility reduces from 100 percent to around 50 percent. In another stage, it slowly increases from around 8 to 8.6.

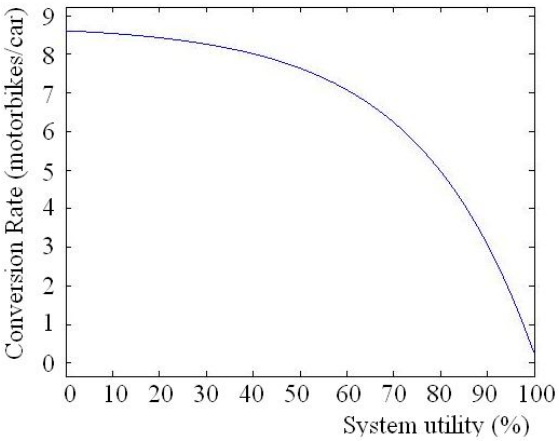


Figure 7: The Passenger Car Equivalent of Motorbikes in Mixed Case with Car Domination

Flexibility, a special characteristic of motorbikes, allows them to travel in unfixed lanes, easily change to a suitable lane, and exhibit high acceleration because small size is its advantage. In addition, car's speed is adjusted by the number of motorbikes traveling in the system, so that the affected level increases synchronously. Therefore, when car's utility does not dominate motorbike's one, lower than 50 percent, the conversion rate increases slightly although the car's utility decreased in the same rate.

When the system's saturated conditions are generated by motorbikes instead of cars, their effects on the system are clearly shown. Although the number of cars entering the system is increased, the other factors of both populations still vary around. The interrelations of Volume IN factor and system utility factor of two populations are shown as on the Figure 8, where two populations are stable in the mixed system with motorbike domination. Each population is shown more clearly in the Figure 9 and Figure 10 for car population and motorbike one in respectively. The conversion rate is determined based on the variations of both populations as shown on the Figure 11. From the simulation results the authors conclude that the passenger car equivalent of motorbikes follows a normal distribution with mean = 23.279 and standard deviation = 1.539.

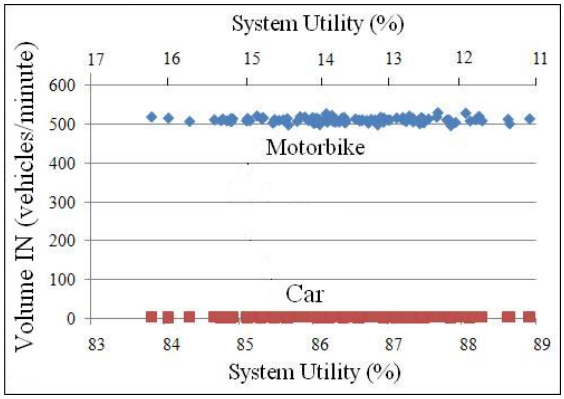


Figure 8: Two Populations in the Mixed System at Saturated Conditions with Motorbike Domination

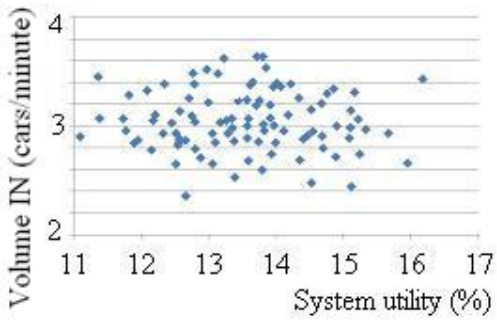


Figure 9: Car Population in the Mixed System at Saturated Conditions with Motorbike Domination

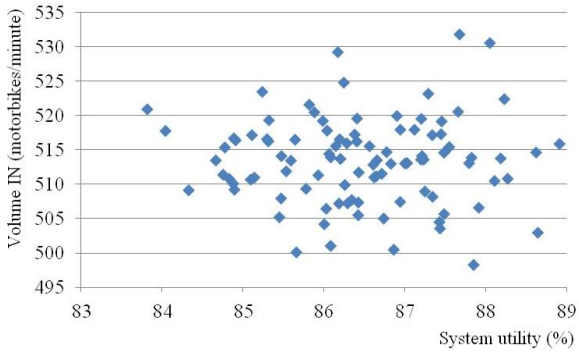


Figure 10: Motorbike population in the mixed system at saturated conditions with motorbike domination

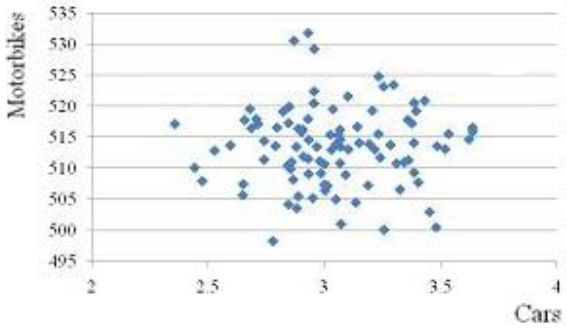


Figure 11: The Passenger Cars Equivalent of Motorbikes in Mixed Case at Saturated Conditions with Motorbike Domination

4. Conclusions

The passenger car equivalent of motorbikes was determined from simulation results. It depends on traffic conditions that whether the traffic system is homogeneous or mixed or how much car's ratio in the mixed system is. The results will be useful for many research applications in developing countries. Mathematical models and useful technologies that facilitate the efficient solution and improvement of traffic problems in developed countries where the car is the main transport mode can be applied in analogous ways.

For example, considering the underdeveloped infrastructure of developing countries and comparing

the ratio between cars and motorbikes on the same system, suitable policies include limiting the number of cars travelling in the traffic system, and replacing the car with the motorbike as the main transport vehicle.

References

[1] D.C. Gazis, "*Traffic Theory*". New York, Boston, Dordrecht, London, Moscow: Kluwer Academic Publishers, 2002. Available at: http://ebooks.kluweronline.com.

[2] F. Balbo, and S. Pinson, "Using intelligent agents for Transportation Regulation Support System design", *Transportation Research Part C,* Elsevier, 2010, pp 140-156.

[3] J. Wang, H. Lu, and H. Peng, "System dynamics model of urban transportation system and its application", *Journal of transportation systems engineering and information technology*, Elsevier, 2008, pp 83-89.

[4] M. Ebben, D. van der Zee, and M. van der Heijden, "Dynamic one-way traffic control in automated transportation systems", *Transportation Research Part B,* Elsevier, 2003, pp. 441-458.

[5] L. E. Owen, Y. Zhang, L. Rao, and G. McHale, "Traffic flow simulation using CORSIM", *Proceedings of the 2000 Winter Simulation Conference*, 2000, pp. 1143-1147. Available at: http://www.wintersim.org/.

[6] Q-L. N. Le, N-H. Do and K-C. Nam, "A simulation model for the mixed traffic system in Vietnam", *International Journal of Simulation and Process Modelling*, Inderscience, 2009, Vol.5, No.3, pp.233-240.

[7] A.M. Law, and W.D. Kelton, *"Simulation Modeling and Analysis*", Singapore: McGraw-Hill, 1999.

[8] W.D. Kelton, R. P. Sadowski, and D. A. Sadowski, *"Simulation with Arena"*, USA: McGraw-Hill, 1998.

[9] H. A. Taha, *"Operations Research: An Introduction"*. USA: Macmillan Publishing Company, 2007.

[10] R. W. Hall, *"Handbook of Transportation Science"*. USA: Kluwer Academic Publishers, 1999.

[11] J. Banks, "*Handbook of Simulation",* USA: John Wiley & Sons, 1998.

[12] H. Bossel, "*Modeling and Simulation*", Germany: A K Peters, 1994.

[13] S. Kara, F. Rugrungruang, and H. Kaebernick, "Simulation modeling of reverse logistics networks", *International journal of production economics*, Elsevier, 2006. Available at: www.elsevier.com.

[14] J. Wahle, O. Annen, Ch. Schuster, L. Neubert, and M. Schreckenberg, "A dynamic route guidance system based on real traffic data", *European journal of operational research,* Elsevier, 2001. Available at: www.elsevier.com.

[15] W. Wen, "A dynamic and automatic traffic light control expert system for solving the road congestion problem", *Expert Systems with Applications,* Elsevier, 2007. Available at: www.elsevier.com.

[16] U. Klein, T. Schulze, and S. Straburger, "Traffic simulation based on the high level architecture", *Proceedings of the 1999 Winter Simulation Conference*, 1999, pages 1095-1103. Available at: http://www.wintersim.org/.

[17] W. Bernhard and P. Portmann, "Traffic simulation of roundabouts in Switzerland", *Proceedings of the 2000 Winter Simulation Conference*, 2000, pp. 1148-1153. Available at: http://www.wintersim.org/.

[18] R. B. Wiley and T. K. Keyser, "Discrete event simulation experiments and geographic information systems in congestion management planning", *Proceedings of the 1998 Winter Simulation Conference*, 1998, pp. 1087-1093. Available at: http://www.wintersim.org/.

[19] M. Lemessi, "An SLX-based microsimulation model for a two-lane road section", *Proceedings of the 2001 Winter Simulation Conference*, 2001, pp. 1064-1071. Available at: http://www.wintersim.org/.

[20] M. P. Hunter, R. M. Fujimoto, W. Suh, and H. K. Kim, "An investigation of real-time dynamic data driven transportation simulation", *Proceedings of the 2006 Winter Simulation Conference*, 2006, pp. 1414-1421. Available at: http://www.wintersim.org/

Biographical notes:

Ngoc-Hien Do is a Doctoral student at the Department of Logistics Engineering, Korea Maritime University, Republic of Korea. He received his BEng in Industrial Systems Engineering from the Ho Chi Minh City University of Technology (HCMUT), Viet Nam, and his MSc in Logistics Engineering from the Korea Maritime University, Korea. He works as a Lecturer and Researcher at the Department of Industrial Systems Engineering Department, HCMUT.

Quynh-Lam Le Ngoc is a lecturer of Department of Industrial Systems Engineering, Ho Chi Minh City University of Technology, Viet Nam. She received her BEng in Electronics Engineering from the Ho Chi Minh City University of Technology (HCMUT), Vietnam, and MEng in Industrial Systems Engineering from the Asian Institute of Technology (AIT), Thailand and PhD in Logistics Department from the Korea Maritime University (KMU), Republic of Korea.

Ki-Chan Nam is a Professor at the Department of Logistics Engineering, Korea Maritime University, Republic of Korea and works as an administrator of the Center for Integrated Logistics Management & Technology Support (CILS). He was also a member of a committee in the Busan Port Authority (BPA). He received his BEng in Navigation from the Korea Maritime University, and his MEng and PhD in Transportation Planning from the University of Wales, UK.

AKAROA2: A CONTROLLER OF DISCRETE-EVENT SIMULATION WHICH EXPLOITS THE DISTRIBUTED COMPUTING RESOURCES OF NETWORKS

Don McNickle
Management Department
University of Canterbury
Private Bag 4800, Christchurch
New Zealand
Don.McNickle@canterbury.ac.nz

Krzysztof Pawlikowski and Greg Ewing
Computer Science and Software Engineering
University of Canterbury
Private Bag 4800, Christchurch
New Zealand
Krys.Pawlikowski@canterbury.ac.nz

KEYWORDS
Discrete-event simulation, sequential simulation, statistical analysis, multiple replications in parallel.

ABSTRACT

This paper describes and summarises our research on enhancing the methodology of automated discrete-event simulation and its implementation in Akaroa2, a controller of such simulation studies. Akaroa2 addresses two major practical issues in the application of stochastic simulation in performance evaluation studies of complex dynamic systems: (i) accuracy of the final results; and (ii) the length of time required to achieve these results. (i) is addressed by running simulations sequentially, with on-line analysis of statistical errors until these reach an acceptably low level. For (ii), Akaroa2 launches multiple copies of a simulation program on networked processors, applying the Multiple Replications in Parallel (MRIP) scenario. In MRIP the processors run independent replications, generating statistically equivalent streams of simulation output data. These data are fed to a global data analyser responsible for analysis of the results and for stopping the simulation. We outline main design issues of Akaroa2, and detail some of the improvements and extensions to this tool over the last 10 years.

INTRODUCTION

Quantitative discrete-event stochastic simulation is a useful tool for studying performance of stochastic dynamic systems, but it can consume much time and computing resources. Even with today's high speed processors, it is common for simulation jobs to take hours or days to complete. Even then, the results may not satisfy the decision-maker's objectives unless a proper experimental framework is set up which guarantees the results with an appropriate degree of accuracy.

Processor speeds are increasing as technology improves, but there are limits to the speed that can be achieved with a single, serial processor. To overcome these limits, parallel or distributed computation is needed. Not only does this speed up the simulation process, in the best case proportionally to the number of processors used, but the reliability of the program can be improved by placing less reliance on individual processors.

One approach to parallel simulation is to divide up the simulation model and simulate parts of it on different processors. However, depending on the nature of the model it can be very difficult to find a way of dividing it up, and if the model does not divide up readily, the overhead of communication between dependent parts of a given simulation model can even make simulation longer. Akaroa2 takes a different approach, applying *multiple replications in parallel* or MRIP. Instead of dividing up the simulation model, multiple independent instances of a given model are executed simultaneously on different processors. These instances continuously send output data (observations) back to a central controller/analyzer, measuring the performance parameters of interest. The central analyzer calculates overall estimates from these observations, e.g. the mean values of the parameters of interest. When it judges that it has enough observations to form all estimates with the required accuracy, it halts the simulation.

Since the simulation replications run independently, n copies of the simulation running on n processors will on average produce observations at n times the rate of a single copy. Therefore final results with an acceptably small statistical error can be produced much faster than from a single instance of the simulation. The total speedup depends on the type of simulation (terminating or steady-state), the type of estimators, and on the method of estimation. A Truncated Amdahl's Law which captures this is formulated in Pawlikowski and McNickle (2001).

MRIP also provides some degree of fault tolerance. It does not matter which instance of the simulation the estimates come from, so if one processor fails, the program it was running can be restarted and the simulation continued without penalty. Alternatively, the simulation will simply continue with one less processor and take proportionately longer to complete. If a simulation is taking an unacceptably long time to complete, additional processors can be called in at any time (note the "Add Engines" button in Figure 3.)

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

PROGRAM ARCHITECTURE

The main components of Akaroa2 are the *akmaster*, the *akslaves*, *akrun* and the *simulation engines*. The relationships between these components are shown in Figure 1.

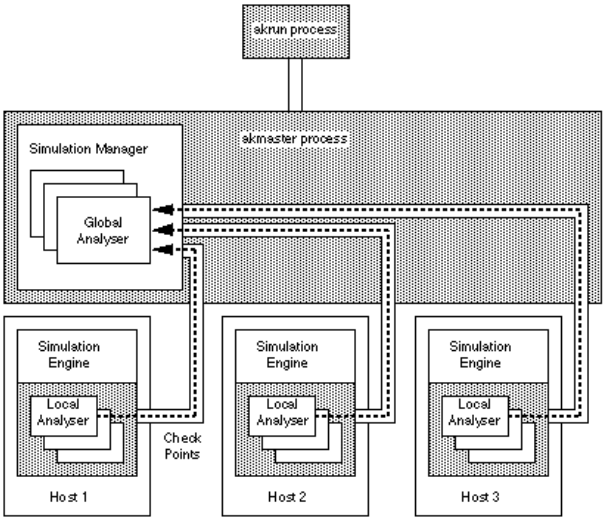


Figure 1. The basic structure of Akaroa2

The *akmaster* process coordinates the activity of all other processes initiated by Akaroa2. It launches new simulations, maintains state information about running simulations, performs global analysis of the data produced by simulation engines, and makes simulation stopping decisions.

Akslave processes (not shown) run on the hosts which are to run the simulation engines. The only function of the akslave is to launch simulation engine(s) on its host as directed by the akmaster. The akslave processes have been introduced because other methods of launching remote processes under UNIX (for example, by using *rsh*) tend to be slow and unreliable.

The *akrun* program is used to initiate a simulation. It first contacts the akmaster process, obtaining its host name and port number from a file left by the akmaster in the user's home directory. For each simulation engine requested, the akmaster chooses a host from among those hosts on the LAN which are running akslave processes. It instructs the akslave on that host to launch an instance of the user's simulation program, passing on any specified arguments. The first time the simulation program calls one of the Akaroa2 library routines, the simulation engine opens a connection to the akmaster process and identifies the simulation to which it belongs, so that the akmaster can associate the connection with the appropriate simulation data structure.

Each engine performs sequential analysis of its own data to form a local estimate of each performance measure. At more or less regularly determined *checkpoints*, the engine sends its local estimates to the akmaster process, where the local estimates of each performance measure from all engines are combined to give a set of *global estimates*.

Whenever a new global estimate is calculated, the relative half-width of its confidence interval at the requested confidence level is computed, and compared with the requested precision. When the precision of all analysed performance measures becomes satisfactory, the akmaster terminates all the simulation engines, and sends the final global estimates to the akrun process, which in turn reports them to the user. Several different simulation experiments may be run simultaneously, using separate *akrun* processes. They may launch instances on the same or different hosts.

In Akaroa2, all interprocess communication is via TCP/IP stream connections, which provide reliable, sequenced, non-duplicated delivery of messages.

An earlier version, Akaroa, used UDP/IP datagrams to communicate between processes. Since the UDP protocol does not guarantee reliable packet delivery, Akaroa spent a great deal of effort attempting to deal with issues of packet loss and duplication. The system was unreliable and difficult to manage. If a process failed to respond within an arbitrary timeout, it was hard to tell whether it had died or was simply taking longer than usual to respond.

The program is implemented in the GNU dialect of C++, and has been tested at the University of Canterbury under the SunOS 4, Solaris 2 and Linux operating systems. Further operating details can be found in the manual, downloadable from the Akaroa2 website at http://www.akaroa2.canterbury.ac.nz/.

The User Interface

Originally, Akaroa2 could only be run by running akrun directly with command-line arguments. Now a graphical user interface *akgui* is provided. The main screen for the akmaster process is shown in Figure 2. This shows a single simulation experiment, being run on 5 engines.

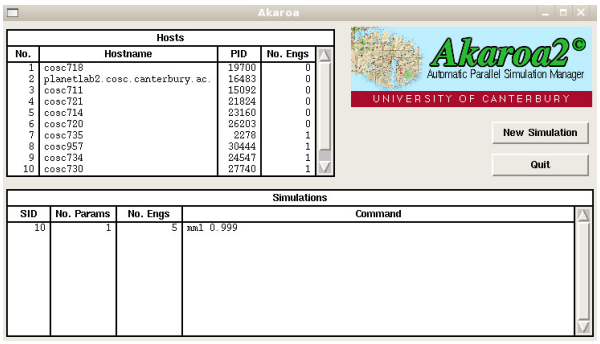


Figure 2. The Main Screen for the Akmaster Process

The progress of each simulation being executed is shown in a progress window (Figure 3). The bar graph shows that the relative precision of (in this case a single) performance measure is currently about 0.2, converging towards the requested level (in this case a 95%

confidence level being estimated to a relative precision of 0.05.)

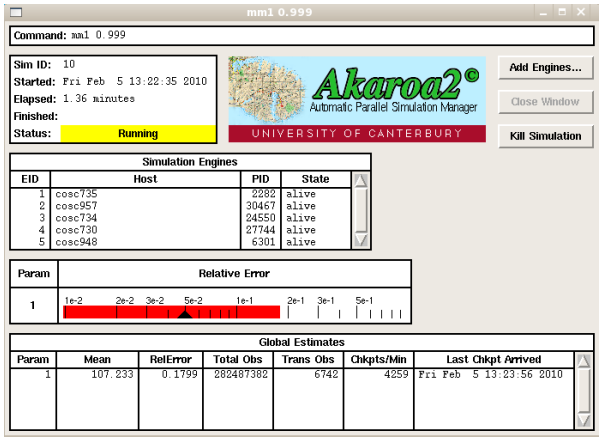


Figure 3. The Akgui Progress Window

IMPROVEMENTS TO INTERFACES AND OPERABILITY

Interfaces with Other Simulation Packages

A popular application of Akaroa is to use it as a controller controlling execution of simulations based on models built with help of other simulation packages. Currently interfaces allow linking Akaroa2 with Ptolemy Classic, the Network Simulator NS2, and OMNeT++. Details of these interfaces can be found from the Akaroa2 website.

Simulation on wide area networks

While Akaroa2 is normally used on a local network, there is nothing in principle to prevent a simulation from being distributed over any set of hosts that can reach each other via the Internet. The communication between processes in an Akaroa2 simulation only requires a very low bandwidth, and is mostly one-directional, so network delays have little influence. In theory, therefore, there should be little difference in performance between using a local, or a wide area network. To test this, Akaroa2 simulations have been run on geographically separated PlanetLab research network nodes (Yasmeen, Ewing, Pawlikowski, and Yamada, 2009). Because of the use of TCP/IP as the communication protocol implementation of this was straightforward. Apart from a slight increase in the time taken to start the simulation up, there was no discernable difference in simulation speed, as predicted.

IMPROVEMENTS TO SIMULATION AND STATISTICAL ASPECTS

The basic approaches to statistical problems of estimation and control, which were adopted in Akaroa2, are described in Pawlikowski (1990). Further details are given in Ewing, Pawlikowski and McNickle (1999). However there have been a number of substantial modifications since these papers.

The Internal Simulation Routines

The simulation library in Akaro2 is now based on a process-interaction approach. This replaces the previous event-scheduling approach, although this remains available. A Process Manager creates processes for each class of entities, describing the life cycle through which these entities go. The class Resource is used to represent finite classes of (usually permanent) entities, to model competition for access to the resource. The class Queue implements a queue of entities of some type, which entities join and leave, with priorities if desired. The usual range of random-variate generators are available. More complex simulations can be written in any language as long as the program is capable of calling routines written in C. Akaroa2 simply picks up, analyses and controls the output observations from this program.

Random Number Generation

In MRIP, each simulation engine must use pseudo-random numbers (PRNs) independent from those used by other engines. Consequently, in Akaroa2, random number generation is not left to the user's simulation program. Instead, the akmaster process is given full control over PRNs used by different simulation engines. Currently Akaroa2 uses a Combined Multiple Recursive PRN Generator with a period of approximately 2^{191}. This sequence is divided into blocks of 2^{128}, and different blocks of PRNs are assigned to different simulation engines. Thus, one could concurrently use up to 2^{67} replications, providing that none of them requires more than 2^{128} PRNs. If fewer replications are used and a simulation engine requires more PRNs, it would receive next block of PRNs. The particular generator used by Akaroa2 is known as MRG32k3a (see, for example L'Ecuyer (1999)). Since it would be very inefficient for a simulation engine to have to communicate with the akmaster process every time it wanted a PRN, each engine generates its assigned block of PRNs by itself, initialising its own copy of the generator from the starting values assigned by the akmaster.

Confidence Interval Estimation

Akaroa2 provides two methods of sequential analysis of output data from steady-state simulation: an automated non-overlapping Batch Means algorithm, and a Spectral Analysis method based on Heidelberger and Welch's (1981, 1983) results. Its implementation in MRIP is described in Ewing, McNickle and Pawlikowski (2002). Some further modifications which improve its performance for sequential analysis are described in McNickle, Ewing and Pawlikowski (2004). Figures 4

and 5 compare the results for simulating an M/M/1 queue using our Batch Means method (Figure 4) and the modified version of Spectral Analysis (Figure 5). What is plotted here is the coverage – that is the actual estimated size of (in this case supposedly 95%) confidence intervals for the mean waiting time, produced by calculating from typically 10,000 separate experiments, the fraction of confidence intervals that actually contain the true mean waiting time. The fall-off in the batch means coverage with load can be entirely explained by correlation between the batch means, (the dashed lines in Figure 4) show the theoretical loss of coverage) which remains a risk with most batch-mean methods. In contrast the coverage from modified Spectral Analysis is almost always at the level specified. Thus our research leads us to strongly favour the Spectral Analysis approach, see also Pawlikowski, McNickle and Ewing, (1998).

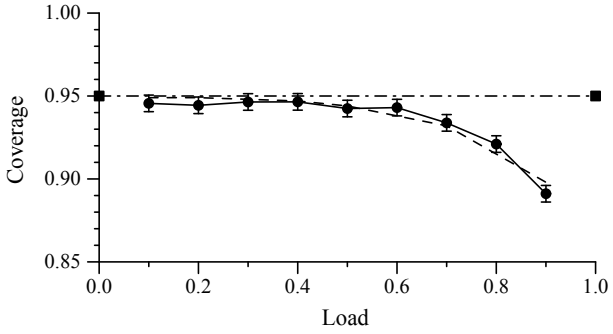


Figure 4. Coverage with Batch Means

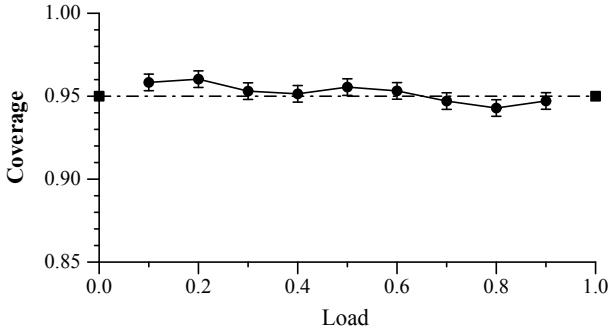


Figure 5. Coverage with Modified Spectral Analysis

An additional attraction of Spectral Analysis is that the same algorithm applies exactly to batched data. This batching of data reduces storage and communications costs.

Transient Period Detection

When Akaroa2 controls a steady-state simulation, an automated method is used to determine the length of the initial transient period. It begins with a heuristic proposed by Gafarian, Anker and Morisaku (1978) to decide when to start testing for stationarity. Its use in a sequential context is described in detail in Pawlikowski. In this heuristic, the length of initial transient period is first taken to be over when the sequence has crossed its running mean 25 times. Then a sequential version of Schruben's test (Schruben, Singh and Tierney, 1983) is used to test for stationarity. If the null hypothesis of stationarity is rejected, the length of the potential transient period is doubled and the test repeated (Pawlikowski). Comparisons with a limited range of other transient deletion methods can be found in Pawlikowski, Stacey and McNickle (1993) which show that this method, although simple, does appear to work well, at least for basic queueing models.

However detecting the length of the transient period remains an uncertain area of discrete-event simulation theory. A large number of methods for selecting the number of observations to delete, and for testing if the system is adequately close to "steady state", have been proposed. Hoad, Robinson and Davies (2008) list 42 methods. Some of these proposals appear to have had limited testing, so their validity remains in question. While our current method appears to be adequate for most simple models, we are considering a sequential version of MSER-5, given its good performance reported in Hoad, Robinson and Davies.

Estimating Quantiles

Mean values provide very limited information about the analysed processes. Much more meaningful insight is provided by quantiles, especially if several quantiles can be estimated simultaneously. For example, 90th or 95th percentiles are often specified by decision makers as the criteria of quality when considering delays or overflow probabilities in manufacturing, customer service, emergency response or telecommunication systems. Estimation of quantiles is yet another order of magnitude harder than estimation of means or variances, as now large amounts of data need to be stored and efficiently sorted. As well as the usual problems due to serial correlation of output, the estimates of multiple quantiles estimated from a single run will also be correlated. Thus, the use of independent replications on this problem is especially attractive.

Lee, Pawlikowski and McNickle (2000) reports on using the existing methods in Akaroa2 to directly estimate quantiles, with reasonable success. However a more advanced method is described below.

Avoiding Premature Stopping

A chronic problem of sequential simulation is that some of the simulation experiments may stop with an insufficient number of observations because, by chance, the required accuracy is apparently temporarily attained. As a result the actual precision of the results obtained is less than specified.

Figure 6 illustrates the problem. It plots the estimated relative precision of a simulation estimating the mean waiting time in a queue, against the (geometric)

checkpoint number. The horizontal lines show stopping criteria of relative precisions of 0.1 and 0.05. Using a relative precision of 0.1 would result in the simulation very nearly stopping at about the 30th checkpoint, whereas data from at least 200 checkpoints are needed. It is also worth noting, as Figure 6 suggests, that this problem is reduced if very high accuracy (low relative precision) is specified – note how the eventual convergence to a relative precision of 0.05 is much less erratic than that to 0.1. High accuracy implies a large amount of data of course, so MRIP has a valuable role in ameliorating this problem by providing large amounts of data, and hence accurate and reliable results, in reasonable elapsed time.

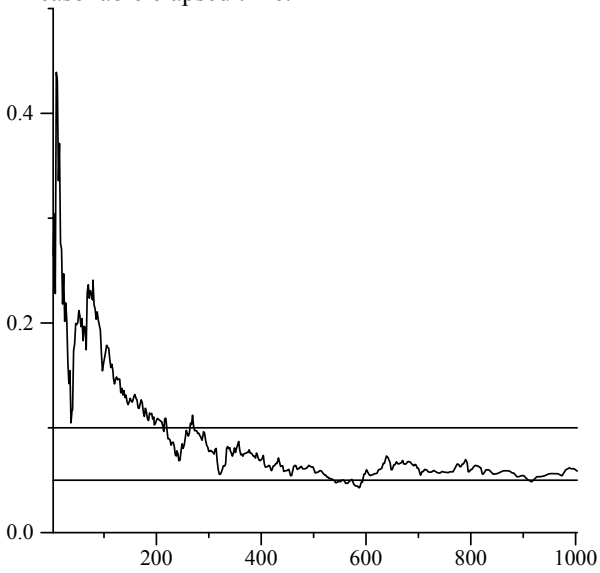


Figure 6. The Risk of Premature Stopping in Sequential Simulation

A recent study (McNickle, Pawlikowski and Ewing, 2010) has suggested that at the run lengths required by common levels of accuracy, the bias in the mean may be negligible. However the coverage can be seriously affected due to prematurely stopped runs. Using a reliable transient deletion technique, and less certainly, initial loading, turns out to help here. The improvements to the Spectral Analysis technique in McNickle, Pawlikowski and Ewing (2004) also help with this problem. In addition a range of simple rules of thumb are described in Lee, Pawlikowski and McNickle (1999) that make use of the multiple replications provided by Akaroa2. The best of these (in its simplest form: take the values from the longest of a number of completely independent runs) has been implemented in the program.

IMPROVEMENTS CURRENTLY IN PROGRESS

Quantile Estimation

In Eickhoff, Pawlikowski and McNickle (2007a) two methods for estimation of multiple quantiles using parallel replications on networks of computers are described (see also Eickhoff, McNickle and Pawlikowski, 2006.) These can be used to simultaneously estimate multiple quantiles, with the set of quantiles to be estimated selected automatically if required. They are in the process of being implemented in a newer version of our simulation controller.

Transient Detection

Almost all methods for detecting the initial transient have been tested on mean values only. Once we move to other measures, demonstrating mean-stationarity may not be enough to determine an appropriate transient period to delete, and methods that demonstrate stationarity in distribution are required. A distribution-based method using the techniques reported in Eickhoff, Pawlikowski and McNickle (2007a), which specifically makes use of parallel replications, is under development.

Estimating Variances

Current analysis of output data from discrete event simulation focuses almost exclusively on the estimation of mean values, largely for reasons of speed, ease of analysis and minimal data storage requirements. However there are a number of applications for which we require the variance – for example jitter in video streams, safety stock in inventory problems, etc. In Schmidt, Pawlikowski and McNickle (2009) three methods of point and interval estimation of the steady-state variance are considered. One, based on splitting of sums of squares, appears to be superior. Estimating variances involves considerably more observations than estimating means. Thus, selecting estimators with good performance characteristics, and using MRIP, is even more important.

CONCLUSIONS

Akaroa2 is downloadable for teaching, and for non-profit research (by universities only) from http://www.akaroa2.canterbury.ac.nz/. It has been downloaded more than 1800 times since the counter was introduced in 2001. Without requiring the use of any parallel programming techniques, it automatically distributes simulation models over an arbitrary number of computers linked by a network, and controls the simulation run length so as to produce final results having a specified precision, both in the case of terminating and steady-state simulation.

REFERENCES

Eickhoff, M., McNickle, D. and Pawlikowski, K. 2006. "Analysis of the Time Evolution of Quantiles in Simulation", *Int. J. Simulation* 7 (6) (2006), 44-55.

Eickhoff, M., Pawlikowski, K. and McNickle, D. 2007a. "Detecting the Duration of Initial Transient in Steady State Simulation of Arbitrary Performance Measures", in

Proceedings ACM ValueTools07 (23-25 October 2007), Nantes, France.

Eickhoff, M., Pawlikowski, K. and McNickle, D. 2007b. "Using Parallel Replications for Sequential Estimation of Multiple Steady State Quantiles", in *Proceedings ACM ValueTools07*, (23-25 October 2007), Nantes, France.

Ewing, G., McNickle, D. and Pawlikowski, K. 2002. "Spectral Analysis for Confidence Interval Estimation under Multiple Replications in Parallel", in *Proc. 14th European Simulation Symposium*, Dresden (October 2002), 52-61.

Ewing, G., Pawlikowski, K. and McNickle, D. 1999. Akaroa2: "Exploiting Network Computing by Distributed Stochastic Simulation" in *Proc. 13th European Simulation Multiconference,* Warsaw, Poland (June 1999), SCSC, 175-81.

Gafarian, A. V., Ancker, C. J. and Morisaku, T. 1978. "Evaluation of Commonly Used Rules for Detecting "Steady State" in Computer Simulation", *Naval Research Logistics Quarterly*, 78 (1978) 511-529.

Heidelberger, P. and Welch, P. D. 1981. "A Spectral Method for Confidence Interval Generation and Run Length Control in Simulations", *Communications of the ACM*, 24(4) (April 1981), 233-245.

Heidelberger, P. and Welch, P. D. 1983 "Simulation Run Length Control in the Presence of an Initial Transient", *Operations Research*, 31 (1983) 1109-1144.

Hoad, K., Robinson, S. and Davies, R. 2008 "Automating Warm-up Length Estimation", in *Proc. 2008 Winter Simulation Conference*, S.J. Mason et al. eds., (2008) 532-540.

L'Ecuyer, P. 1999. "Good Parameters and Implementations for Combined Multiple Recursive Random Number Generators", *Operations Research*, 47, 1, Jan-Feb 1999, 159-164.

Lee, J.-S. R., Pawlikowski, K. and McNickle, D. 1999 "Sequential Steady-State Simulation: Rules of Thumb for Improving the Accuracy of the Final Results", in *Proc. ESS99* (Erlangen, Germany, Oct 26-28 1999), 618-622.

Lee, J.-S. R., Pawlikowski K. and McNickle, D. 2000 "Initial Transient Period Detection for Steady-State Quantile Estimation", in *Proc. Summer Computer Simulation Conference* SCSC'2000, Vancouver, Canada, International Society for Computer Simulation, San Diego, July 16-20, 2000, Paper #S213, 1-6

McNickle, D., Pawlikowski, K. and Ewing, G. 2004 "Refining Spectral Analysis for Confidence Interval Estimation in Sequential Simulation" in *Proceedings of the ESS2004*, Budapest, Hungary Oct 2004, 99-103.

McNickle, D., Ewing, G. and Pawlikowski, K. 2010. "Some Effects of Transient Deletion on Sequential Steady-State Simulation", to appear in *Simulation Modelling Practice and Theory*, paper SIMPAT864.

Pawlikowski K., Stacey, C. and McNickle, D. 1993. "Detection and Significance of the Initial Transient Period in Quantitative Steady-State Simulation", in *Proc. Eighth Australian Teletraffic Research Seminar*, RMIT Melbourne, (6-8 December 1993) 193-202.

Pawlikowski, K., McNickle, D. and Ewing, G. 1998. "Coverage of Confidence Intervals from Sequential Steady-State Simulation", *Simulation Practice and Theory*, 6 (1998), 255-267.

Pawlikowski. K. And McNickle, D. 2001. "Speeding up Stochastic Discrete-Event Simulation", in *Proc. European Simulation Symposium, ESS'01*, Marseille, France, 18-20 October, ISCS Press, 132-138.

Pawlikowski, K., Schoo, M. and McNickle, D. 2006. "Modern Generators of Multiple Streams of Pseudo-Random Numbers", in *Proc. Int. Mediterranean Modelling Multiconference* (ESM06), Barcelona, 553-559.

Pawlikowski, K. 1990. "Steady State Simulation of Queueing Processes: a Survey of Problems and Solutions", *ACM Computing Surveys*, (22, June 1990) 123-170.

Schmidt, A., Pawlikowski, K. and McNickle, D. 2009. "Sequential Estimation of the Steady-State Variance in Discrete Event Simulation", in *Proc. ECMS 2009*, 630-635.

Schruben, L., Singh, H. and Tierney, L. 1983. "Optimal Tests for Initialisation Bias in Simulation Output", *Operations Research*, 31(6) (1983) 1167-1178.

Yasmeen, F., Ewing, G., Pawlikowski, K. and Yamada, S. 2009 "Distributing Akaroa2 on PlanetLab". Proceedings of IEICE General Conference, Matsuyama City, Japan, March 17-20, 2009.

AUTHOR BIOGRAPHIES

DON MCNICKLE is an Associate Professor of Management Science in the Department of Management at the University of Canterbury. His research interests include queueing theory; networks of queues and statistical aspects of stochastic simulation. He is a full member of INFORMS.

GREG EWING is an Adjunct Research Associate in the Department of Computer Science and Software Engineering at Canterbury; where received a Ph.D. His research interests include simulation; distributed systems; programming languages, 3D graphics and graphical user interfaces. He has made contributions to the Python programming language.

KRZYSZTOF PAWLIKOWSKI is a Professor of Computer Science at the University of Canterbury. His research interests include quantitative stochastic simulation; and performance modelling of telecommunication networks. He received a PhD in Computer Engineering from the Technical University of Gdansk, Poland. He is a Senior Member of IEEE and a member of SCS and ACM. His web page is <http://www.cosc.canterbury. ac.nz/"'krys/>.

ACKNOWLEDGEMENT

This research was supported in part by grants from the College of Business and Economics, and the College of Engineering, University of Canterbury.

APPLICATION OF LOW-COST COMMERCIAL OFF-THE-SHELF (COTS) PRODUCTS IN THE DEVELOPMENT OF HUMAN-ROBOT INTERACTIONS

Ottar L. Osen
Helge T. Kristiansen
Webjørn Rekdalsbakken
Department of Information and Communication Technology
Aalesund University College
N-6025 Aalesund, Norway
E-mail: oo@hials.no

KEYWORDS

Human-robot interaction, game technology, cell phone interface, Wii Remote, iPhone.

ABSTRACT

In the effort of developing sensible ways for interaction between humans and automated equipment the use of commercial off-the-shelf (COTS) products is shown to be fruitful in the learning process. The development in the field of consumer electronics has lead to increasingly more elaborate facilities for interaction with the human user. Modern cell phones and game technology are typical representatives of this trend. In this work such equipment has been explored in the aim to achieve easy and natural interfaces between humans and

INTRODUCTION

This document presents a work on human-robot interaction and communication performed at Aalesund University College (AUC) among lecturers and students at the Bachelor of engineering program in Cybernetics. The work is integrated in the main topics of the program. A central part of this education is represented by a broad use of student projects in close cooperation with the scientific staff. The students get the means and support to investigate commercially available electronics and computer equipment related to automatic systems and autonomous devices. AUC is situated in one of Norway's most vital industrial areas with a strong cluster of ship builders, and a variety of producers of all kinds of ship equipment. This situation has resulted in a strong collaboration between AUC and many of these companies. This cooperation has given offspring to many new products for use onboard ships, examples of which are 3DOF and 6DOF motion stabilized platforms, heave-compensated winches, and remotely controlled tracking systems for light sources and cameras. Two of the companies with which AUC has the closest contact, are Rolls Royce Marine Dept. Inc. (RR) and Offshore Simulation Center AS (OSC). These two companies maintain an extensive activity in the development of ship simulators for a variety of purposes, including high-speed craft simulators, and simulators for anchor handling and crane and winch operations. OSC is today an international supplier of the most modern simulators. In this development AUC has played an important role both in mathematical modeling and in program construction (Rekdalsbakken 2006) (Rekdalsbakken 2007). An important aspect of such simulators is of course the human interface. It has to be as realistic as possible, and at the same time not too expensive. This is the background for the current work on exploring COTS products as part of the human interface towards the advanced operating equipment onboard ships (Rekdalsbakken and Osen 2009). The two main products explored in this context are the Wii Remote controller from Nintendo and the iPhone from Apple. The experiments have been performed on small-scale models of remotely controlled vehicles and include a total communication chain from the human operator to sensors and actuators installed on the vehicles. It is the intention that these experiments shall give insight into new developments in the field of human-robot interactions, which may lead to the integration of said technologies into products that are used in real user interfaces onboard ships.

INTERFACES

The Wii Remote Controller

The Wii Remote controller from Nintendo (2010) was developed to give the player a new dimension of interactivity with games. It may be said that it has revolutionized the human feeling of being part of the game context. The Wii Remote interface consists of several parts; a 3-axis accelerometer, an infrared camera, vibration function, and several buttons for input signals. It communicates over a wireless connection using the Bluetooth (Bluetooth 2010) protocol. The Wii Remote is perhaps the best example of gadget that has become a universal control tool with a wide range of applications.

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

The iPhone

The iPhone from Apple (2010a) has set a new trend in modern cell phone development. It is a stand-alone device for telephone and general internet communication and services. It also offers a lot of auxiliary tools and facilities, including GPS and a 3-axes accelerometer.

From a HMI perspective the iPhone is very interesting since it has a high resolution color multi touch screen. Although the idea of multi touch dates back to mid 80's (Lee, Buxton and Smith 1985) multi touch technology is still in its infancy and usage patterns established amongst users of mass market devices such as iPhone will impact the future of multi touch technology. Multi touch is probably best known to the public through Sci-Fi TV series and movies such as "Minority Report", but the growing market penetration of products such as iPod, iPhone and other smart phones are rapidly making multi touch a part of everyday life for millions. Multi touch is subject to substantial research activity and IC manufacturers are inventing new multi touch devices such as Atmel's (2010) maXTouch technology.

The iPhone communicates on common standards like wireless local area networks (Wi-Fi), Bluetooth and USB. With its comprehensive set of interfacing facilities the iPhone represents a great potential as a tool for human-robot interactions.

Communication standards

Several kinds of communication standards are involved in these experiments on human-robot interactions. The handheld devices for consumer applications are equipped with wireless interfaces like Wi-Fi and Bluetooth. These facilities have been utilized in this work in addition to infrared remote control and USB connection. In the communication chain between the different parts in the experiments Wi-Fi and radio communication (RC) have been used in addition to cabled Ethernet.

Controllers

In these experiments several kinds of embedded microcontrollers have been explored. Especially three controllers have been much used, because of their qualities regarding communication, concurrent programming, and the diversity of interfacing possibilities. These controllers are the Muvium (2005) SBC65EC SBC from Modtronix, the Javelin Stamp Demo Board (JSDB) from Parallax Inc (2006) and the Arduino (2009) Duemilanove ATmega328. These controllers have been built into the test equipment, i.e. the remotely controlled vehicles and the communication links, and perform the communication tasks in addition to controlling the motion of the vehicles and their peripherals.

HUMAN – ROBOT INTERACTION

Wii Remote and Bluetooth

Bluetooth has become a very popular wireless communication standard, especially among producers of consumer electronics. Bluetooth follows the IEEE 802.15 standard for Wireless Personal Area Network (WPAN) and operates in the 2.4 GHz frequency band. Power class 2 devices (which is the most common) has a maximum sending power of 2.5 mW and has a reach of 10 meters. For the other 2 power classes 1 and 3 the ranges are 100 meters and 1 meter respectively and maximum sending powers are 100mW and 1 mW. The Bluetooth architecture uses different protocols and gives a lot of possibilities for interconnection of different kinds of electronic units. How devices use Bluetooth is standardized in so called profiles. The profiles specify dependencies, interface formats and protocol stack. There are profiles for images, printing, ISDN, fax, file transfer, hands free, headset and so on. The Wii Remote controller uses Bluetooth in its communication with other devices. The profile used is called Human Interface Device Profile (HID). HID is used for mice, joysticks, keyboards and similar devices. HID is designed for low latency. In addition the Wii Remote controller includes an infrared (IR) camera. The camera is used to decide the position of the Wii controller relative to two fixed IR beacons (known to Wii users as the sensor bar that resides on top of their TV set, but which in fact is just 2 IR diodes placed in each end of the bar). The sensor (camera) is inside the Wii Remote.

Wi-Fi Communication and the iPhone

Wi-Fi is a common term for wireless local area networks based on the IEEE 802.11 standard (IEEE 2007). To use a Wi-Fi one needs a device with a wireless network card and a wireless access point. The iPhone has implemented this network as an interface in its human-machine interconnections. This interface is used to collect the information from the iPhone on a PC for use in the control of a remote vehicle.

VEHICLE COMMUNICATION

Wi-Fi Network

For the remote communication between the central PC and the performing equipment wireless connections have been used. The Wi-Fi network for communication with the remote vehicles is implemented on the vehicle by use of a Nano WiReach (Connect One 2010) serial to Wi-Fi bridge. This very small component is compatible with the wireless network standard 802.11b/g and can roam between different networks. The WiReach is based on the chip sets iChip CO2144 IP and Marwell 88W8686. There is no need for software drivers to use this component, and it is easily configured over a serial communication from a terminal program.

Radio Communication

Another communication standard used in these experiments is the traditional 433 MHz radio connection. This is established by the use of an RF transceiver of the type ER400TRS from Easy-Radio (LPRS 2005). This is a half duplex FM (FSK) radio transceiver with 10 mW sending power and a typical range of 250 meters. The component functions as a transparent serial communication link. A 50 ohms antenna was mounted on the transceiver output port to improve signal reach.

SYSTEM INTEGRATION

Connecting the Pieces

The overall purpose of this work has been to purchase and test the COTS products necessary to build and operate remotely controlled automatic equipment. These devices are characterized by a chain of operations reaching from human interactions, through wireless communication, to automatic control and data acquisition.

Software Integration

Several programming tools and libraries have been explored in this investigation. The programming languages used have been Java and C++, with different IDEs adapted to the different controllers and situations. An aim has been to find free open source software, which could be modified and further developed to suit the specific needs of the operations.

WII REMOTE AS A HUMAN-ROBOT INTERFACE

Wii Remote as Interface

In this project a Wii Remote control was used as interface for controlling the motion and functions of a radio controlled (RC) car. The Wii Remote has been developed to become one of the most popular and versatile game interfaces, and its applications in all kinds of games are increasing continuously. Other console gaming platforms such as Sony's PS3 and Microsoft Xbox 360 have lately released their wireless remote counterparts. Recently these devices have found a segment as a more general interface towards other equipment, e. g. as an interface towards robots and other automatic machinery. National Instruments has released a LabView interface for the Wii Remote. In this application the 3-axis accelerometer of the Wii Remote is used as control input for the car. The Wii Remote communicates with a local PC through the Bluetooth interface.

Wii Remote Accelerometer

The Wii Remote uses the popular ADXL330 accelerometer from Analog Devices (2007). This accelerometer measures three axes with a range of +/- 3g. Since the device can measure the static acceleration of gravity it can be used as a tilt sensor in addition to measure motion.

Software and Communication

Some very useful software packages have been developed for the purpose of collecting the Wii Remote signals sent over the Bluetooth connection. Two of these packages have been explored; the Java package WiiuseJ (2009), and a Visual Studio program under Creative Common License written by Edgar Barranco (2009). The last program uses a .NET library for interaction with the Wii Remote and creates a GUI for reading the input signals. Both of these systems are open source programs. With the WiiuseJ package one can set up a WiimoteListener as an interface program towards the Wii Remote. The WiimoteListener is implemented as an interface in Java, and the methods to be used must be overrided by the application programmer. With basis in these program packages GUIs were created both in Java and in Visual Studio's C# for collecting the Wii Remote signals and forwarding them in an adequate format to the remote vehicle.

As mentioned earlier the Wii Remote uses the HID profile. Unfortunately it does not conform to the standard data types and HID descriptor. This calls for non standard drivers that have been developed by members of several internet user groups. Since Nintendo does not document this protocol the developers of said drivers must use reverse engineering to understand how the Wii Remote communicates (WiiBrew 2009). In 2009 Nintendo released an add-on to the Wii Remote called Wii MotionPluss which combined with the built in accelerometer increases the accuracy and enables tracking of more complex movements. The Wii MotionPlus add-on is a tuning fork gyroscope implemented as a MEMS (micro-electrical-mechanical system) produced by InvenSense (2008). The IDG-650 gyroscope (InvenSense 2009) measures angular velocity in the pitch/roll (X/Y) axis and ISZ-650 in the Yaw (Z) axis, see figure 1.

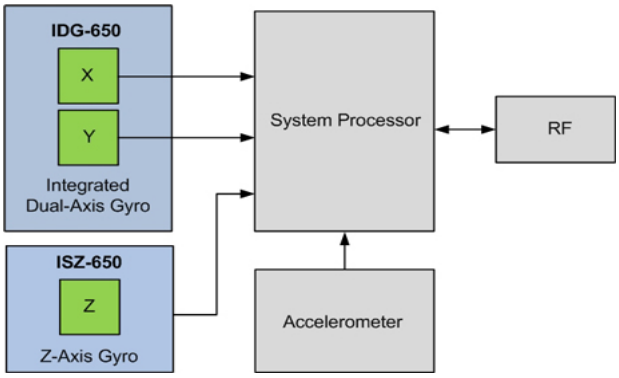


Figure 1: Game Controller System Diagram

Practical Implementations

The primary goal of this application has been to control the motion of the remotely operated car, i.e. speed and direction, by use of a handheld Wii Remote controller. In addition the application has been extended through implementation of control signals from the Wii Remote push buttons to the car. The car will be used for remote data acquisition from sensors mounted onboard the car. The communication from the PC to the remote car was implemented in two different ways, either by a local Wi-Fi network or by radio communication. The implementation of the radio link was performed by the use of two Arduino ATmega328 microcontrollers, one as a transmitter link and one as a receiver onboard the car, see Figure 2 for the transmitter device. The transmitter reads serial data from the host PC over a USB port and transfers the data to the ER400TRS radio component. This communication is fully transparent as the TX and RX pins on the Arduino microcontroller are connected both to the USB port and the ER400TRS. Thus all the data traffic goes through both these channels. It is only necessary to choose the direction of communication for the ER400TRS. In this way the data is also returned to the PC and is used as a control of the communication line. As an alternative to the Arduino microcontroller the Muvium SBC65EC microcontroller was also used in this experiment. A wired TCP/IP connection was established to the SBC65EC microcontroller. By use of the ER400TRS RC transceiver communication was established and another SBC65EC microcontroller on the vehicle. This is shown in Figure 3. Communication over the Wi-Fi network was set up directly from the PC to a Muvium SBC65EC microcontroller. This controller is furnished with a 10 Mbs Ethernet connection with a TCP/IP stack supporting socket connections for TCP and UDP. It has, however, not support for Wi-Fi communication. It was therefore decided to provide the controller with a Nano WiReach LAN to Wi-Fi Bridge. This component connects to the Ethernet port and supports wireless networks of type 802.11b/g. In this way the Muvium SBC65EC can communicate directly with a PC over a Wi-Fi network.

Program Architecture

The PC application is implemented as a program with three major classes. In Java these are realized as concurrent threads. The three classes are called *ComputerApp*, *DataSender* and *WiiMoteHandler*. The *ComputerApp* class contains the *main()* method and instantiates objects of the two other classes, which are run as independent threads. *ComputerApp* then generates a GUI for user interaction through the PC, and also establishes the communication channel to the Wi-Fi network on the given IP address and port number. A socket class connects to the given IP and port address and supports the methods for communicating on the associated input and output streams. The *WiiMoteHandler* class takes care of the communication against the Wii Remote controller over the Bluetooth transceiver on the PC. This is done by implementing the WiiMoteListener interface in the WiiuseJ library. The methods in this interface have to be overwritten. The vital method in this context is *onMotionSensingEvent()* which reads motion data from the Wii Remote's accelerometers. These motion values are verified and properly scaled before being transferred to public variables in the *ComputerApp* object. The *DataSender* class has as its responsibility to continually check for updates of these variables and transmit new incoming values to the Wi-Fi output stream. The relation between the three classes is shown in Figure 4.

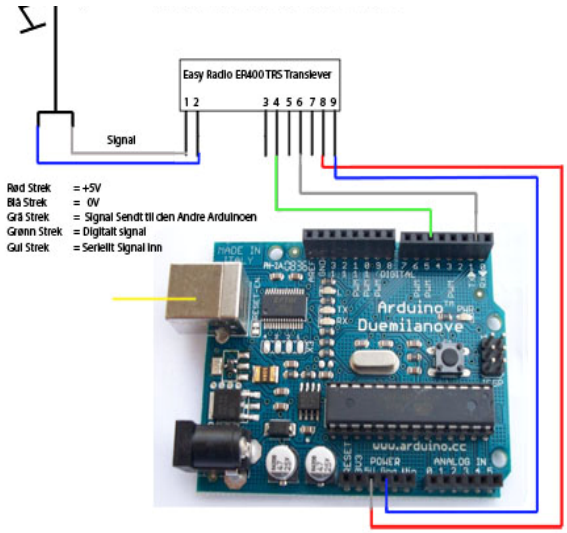


Figure 2: Arduino ATmega328 used as a RC communication link

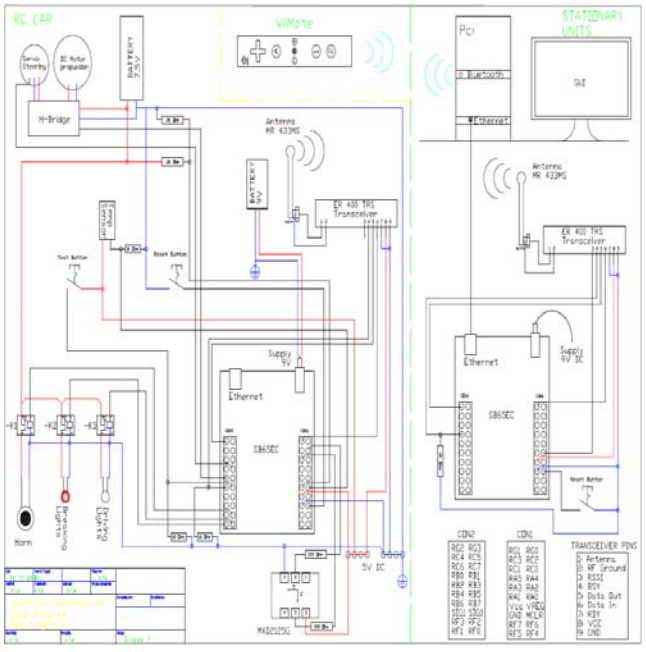


Figure 3: Electronic drawing of RC communication using SBC65EC

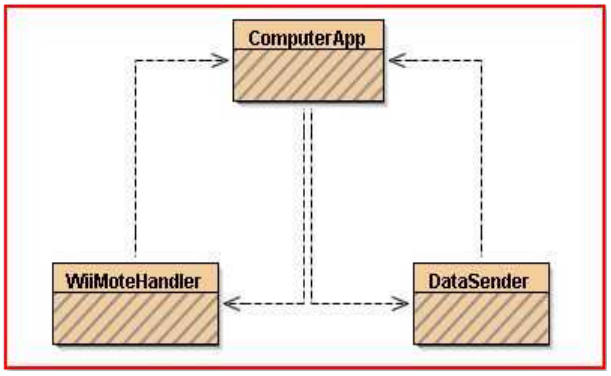


Figure 4: Class relations of the PC application

The application run on the Muvium SBC65EC microcontroller also consists of three classes; *SocketServo*, *MultiServoController* and *ServoService*. *SocketServo* runs as the main thread by inheriting from the native UVMRunnable class. First it establishes a socket connection to the Wi-Fi network on a given IP address and port, and then it instantiates objects of the two other classes. *MultiServoController* provides the methods necessary for sending PWM signals to the Muvium ports. The *ServoService* object is enabled to inherit from the native class SocketService by use of the factory method connection.addSocketService (*ServoService)*. *ServoService* implements the SerialPortListener interface and creates a Listener on the Wi-Fi connection port. This gives access to the method serialEvent(SerialPortEvent serialEvent), which starts when incoming packages arrives at the connection. These packages consist of only two bytes giving the output port number and percentage of signal value. *ServoService* forwards these values to the *MultiServoController* object for port handling.

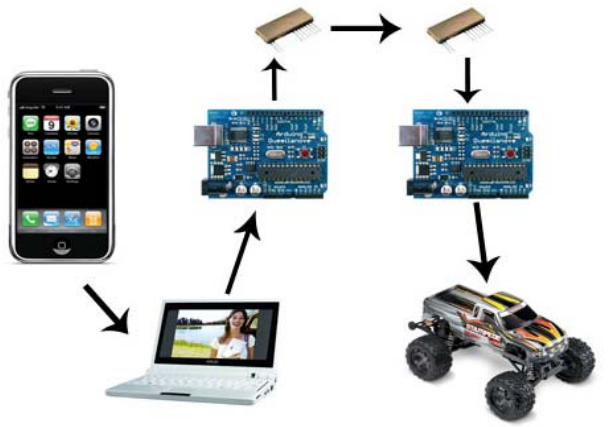


Figure 5: The communication chain.

IPHONE AS HUMAN-ROBOT INTERFACE

iPhone as Interface

In this project an iPhone was used as interface for controlling the motion and functions of a car over a wireless connection. The iPhone communicates with a PC over a local Wi-Fi network. The necessary control signals come from the 3-axes accelerometer readings on the iPhone. These data are further transmitted to the car over a wireless connection. Figure 5 shows the communication chain.

iPhone accelerometers

The iPhone uses a LIS302DL 3-axis accelerometer from STMicroelectronics (2008) with I^2C bus. This accelerometer is also able to measure static acceleration such as gravity and may therefore be used as a tilt sensor in addition to motion detection.

Software and Communication

A growing amount of public software is available for the iPhone. The iPhone has a great potential as an interface towards other equipment and this segment is still quite new and poorly investigated. Only a few Bluetooth profiles are supported (Apple 2009), hence there is a great demand for more Bluetooth support on the iPhone. As a consequence so far the focus has been on using iPhones Wi-Fi interface.

Two program systems have been used in this work; one is the OSCemote (Minor 2009) remote control application, the other is the Max5 communication package (Cycling ’74 2010). OSCemote runs on the iPhone, while Max5 is installed on the host PC. With the OSCemote program one can transmit Open Sound Control (OSC) messages over the Wi-Fi network (Freed and Schmeder 2009). OSC is a modern net based offspring of MIDI that can be used for serial communication. OSC is used for transmitting accelerometer readings from the iPhone. The Max5 program collects the data from the Wi-Fi input. It receives and unpacks UDP packages sent by OSCemote containing the accelerometer data, scales the numerical values and transmits them over a serial line to an Arduino ATmega328 microcontroller for further RC communication. Alternatively the communication goes directly over a Wi-Fi network from the PC to a Muvium SBC65EC microcontroller using a Nano WiReach LAN to Wi-Fi bridge.

The OSCemote software was selected since it is available through Apples App Store, but other software alternatives do exist and more will probably come. One interesting possibility is to write one’s own software, to facilitate that one needs the iPhone SDK which is available at Apples developer site (Apple 2010b). This approach would allow usage of industrially acknowledged protocols such as Modbus (2010) TCP.

Practical Implementations

The Max5 program collects the data from the WiFi input port. It receives and unpacks UDP packages sent by OSCemote containing the accelerometer data, scales the numerical values and transmits them over a serial line to an Arduino ATmega328 microcontroller. The ATmega328 controller operates as a transparent serial link for wireless RC communication. Another ATmega328 embedded into the remotely controlled car, receives and effectuates the radio signals. As an alternative the communication line was also realized over a WiFi network from the PC to a Muvium SBC65EC microcontroller furnished with a Nano WiReach LAN to WiFi bridge.

CONCLUSION

The primary goal of this work was to control the motion of a remote car, i.e. speed and direction, by use of commercially available handheld devices like the Wii Remote and iPhone over a wireless radio connection. However, this is only the basis for further utilizations of the car, as a tool for exploring its environment. Experiments with cameras are currently being performed and other types of sensory devices will be investigated. In this way the possibilities of using the Wii Remote and iPhone as two way communication interfaces towards remotely controlled equipment, will be investigated.

The experiments of using consumer electronics as interfaces in interactions with automatic equipment have shown that these devices have the capability to satisfy the demands raised in such situations. Many of these devices are growing more and more versatile with a comprehensive set of possibilities for easy and safe interaction with other equipment. The lack of documentation from equipment vendors may be a challenge, but that is often solved by the internet community using reverse engineering.

Due to the vast quantities these devices are produced in the prices are very low. Analysts predict sales of Wii MotionPlus at ten million units in US and Europe, and in Japan it sold something like 600 000 units the first week. The large production quantities does not only result in low cost, it also improves component quality (poor quality would have serious consequences for supplier due to the large amount of units) and finally there will be a large amount of users trying to stretch the technology to its limits and coming up with great new applications.

As a summary this work has revealed the great potential inherited in this new consumer electronic devices. The experiments encourages further work towards the implementation and use of such devices as human-robot interfaces in real ship operations.

ACKNOWLEDGMENT

We will like to express our grateful thanks to our hardworking and clever last year students in cybernetics.

REFERENCES

Analog Devices 2007. “Small, Low Power, 3-Axis ±3 g *i*MEMS Accelerometer”. Analog Devices, Inc. http://www.analog.com/static/imported-files/data_sheets/ADXL330.pdf

Apple 2009. “iPhone and iPod touch: Supported Bluetooth profiles”. Apple Inc. http://support.apple.com/kb/HT3647

Apple 2010a. “iPhone”. Apple Inc. http://www.apple.com/iphone/

Apple 2010b. “iPhone SDK 3.1.3”. iPhone Dev Center, Apple Inc. http://developer.apple.com/iphone/index.action

Arduino 2009. “Arduino open-source electronics prototyping platform”. http://www.arduino.cc

Atmel 2010. “maXTouch the worlds leading touch technology”. Atmel Corporation. http://www.atmel.com/products/touchscreens/default.asp

Bluetooth 2010. “Bluetooth Basics”. Bluetooth SIG, Inc. http://www.bluetooth.com/Bluetooth/Technology/

Barranco, E. 2009. “Wii Mote + Arduino + RC Car”. http://r00t-ed.homeip.net/projects.php

Connect One 2010. “Nano WiReach wireless LAN bridge”. Connect One Ltd. http://www.connectone.com/products.asp?did=73&pid=80

Cycling ’74 2010. “Max, Interactive visual programming environment for music, audio, and media. Cycling ’74. http://cycling74.com/products/maxmspjitter/

IEEE 2007. “Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications”. IEEE Computer Society, IEEE Std 802.11-2007, Revision of IEEE Std 802.11-1999.

Freed, A. and Schmeder, A. 2009. “Features and Future of Open Sound Control version 1.1 for NIME”. Conference Paper at NIME09. http://cnmat.berkeley.edu/node/7002

InvenSense 2008. “INVENSENSE IDG-600 Motion Sensing solution showcased in Nintendo’s new Wii Motion Plus Accessory”. InvenSense, Inc. http://invensense.com/mems/gyro/documents/articles/071508.html

InvenSense 2009. “MEMS Gyroscopes for Gaming”. InvenSense, Inc. http://invensense.com/mems/gaming.html

Lee, SK., Buxton, W. and Smith, K.C 1985. “A multi-touch three dimensional touch-sensitive tablet”. CHI’85 Proceedings April 1985, 21-25, ACM 0-89791-149-0/85/004/0021 http://www.billbuxton.com/leebuxtonsmith.pdf

LPRS 2005. “ER400TRS Transceiver”. Low Power Radio Solutions Ltd. http://www.lprs.co.uk/datasheets/ER400TS-02.pdf

Minor 2009. “OSCemote – iPhone Application. Open sound control in the palm of your hand”. http://pixelverse.org/iphone/oscemote/

Modbus 2010. “Modbus TCP Toolkit”. Modbus Organization, Inc. http://www.modbus.org/toolkit.php

Muvium 2005. “µVMDK Muvium Device Development Kit.” http://www.muvium.com

Nintendo 2010. “Wii Controllers”. Nintendo. http://www.nintendo.com/wii/what/controllers

Parallax Inc. 2006. “Javelin Stamp manual version 1.0a” http://www.parallax.com/Portals/0/Downloads/docs/prod/javelin/JavelinStampMan1-0.pdf

Rekdalsbakken, W. 2006. “Design and Application of a Motion Platform for a High-Speed Craft Simulator.” Proceedings of ICM 2006, IEEE 3rd International Conference on Mechatronics, 38-43, IEEE Catalog Number of Printed Proceedings: 06EX1432.

Rekdalsbakken, W. 2007. "The Use og Artificial Intelligence in Controlling a 6DOF Motion Platform." Proceedings of ECMS 2007, the 21th European Conference on Modelling and Simulation, 249-254. European Council for Modelling and Simulation, printed ISBN: 978-0-9553018-2-7, CD ISBN: 978-0-9553018-3-4.

Rekdalsbakken, W and Osen, O. L. 2009. "Teaching embedded control using concurrent Java" *Proceedings of ECMS 2009, the 23th European Conference on Modelling and Simulation*. European Council for Modelling and Simulation, printed ISBN: 978-0-9553018-8-9, CD ISBN: 978-0-9553018-9-6.

STMicroelectronics 2008. “LIS302DL MEMS motion sensor” http://www.st.com/stonline/products/literature/ds/12726.pdf

WiiBrew 2009. “WiiMote Accelerometer”. http://wiibrew.org/wiki/Wiimote#Accelerometer

WiiuseJ 2009. “WiiuseJ, Java API for Wiimotes” http://code.google.com/p/wiiusej/

AUTHOR BIBLIOGRAPHY

OTTAR L. OSEN is MSc. in Cybernetics from the Norwegian Institute of Technology in 1991. He is a senior instrument engineer at Offshore Simulation Centre AS, and assistant professor at Aalesund University College.

HELGE T. KRISTIANSEN is MSc. in Electrical engineering from the Norwegian Institute of Technology in 1992 and Master of Information Technology from Aalborg University in 2003. He is an assistant professor at Aalesund University College.

WEBJØRN REKDALSBAKKEN got his MSc. in Physics at The Norwegian Institute of Technology in 1977. He has been rector at Aalesund Engineering College, and is now assoc. professor and leader of the BSc. programme in Automation at Aalesund University College.

ELECTRICALLY DRIVEN AND CONTROLLED LANDING GEAR FOR UAV UP TO 100KG OF TAKE OFF MASS

Zbigniew SKORUPKA, Wojciech KOWALSKI and Rafał KAJKA
Institute of Aviation
02-256 Warsaw, Poland
E-mail: zbigniew.skorupka@ilot.edu.pl
rafal.kajka@ilot.edu.pl
wojciech.kowalski@ilot.edu.pl

KEYWORDS
UAV, Unmanned Aerial Vehicle, landing gear, IA.

ABSTRACT

UAV - Unmanned Aerial Vehicle is very rapidly growing area of flying objects. Various tasks and possible areas of UAVs use are generating increasing demand for various types and designs. Thanks to simpler certification requirements UAV market could be used as a test field for various new technologies like wide use of modern electric systems. One of the area where electric systems could be implemented is a landing gears with its steering, braking or retraction / extension systems.

1. WHAT IS UAV?

An Unmanned Aerial Vehicle is an aircraft without any human crew. It is controlled remotely from stationary or mobile command center. UAV control is performed by radio (directly from ground or via satellites for wider ranges of operation) or autonomously (by on board computers). Autonomous control of the UAV is based on pre-programmed flight plans using complex automation systems.

Figure 1. UAV monitoring and control at CBP (photo: Wikipedia)

UAV main use is military. It can perform attack or reconnaissance missions.
For reconnaissance missions UAV's are often equipped with variety of cameras (from normal video cameras to high resolution infrared ones) and sensors ex. for sensing nuclear or chemical weapons.

Figure 2. A Rheinmetall KZO of the German Army, used for target acquisition and reconnaissance (photo: Wikipedia)

For attack missions UAVs can carry full range of aerial weapons such as missiles, bombs etc. It can be also equipped with systems enabling targeting for artillery and missile attacks (such as laser targetters).
Non military usage of UAVs is jet not so wide but this type of air vessel is in interest of fire fighting brigades (for spotting fires or even to carry extinguishers to areas where extreme danger is present), police operations, media (filming from distance), security work (ex. surveillance of pipelines), rescue missions. In other words UAVs are preferred for missions that are too dull, dirty, or dangerous for manned aircraft.
UAVs come in shapes of regular airplanes or helicopters. Most used propulsion system in aircraft-like UAVs is propeller system. Jet propulsion system is used in systems operating on higher flight levels mainly for wide area reconnaissance.
Size of the UAV depends on what mission it is designed for. UAV take off mass can be as less as 50[kg] and as large as few tons depending of type.
All of UAVs are equipped in features as normal aircrafts. Steering system, landing gear and shape of vessel itself is similar to normal aircrafts.

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

Figure 3. RQ-4 Global Hawk, a high-altitude reconnaissance UAV capable of 36 hours continuous flight time. (photo: Wikipedia)

2. LANDING GEARS.

Dissipation of the energy during landing process can be achieved in different ways:
- spring L/G (in aeronautics L/G is an abbreviation for landing gear) made as a spring beam (fig. 4)
- flexible elements the most often rubber or different elastomers built-up in the L/G structure.
- steel disc springs
- oleo-pneumatic shock absorbers (fig. 5)

Figure 4. Spring beam L/G for Cessna 195 airplane (photo: Wikipedia)

The first three solutions are preferred in light airplanes because of their low cost in connection with high efficiency rates and low weight. Yet use of the oleo-pneumatic shock absorber is the most effective solution during landing.
Because of the largest efficiency in energy dissipation use of oleo-pneumatic shock absorbers is general in military and commercial airplanes where cost of the construction is not the most important criteria.
Oleo-pneumatic shock absorber absorbs energy by "pushing" a volume of hydraulic fluid against volume of gas (usually nitrogen but can be dry air.)
Oleo-pneumatic shock absorbers carry out two functions:
- a spring or stiffness function, which provides the elastic suspension by the compression of a gas volume.
- a damping function, which dissipates energy by forcing hydraulic fluid through one or more small orifices.

UAV landing gear is constructed same way as landing gear for normal aircraft. Landing loads absorption has to be even more effecitve because of delicate sensors onboard UAV.
In small UAVs fixed spring landing gear is most often used. For special applications miniature (but as effective as full sized) retractable landing gear is used. Institutes' of Aviation Landing Gear Departments' small UAV landing gear is a fully operable landing gear with shock absorber which can absorb energy of permissible vertical landing speed up to 3.05 m/s

Figure 5. M-28 „Skytruck" main L/G with oleo-pneumatic shock absorber (design Institute of Aviation L/G Department, photo: IA archive)

3. CONTROL SYSTEMS OF THE UAV'S LANDING GEAR.

In UAVs the same variety of control systems (electric, pneumatic or hydraulic) can be used as in normal air vessels. In big constructions with adequate thrust power and enough space, hydraulic system for steering and landing gear controlling is used.
In smaller constructions fitting hydraulics is a problem of space and mass. Also having enough electric power for sensors, computers and cameras doesn't explain putting other control systems than electric in such UAV.
The main problem of the electric control systems is that they are not as compatible with mechanics as everybody think. Main problem is that electric power can be easily transferred into mechanical movement for rotation. Best example is ordinary electrical motor. In airplane control systems almost always linear movement is needed. Of course there is a variety of electrical linear motors but they are expensive and not light enough. Best way is to make rotary movement into linear. This can be achieved by putting some mechanical gears but this also has it's downback in lowering power and torque with every gear level used. Of course special linear actuator can be built but there are two main problems: price and availability. Price of custom made linear actuator can be much higher than UAV itself. Availability of such actuator is also a problem during operation, there can be situation

in which UAV is grounded for weeks because of time for actuator manufacturing time.

Figure 6. Nose Landing Gear on test stand in Laboratory (photo: IA archive)

These aspects were taken into account during landing gear planning phase in Landing Gear Department. First of all low (100kg) take off mass and plenty of electric power onboard made not reasonable of taking into the account hydraulic control systems.
Second cost of landing gear system had to be acceptable and active parts of the landing gear had to be easily available.
Also there was no restriction of using aviation grade parts because these regulations only apply when air vessel is carrying humans. For non-human aircrafts there is only need of using parts which are reliable enough.

Figure 7. Main Landing Gear on test stand in Laboratory (photo: IA archive)

When all restrictions were taken into the account, Landing Gear Department engineers decided to use common use servos for modelers. Servos had to be durable enough to meet safety requirements and loads generated in the landing gear system. Because of these loads, servos with titanium gears were used.
Exemplary specification of servo that can be used in small UAV's landing gear.

Table 1. Exemplary Specifications of Possible Landing Gear Servo

Control System:	Pulse Width Control 1500usec Neutral
Required Pulse:	3.3-7.4 Volt Peak to Peak Square Wave
Operating Voltage Range:	4.8-7.4 Volts
Operating Temperature Range:	-20 to +60 Degree C (-68F to +140F)
Operating Speed (6.0V):	0.15 sec/60° at no load
Operating Speed (7.4V):	0.12sec/60° at no load
Stall Torque (6.0V):	333.29 oz-in. (24kg.cm)
Stall Torque (7.4V):	416.61 oz-in. (30kg.cm)
Standing Torque (6.0V):	433.27 oz-in. (31.2kg.cm) 5 degree deflection
Standing Torque (7.4V):	541.59 oz-in. (39kg.cm) 5 degree deflection
Operating Angle:	60 Deg. one side pulse traveling 450usec
Direction:	Clockwise/Pulse Traveling 1500 to 1950usec
Idle Current Drain (6.0V):	3mA at stop
Idle Current Drain (7.4V):	3mA at stop
Current Drain (6.0V):	300mA/idle and 4.2 amps at lock/stall
Current Drain (7.4V):	380mA/idle and 5.2 amps at lock/stall
Dead Band Width:	2usec
Motor Type:	Coreless Metal Brush
Potentiometer Drive:	6 Slider Indirect Drive
Bearing Type:	Dual Ball Bearing MR106
Gear Type:	4 Titanium Gears
Connector Wire Length:	11.81" (300mm)
Dimensions:	1.57" x 0.78"x 1.45" (40 x 20 x 37mm)
Weight:	2.18oz (62g)

Figure 8. Example of Servo (photo: HITEC RCD)

Retraction system is a simple system with servo as actuating part and few gears for multiply lifting power of the servo itself. Notification of the open/close state is via computer signal based on position signal from the servo. Closed position is maintained by gas spring and

by computer signal to servo in case of unwanted retraction. All landing gears have also gas spring used as mechanical help during retraction/extension of the landing gear. In retraction system servos were used without any reconfigurations.
Main landing gear is equipped with electrically actuated disc brakes. Brakes themselves are mechanical disc brakes working on the same principle as any disc brakes (ex. car disc brakes). Actuating device is the same servo as one used in retracting and steering systems. Braking is performed by brake lever by servo what is transposed to linear movement of brake pads. Braking force is maintained by spring between actuator and brake mechanism.

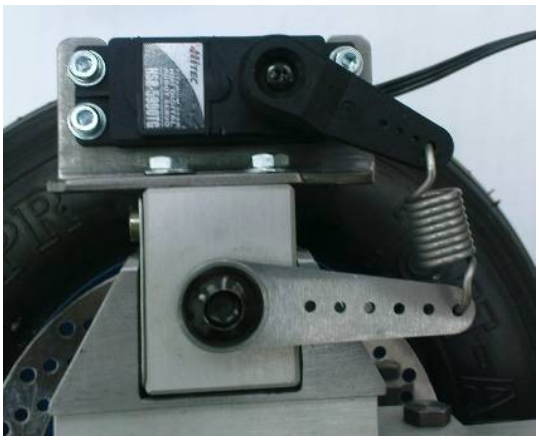


Figure 9. UAV's Main Landing Gear mechanism (photo: IA archive)

Nose landing gear has steering system for ground movement (taxiing). This system is also actuated by the same servo as retraction system. Main difference is that there is a need of knowing current position of the landing gear during whole time of taxiing so the potentiometer form the servo is mounted to front landing gear axis of revolution. During closing of the landing gear front landing gear is rotated onto central position by signal from commanding computer. All landing gears control is purely electric and for operation needs only electrical power source and steering computer. This landing gear can be used in any UAV which take off mass is no more than 100kg. Only need is to plug it to the computer and proper power source.

4. MAIN CHARACTERISTICS OF IA'S UAV LANDING GEARS.

Table 2. Nose Landing Gear Components

Structural parts	Aluminum Alloy.
Wheel	200x50.
Steering system servo	HSR 5990TG.
Steering system potentiometer	EXTERNAL
Retraction / Extension system servo	HSR 5990TG.
Retraction / extension potentiometer	EXTERNAL
Shock absorber	Oleo / pneumatic
Retraction / extension support	Gas spring
Total mass of one nose landing gear	4,8 kg.

Table 3. Main Landing Gear Components

Structural parts	aluminum alloy.
Wheel	200x50 and disc brake size 100 mm
Braking system servo	HSR 5990TG.
Brake pads	Accent FREEZER
Brake spring	Helical
Retraction / Extension system servo	HSR 5990TG.
Retraction / extension potentiometer	EXTERNAL
Shock absorber	Oleo / pneumatic
Retraction / extension support	Gas spring
Total mass of one main landing gear	4,1 kg.

5. SUMMARY

UAVs are most likely to have only electrical control systems in landing gears.
Due to the lower certification requirements and lower development costs, electrically controlled landing gears on UAV are easier and cheaper way for future implementation for bigger, passenger aircrafts. All new electrical technology could be quickly tested on less demanding unmanned vehicles.
The knowledge acquired during the development of presented project, results in better understanding of problems and design principles of electrical loading systems, especially for aviation applications where landing gears systems are special and very challenging system responsible for the most dangerous phase of the flight process.
Described landing gear is currently mounted in research UAV for NACRE project which is a part of UE 6th Framework Programme. All landing gear systems were verified during laboratory tests as well as UAV final flight tests.

6. REFERENCES

Wołejsza Z., Kowalski W., Lafitte A., Mikułowski G. and Remmers L., 2005 “State of the Art. In Landing Gear Shock Absorbers”, Transactions of the Institute of Aviation

Currey N. 1989 “Aircraft Landing Gear Design: Principles and Practices”, AIAA Education Series

Wikipedia, en.wikipedia.org/wiki/Unmanned_aerial_vehicle

Institute of Aviation website, www.ilot.edu.pl

Landing Gear Department website, www.cntpolska.pl/index.php/landing-gears-department/about-us

AUTHOR BIOGRAPHIES

ZBIGNIEW KONRAD SKORUPKA, born in 1977 in Warsaw, Poland. Finished Warsaw Academy of Technology, where he studied Shaping Machines Design and Steering. He received his M.Sc. in 2004. Currently he is working at Landing Gear Department of Warsaw Institute of Aviation. His current professional area are: use of smart materials in landing gear, general design and testing of landing gear, brakes and test stands.

E-mail: zbigniew.skorupka@ilot.edu.pl

WOJCIECH KOWALSKI, born in Warsaw, Poland and went to the Warsaw Academy of Technology, received his M.Sc. in Engineering in 1970, Faculty of Power and Aeronautical Engineering in Applied Mechanics and 2000 PhD from Warsaw University of Technology, Faculty of Automotive and Construction Machinery Engineering. Now he is working at Warsaw Institute of Aviation. His current professional area are: load and stress analysis, construction dynamics, test parameters determination, results analysis and interpretation of tests and calculations

E-mail: wojciech.kowalski@ilot.edu.pl

RAFAŁ KAJKA born in 1977, received his M.Sc. in Engineering (2000) from Warsaw University of Technology, (Poland). Directly after graduation he started his doctoral studies and in 2005 he received PhD from Warsaw University of Technology, Faculty of Automotive and Construction Machinery Engineering. Since 2001 he is working at Institute of Aviation, Warsaw initially as a strength engineer and from 2007 as a Manager of Landing Gear Department, Institute of Aviation, Warsaw, Poland.

E-mail: rafal.kajka@ilot.edu.pl

LOW-COST SYSTEM FOR DETECTING TRAFFIC OFFENCES

Łukasz Kamiński
University of Warsaw
ul. Banacha 2
02-097 Warszawa, Poland
kamis@mimuw.edu.pl

Michał Popławski
Nicolaus Copernicus University
ul. Gagarina 11
87-100 Toruń, Poland
michalpopl@fizyka.umk.pl

Michał Łyczek
University of Warsaw
ul. Banacha 2
02-097 Warszawa, Poland
michal.lyczek@students.mimuw.edu.pl

KEYWORDS

road traffic analysis, real-time image analysis, ANPR, detection of traffic offences.

ABSTRACT

This paper describes the implementation of a prototype installation for automatic detection of traffic offences with the use of video camera real time analysis. In this paper, we focus on the technical aspect of the installation as well as on the possibilities to implement algorithms to detect offenders. The project of the installation assumed the use of low-cost series of components and reducing infrastructure requirements in the place of assembly of the equipment.

INTRODUCTION

There are many solutions for the detection of traffic offences worldwide. In Poland, the traffic enforcement camera is the most popular device. In the world there are networks of automated traffic offences detection systems. An example of the speed control system and compliance with traffic lights was launched in the UK (SAFECAM). Many manufacturers offer solutions to control both speed and other offences (e.g., SODI, RADAR).

Speed measurement systems, which currently applies worldwide rely on specialized equipment to measure vehicle speed. Most arrangements are based on a fixed frequency wave emissions and measure their reflection from a moving vehicle. Other solutions include laser meters or inductive loops placed in the road.

The aim of this work is to describe a prototype and a concept of the installation to detect offenders. Installation is intended to provide the ability to detect violations at a low cost. Regards both the purchase price of equipment, price of the hardware, and the cost of local infrastructure.

The paper will present a implementation of the measuring device both the hardware and analysis software for the device.

CONCEPT

The aim is to create system based on standard, commonly available components, for detecting a variety of traffic offences. The system is intended to detect following offences:

1. Breaking speed limit - in a particular place or along the road.
2. Driving on a red light.
3. Overtaking in unauthorized locations
4. Parking in unauthorized locations
5. Ignoring the STOP sign.

Implementation of the system of different offences is possible by means of dedicated algorithms for their detection.

This work presents a hardware implementation of the concept and principles of implementation of basic elements of software installations to the detection of these offences.

For each offence system must identify the vehicle in the testing area and must automatically locate and recognize the licences plate.

REALIZATION

Design concept aims to minimize the need for creating dedicated infrastructure for the system. Therefore, to communicate with the device GSM network will be used. Currently, network coverage in Poland, where HSDPA transmission is available is still negligible. By contrast, it can be assumed that almost everywhere (with appropriate antennas) can be reached EDGE connection.

Establishment of a transmission in EDGE forces reducing the amount of information sent from the device describing detected events. Therefore, the system assumes the need for processing the video signal in real-time in the immediate proximity of the video camera.

The exact method of implementation to communicate with an external system, the type of data collected, and how they are transmitted is described in the work.

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

Traffic analysis device consists of a image processing unit for image analysis and two video cameras (for different offences amount of cameras may vary). Cameras are connected with wires to the computer. These devices should be installed in close proximity. You can use a common assembly point for all the elements. All devices are powered from one power source 230V (standard in Poland).

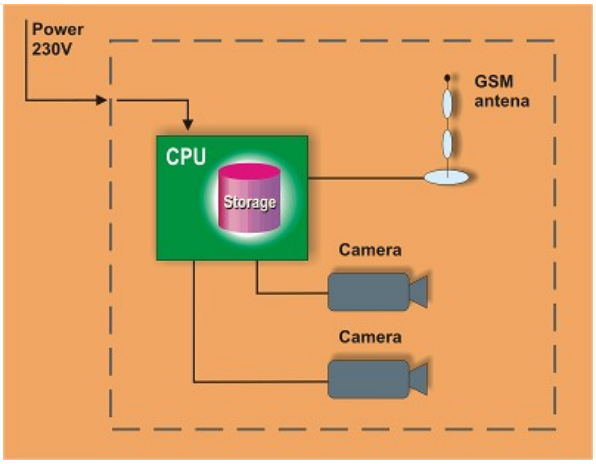


Figure 1: Diagram of the device.

Hardware architecture:

1. Computer for real-time image processing.
2. Wireless communication module (GSM).
3. Camera for offences registration
4. Camera for vehicle and driver identification.

Examples of installations for the detection of speeding and failure to detect the traffic light.

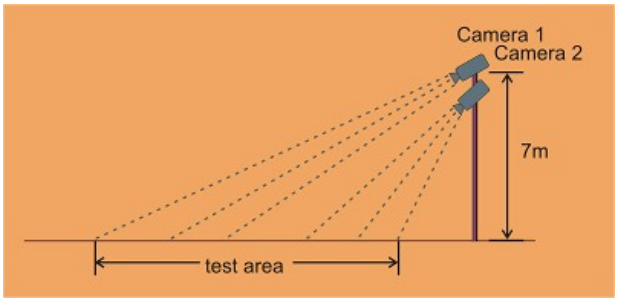


Figure 2: Detecting breaking speed limit

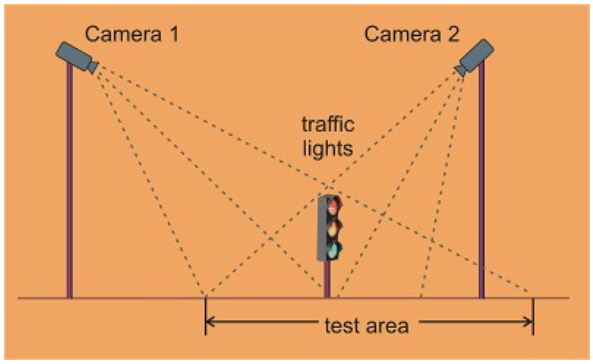


Figure 3: Detecting traffic lights violation.

Software - the basic algorithm for image analysis.

1. Image processing from two cameras in real-time.
2. Vehicle detection and tracking.
3. Licences plate detection.
4. Licences plate identification.
5. Offences detection

A key element is the installation of appropriate camera (performance and image quality). The installation uses 3.1Mpix camera with a resolution of 2048px x 1536px. Depending on camera parameters it would be possible to recognize face of the driver and vehicle registration plates. An important parameter characterizing camera is the exposure time for a single frame. For sharp image of a fast-moving vehicles it is necessary to have a short exposure time. Conducted tests showed that it is important to scan the camera matrix consistent with the expected direction of movement of vehicles. Another important factor is that camera should scan all lines at once and not the even and odd lines in separate passes (interlacing). Tested camera for a single frame, with exposure time of 10ms, gave for vehicles travelling at speeds of around 100km/h picture completely unreadable (the apparent difference between adjacent lines of the image).

Tests showed that for single frame at 1ms (which gives about 0.5μs for exposure of a single line of the image) there was no change in the image quality for a range of speeds for moving vehicles. This allows to say that the image quality is not dramatical, even for vehicles travelling at speeds around 200 km/h, which allows for effective application of this solutions anywhere.

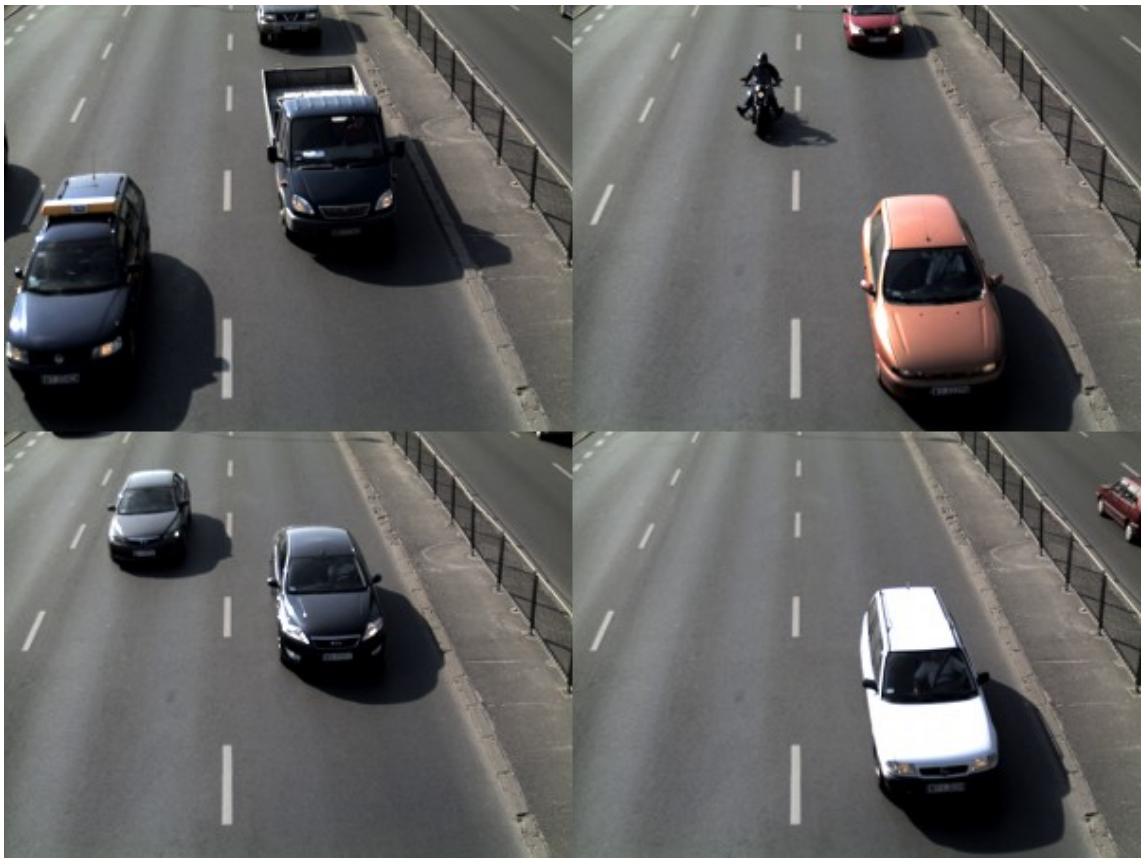
Device parameters:
Processor CORE 2 QUAD Q8200 2.33GHz LGA775 BOX AV3100DN + single lens 12VM412ASIR
Bitrate: 4MB/s (32Mbit/s).

Vehicles are identified using algorithm for vehicle detection presented in this paper.

TESTS

System tests were carried out in two stages. The first stage consisted of recording a large amount of test material in place of measuring the different lighting conditions.

For registration materials have been used existing infrastructure on Puławska street in Warsaw.



In the second phase we tested algorithm on the recordings. Tests were conducted using the library libanpr. The last phase of testing was:

1. Verification whether the equipment will be possible to analyse images from cameras in real time (to examine sufficient capacity).

2. Check if the image quality from the cameras is sufficient to achieve the detection algorithm of offence (whether it is possible to locate the vehicle and the automatic location of the licence plate).

SUMMARY

Presented results show that the implementation of cheap and simple system described by the concept is possible. In order to achieve the detection of specific offences would be necessary to implement dedicated algorithms. The creation of such algorithms may be the subject of future work. It is possible to implement specific extension of the algorithm to detect other offences. It should be also confront the idea of a solution to the existing legal regulations in different countries of application.

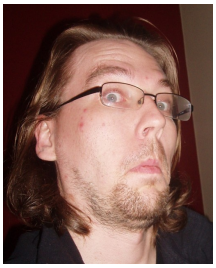
Description of the exact idea of the system in the realities of the organization in Poland can be found in the work.

In the future, one can also consider the integration of solutions from other sources of information of traffic speeds, such as inductive loops.

Another aspect for further research is to examine the potential for reducing energy consumption of the entire device in order to be able to supply it from it's own energy source or from solar or wind generators.

REFERENCES

1. Łyczek M., Butryn P., Chrobak A., Kulka A., 2008. "Information acquisition for city traffic models based on image analysis" in *Proceedings of 22nd European Conference on Modelling and Simulation*.
2. Łyczek M., Kamiński Ł., Chrobak A., Kulka A., 2009. "Automatic Traffic Detection System" in *Proceedings of 23rd European Conference on Modelling and Simulation*.
3. C.N. Anagnostopoulos, I. Anagnostopoulos, V. Loumos, E. Kayafas, 2006. "A license plate recognition algorithm for Intelligent Transportation System applications" in *ITS*, No. 3.
4. Cheolkon Jung, Qifeng Liu, Joongkyu Kim, 2009. "A stroke filter and its application to text localization" in *Pattern Recognition Letters*.
5. Datong Chen, Jean-Marc Odobez, Herv/e Bourlard, 2004. "Text detection and recognition in images and video frames" in *Pattern Recognition*.
6. Zurad - http://www.zurad.com.pl
7. Safecam - http://www.safecam.org.uk/
8. Sodi - http://www.sodi.com/
Radar - http://www.radardetectors.co.uk/



ŁUKASZ KAMIŃSKI was born in Brodnica. He studied mathematics and computer science and obtained his Master degree from University of Warsaw. He won Polish Academic Collegiate Programming Contest and ACM Central European Collegiate Programming Contest. He was ACM International Collegiate Programming Contest finalist. Currently, he is working on his doctoral thesis in the Faculty if Mathematics, Informatics and Mechanics at the University of Warsaw.
l.kaminski@mimuw.edu.pl

MICHAŁ ŁYCZEK was born in Warsaw, Poland. He finished high school with special mathematical profile in 2005. Since then he studies in the Center for Interfaculty Individual Studies in Mathematical and Natural Sciences at the University of Warsaw. Since 2007 he has been working on image processing and visualization software in Interdiciplinary Center for Mathematical and Computational Modeling at the University of Warsaw. His email address is:
michal.lyczek@students.mimuw.edu.pl

MICHAŁ POPŁAWSKI was born in 1980 in Torun. He studied physics and obtained masters degree on Nicolai Copernicus University in Torun. He also graduated on Polish Air Force Academy in Deblin. His thesis included aerodynamics and quantum mechanics' simulations, especially based on real time calculations and OpenGL renderings. Currently He is working on his doctoral thesis on the Faculty of Physics, Astronomy and Applied Informatics, Nicolai Copernicus University in Torun. He is also a full-time military pilot.

FLYING OBJECT ARMOUR CONCEPT ANALYSIS BASED ON HELICOPTER

Włodzimierz GNAROWSKI, Jerzy ZOLTAK, Rafał KAJKA,
Institute of Aviation
02-256 Warsaw, Poland
E-mail: wlodzimierz.gnarowski@ilot.edu.pl, jerzy.zoltak@ilot.edu.pl
rafal.kajka@ilot.edu.pl

KEYWORDS
Helicopter, Armour, Protection, Institute of Aviation.

ABSTRACT

Medium helicopter armour concept is discussed. Basic analysis of armour configuration on helicopter fuselage is described. Important zones and requirements for their protection are analyzed as well. Aerodynamic characteristics of medium helicopter are obtained using numerical simulation and wind tunnel tests.

1. INTRODUCTION

Helicopter armour needs some compromises. Armour must provide proper protection level for crew and helicopter components and has to be as light and as dense as possible due to mass requirements. In Cobra, Apache or Mi-24 helicopters, armour is made as integral part of helicopter fuselage (Mi-24, fig 1 and 2)

Figure 1. Mi-24 helicopter – example of armour integrated with fuselage of the helicopter (photo: Wikipedia)

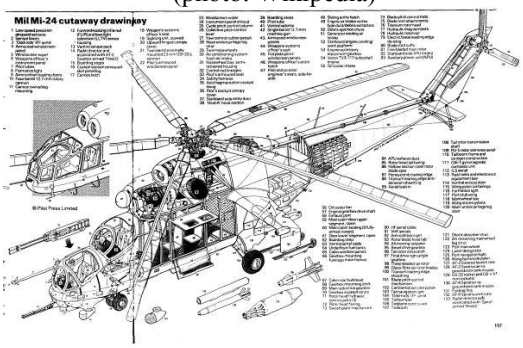


Figure 2. Equipped fuselage of Mi-24 helicopter (photo: Wikipedia)

In case of multipurpose helicopters used both in transport, medical and combat missions, external armours (built around helicopter fuselage) or internal armours (built as protective screens for transported cargo) are used. These armours are generally protecting against fire from small or medium calibre machine guns (Mi-17, fig. 3)

Figure 3. Mi-17 helicopter with Dyneema RQ-4 armour plates (photo: Wikipedia)

2. HELICOPTER ARMOR REQUIREMENT.

For multipurpose helicopter armour protection concept phase, design team based on experience and test methodology recommended by ,,Standardization Agreement - 4569" (STANAG) for light armoured vehicles.

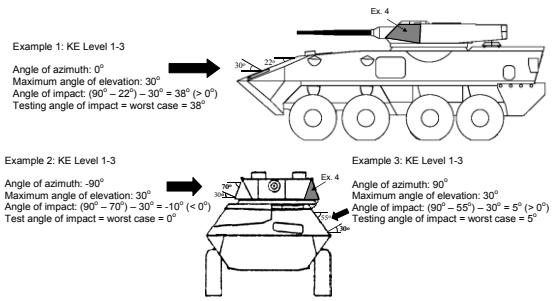


Figure 4. Demonstration of the determination of angles of impact for sloping plates on actual vehicles. The impact angle of the artillery threat may be established using the same methodology but applying 360° of azimuth and the elevation specified for each Protection Level defined in Appendix 1 STANAG 4569 .

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

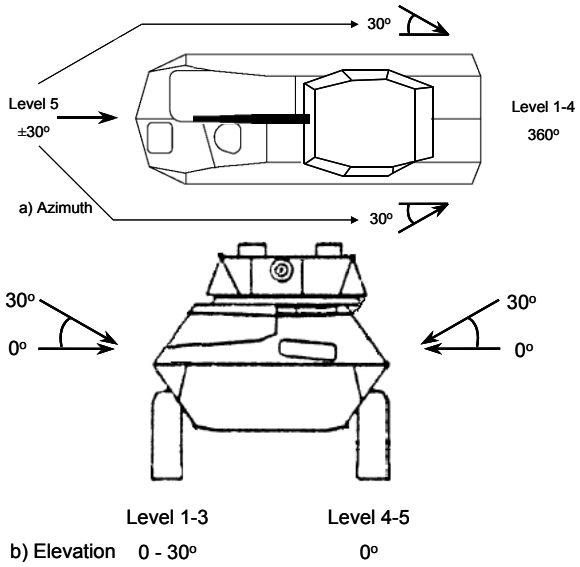


Figure 5. Attack angles defined in Appendix 1 STANAG 4569 for the Kinetic Energy Protection Levels

3. HELICOPTER ARMOUR CONCEPT.

Proposed armour concept concentrates on bullet protection of cockpit (pilot + special instrumentation operator) as well as on protection of main systems of the helicopter such as radio-electronic devices compartments/bays, main gear reducer and engines. It will be possible by using flat plates made of multilayer Ultra-High Molecular Weight Polyethylene (UHMWPE) Dyneema SB 21 with energy absorption coefficient Eabs" > 245 J/(kg/m2) for cockpit protection and radio-electronic devices compartment. Total area of external armour will be equal to 12,48[m^2]

Cargo bay will be protected by soft screen made out of kevlar fabric reinforced with ceramic inserts, which protect from 7,62[mm] and 12,7[mm] calibre machine gun bullets. Total area of soft armour will be equal to 11,12[m^2]

Armour of areas exposed to high temperature engine exhaust gases from turbine engines will be made out of Titanium plates (1,66[m^2] total area).

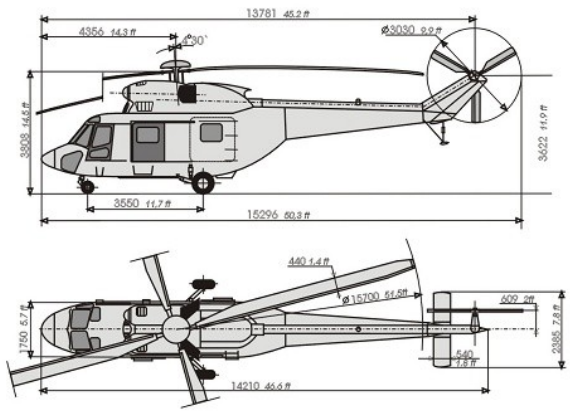


Figure 6. General geometry of Helicopter

In helicopter mass balance, statistically speaking, armour is around 20÷30% of payload. This size of additional mass possibly slightly decrees performance properties such as climb speed and range, which are important parameters in terms of tactical applications.

Transparent surfaces of the helicopter has been reduced to smallest dimensions as possible, reducing the area of hitting inside of the helicopter and these areas will be secured by polycarbonate plates.

The front part of the helicopter, due to the small angular velocity relative to earth movement, when frontal gunfire emerges, requires particularly careful protecting, both for technical and psychological reasons, what gives reason for placing large percentage of the armour in front section of the helicopter (above 90%).

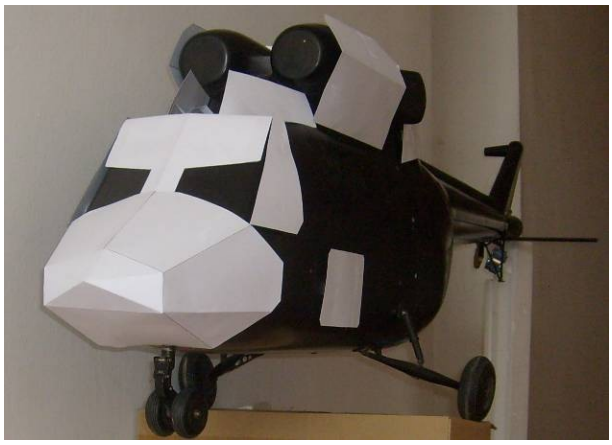


Figure 7. Concept of 6 tons take off mass helicopter armour (model)

Figure 8. Front view of armour installed on helicopter model (model used in wind tunnel tests)

4. ARMOUR INFLUENCE ON AERODYNAMIC CHARACTERISTICS OF HELICOPTER

The additional armour mounted on a helicopter fuselage changes its external geometry. This fact leads to change of object's aerodynamic characteristic.
To obtain aerodynamic characteristics of the tested object the commercial code FLUENT has been used. Simulations were done for Mach number M=0.1 and selected angles of attack. Two configurations are tested. The first one is the clean helicopter geometry - basic configuration. The second one is the armoured helicopter configuration.
In the Figure 9 the numerical meshes for both tested configuration are presented.

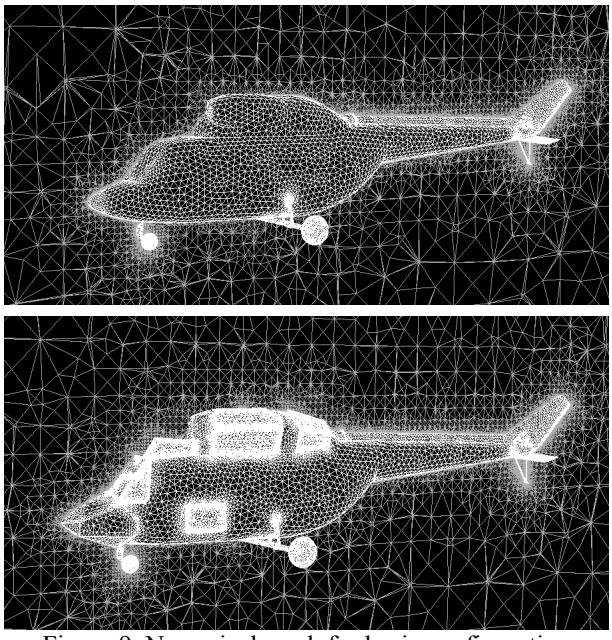


Figure 9. Numerical mesh for basic configuration (upper) and armoured configuration (lower)

The size of numerical meshes are: 0.8 mln cells for baseline and 1.5 mln cells for armoured geometry respectively.
During the simulation basic flow parameters of flowfield over the objects are analyzed. In the Figure 9 the static pressure on the helicopter surface are presented. In the Figure 10 the Mach number distribution in symmetric plane is shown. It can be seen that on the upper surface of fuselage by the front armour there are flow separation region.
For each tested case the basic aerodynamic forces were calculated. The influence of helicopter armour on aerodynamic characteristics is analyzed for selected cases (conditions). For example for angle of attack α=0, the lift coefficient is near for both configurations but drag coefficient for armour configuration is 25% higher than for basic configuration.
In next step of the analysis the wind tunnel test both configurations will be done.

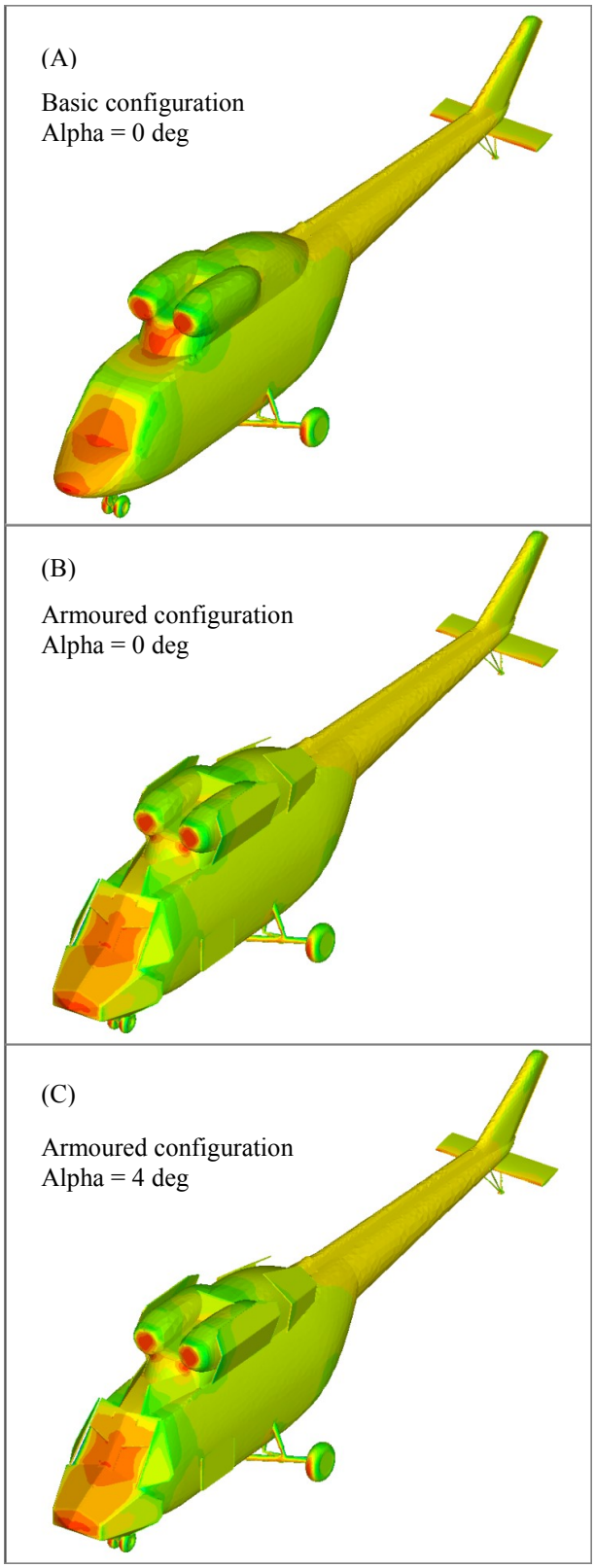


Figure 10. Static pressure distribution on helicopter surface for basic configuration (A) and armoured configuration (B) and (C)

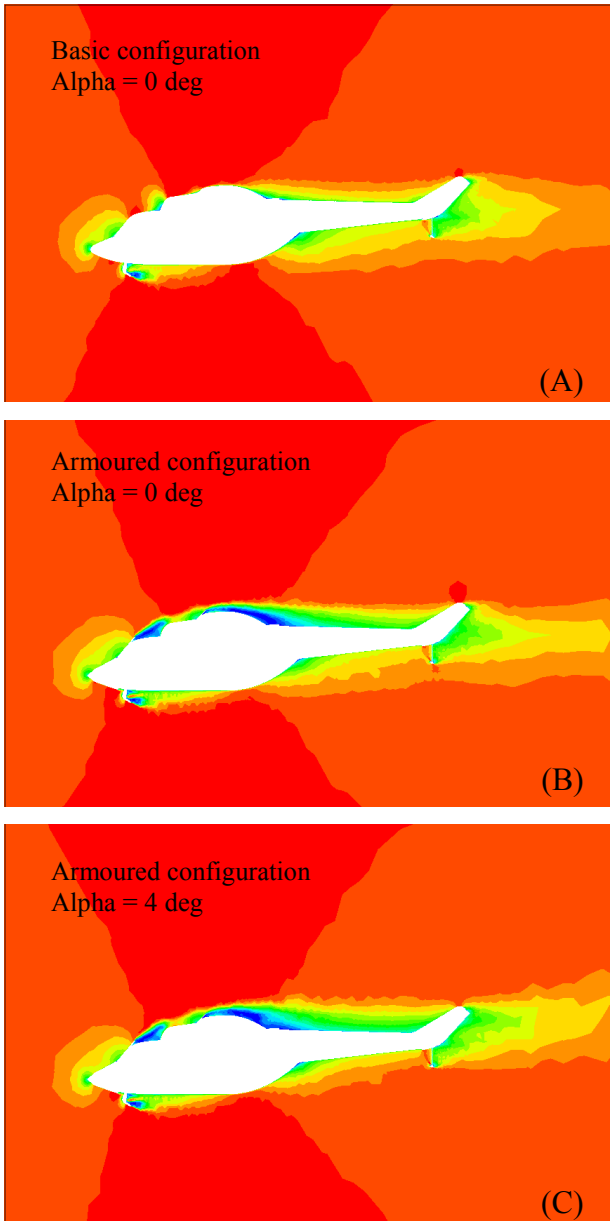


Figure 11. Mach number distribution in symmetric plane for basic configuration (A) and armoured configuration (B) and (C)

5. SUMMARY

Protection of flying objects is very difficult. Due to rigorous mass requirement and specific external shape (profile highly optimized during designing), any even small modification of the mass, redistribution of CG or external shape can lead to dramatic decreasing functionality, safety or performance. From the other side operation condition quite often required additional protection of the flying object to provide acceptable safety level for the crew and important subsystems. Article presents initial stage of armour developing for 6T helicopter. Three faze are described: literature study, armour concept build on model, CFD analysis and initial preparation to model based wind tunnel tests. All these steps was done to provide preliminary answers concerning sensitivity and stability of the flying object under modifications caused by eventual armour installation.

6. REFERENCES

[1] Standardization Agreement (STANAG) – 4569 NATO - January2004
[2] Influence of armour on a helicopter aerodynamic characteristics , April 2010, Instytut Lotnictwa, Internal Report, April 2010, (in edition, polish)
[3] www.wikipedia.org – web resources
[4] Witkowski R.: Introduction to knowledge about helicopters. Research Library of Institute of Aviation, Warsaw, Poland, 1998.
[5] Ball R.E.: The fundamentals of aircraft combat survivability analysis and design, second edition. Blacksburg, American Institute of Aeronautics and Astronautics, Inc. 2003.

WŁODZIMIERZ GNAROWSKI was born in Poznań, Poland. Chief designer in Institute of Aviation, Warsaw, Poland. Profile/ experience: High skilled and experienced engineer, after studying aircraft construction at Technical University of Warsaw, has built his career working at Institute of Aviation in Warsaw. A designer, former leader and chief designer of IRYDA combat trainer program, focused his professional interest on aeronautical construction problems, was responsible for and has realized several engineering projects, author of concept project of ground attack combat aircraft KOBRA characterized by high maneuverability and very good aerodynamic performances confirmed in model tunnel tests for low and high speed, recently actively involved in EU research programme concerning wing high lift devices for a passenger airplane.
E-mail wlodzimierz.gnarowski@ilot.edu.pl

JERZY ŻÓŁTAK has led the Numerical Aerodynamic and Flight Dynamics Group at the Aviation Institute in Warsaw, Poland since 1992. Earlier he was the senior researcher at the High Speed Wind Tunnel Laboratory (1980-1992) and an associate researcher at Warsaw University of Technology (1995-1997) He was also a visiting researcher at the Technical Engineering Department of the University of Manchester Institute of Science and Technology (for nine months in 1996). His current professional area are: numerical aerodynamic, numerical methods, numerical design and optimization
E-mail: jerzy.zoltak@ilot.edu.pl

RAFAŁ KAJKA born in 1977, received his M.Sc. in Engineering (2000) from Warsaw University of Technology, (Poland). Directly after graduation he started his doctoral studies and in 2005 he received PhD from Warsaw University of Technology, Faculty of Automotive and Construction Machinery Engineering. Since 2001 he is working at Institute of Aviation, Warsaw initially as a strength engineer and from 2007 as a Manager of Landing Gear Department, Institute of Aviation, Warsaw, Poland.
E-mail: rafal.kajka@ilot.edu.pl

CAR BRAKE SYSTEM ANALYTICAL ANALYSIS

Wojciech Kowalski, Zbigniew Skorupka, Rafał Kajka, Jan Amborski,
Institute of Aviation
02-256 Warsaw, Poland
E-mail: zbigniew.skorupka@ilot.edu.pl
wojciech.kowalski@ilot.edu.pl
rafal.kajka@ilot.edu.pl
jan.amborski@ilot.edu.pl

KEYWORDS
Car brakes, braking system, brake analysis.

ABSTRACT

This paper contains numerical analysis of brake system for heavy (mass m=6400[kg]) transport car. Analysis was performed in order to correct existing non optimal brakes in mentioned car. Analysis was based on results of the brake system dynamic tests made in Landing Gear Laboratory Institute of Aviation in Warsaw, Poland. Authors describe analytical process which led to generate results for new parameters of more efficient braking system for heavy transport car.

1. INTRODUCTION

In 2009 Institute of Aviation Landing Gear Department was asked for redesigning brake system for heavy transport car. Car brake system was not efficient enough and too costly in maintenance because of extensive wear of braking shoes (rear brake) and braking pads (front brake).
Analyzed car brake system used two types of brakes: rear drum brakes and front disc brakes. Such configuration of brakes is common but not very efficient due to lover efficiency of drum brakes in general. Another problem was that brakes weren't made especially for this car but were taken from another car which was similar in parameters but had much lower nominal mass.
Landing Gear Department was supplied with some brake parameters (made by car testing facility) but full characteristics of brakes were unknown. Landing Gear Department has its own laboratory equipped with test stand capable to perform tests of brake systems.
Based on results from car test facility new set of tests were made in Landing Gear Laboratory in order to achieve full characteristics of existing brakes. Obtained results were used in calculation for new brake system. Calculations assumption was made that disc brakes will be used on both axles.
Below chapters shows analysis and calculations which led to generate parameters for improved brake system for heavy transport car.

2. ANALYSIS AND CALCULATIONS

2.1. Preliminary Data

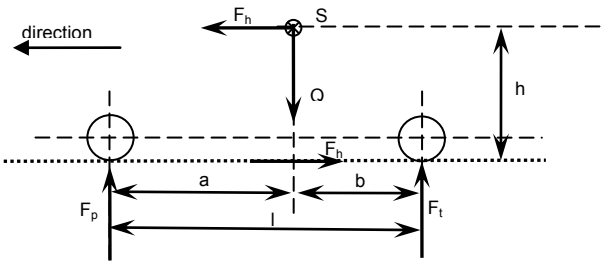


Figure 1. Car Load Distribution

Table 1. Used designations

Q	Vehicle weight	[N]
F_p	Front axle static reaction (for both wheels)	[N]
F_t	Rear axle static reaction(for both wheels)	[N]
F_h	Braking force (four wheels)	[N]
l	Wheel base	[mm]
h	Center of gravity to ground distance	[mm]
a	Front axle to center of gravity distance	[mm]
b	Rear axle to center of gravity distance	[mm]
m_s	Vehicle mass	[kg]
m_1	Vehicle mass for front axle	[kg]
m_2	Vehicle mass for rear axle	[kg]
g	g force = 9,81	[m/s^2]

Table 2. Preliminary Data

m	62931,2	[N]
m_1	27615,2	[N]
m_2	35316,0	[N]
h	1169	[mm]
a	1783	[mm]
l	3191	[mm]
Q	62931,2	[N]
a	1783	[mm]
h	1169	[mm]
b	1408	[mm]
F_p	27767,8	[N]
F_t	35163,3	[N]
r	0,426	[m]
F_p/F_t	0,790	

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

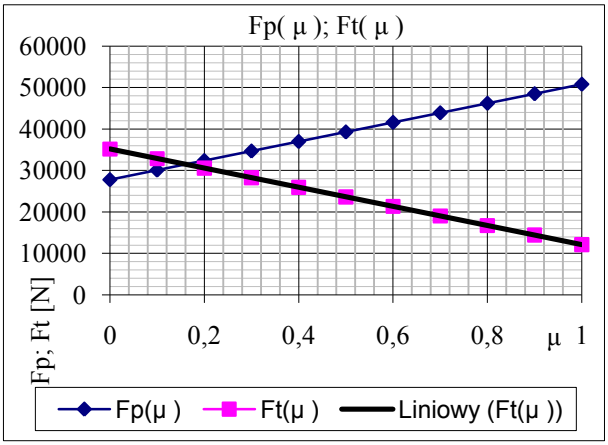


Figure 2. Dynamic Balancing of the Car

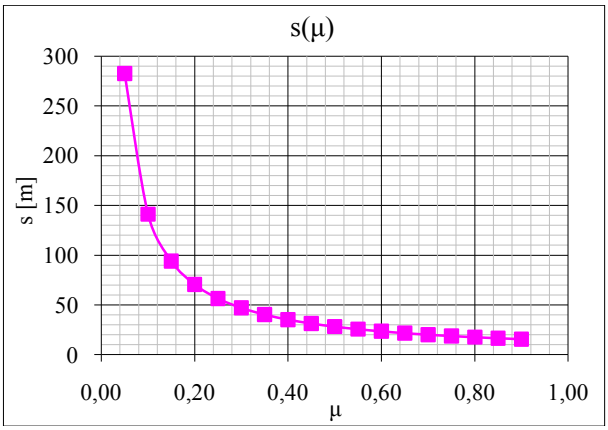


Figure 3. Braking Distance Versus Tyre-Ground Friction Coefficient

2.2. Brake Selection Analysis

Let's assume that brake is optimized for tyre-ground friction coefficient equal to μ=0,8 (see fig. 2) then:

$$F_t = \frac{16720}{2} = 8360 \text{ [N]}$$

$$F_p = \frac{46212}{2} = 23106 \text{ [N]}$$

$$F_{Ht} = \mu * F_t = 6688 \text{ [N]}$$

$$F_{Hp} = \mu * F_p = 18485 \text{ [N]}$$

$$r = 0{,}426 \text{[m]}$$

$$M_{ht} = F_{Ht} * r = 2850 \text{ [Nm]}$$

$$M_{hp} = F_{Hp} * r = 8300 \text{ [Nm]}$$

$$m = \frac{M_{ht}}{M_{hp}} = 2{,}91$$

Let's assume, that every braking pressure will give us constant braking moment ratio m=const

Kinetic energy of the vehicle with mass m_s = 6400 [kg] and speed V=60 $\left[\frac{km}{h}\right]$=16,67 $\left[\frac{m}{s}\right]$ is equal to

$$E = \frac{m_s * V^2}{2} = 888888{,}9 \text{[J]}$$

Second parameter is a braking distance.
Vehicle braking test shown that braking distance from velocity V =60 $\left[\frac{km}{h}\right]$ is equal to s_h=30[m]
For vehicle stop in desired distance (assuming constant decelerated motion) we need braking force equal to:

$$F_h = \frac{E}{s} = 29629 \text{ [N]}$$

or total braking moment equal to:

$$M = F_h * r = 12622 \text{ [Nm]}$$

assuming that

$$m = \frac{M_{Ht}}{M_{Hp}} = 2{,}91$$

result is

$$M_{Ht} = \frac{12622}{2*(1+2{,}91)} = 1614 \text{[Nm]}$$

$$M_{Hp} = 2{,}91 * 1614 = 4697 \text{[Nm]}$$

2.3. Braking Pressures Estimation

After front brake tests, it turned out, (despite first suggestions) that pressure in braking system can be around 3[MPa] and pressure needed to achieve braking moment at level of 3000[Nm] is around 9 ÷10[MPa].
According to results of the tests for front and rear brakes distribution of the rear and front braking moments can be checked in assumption that pressures in both front and rear brakes are equal. Initial parameters are taken from previous chapter
Vehicle energy at the start of braking process:

$$E_s = \frac{mV^2}{2} = 888888{,}9 \text{ [Nm]}$$

All wheels braking force:

$$F_{ch} = \frac{E_s}{s_h} = 29629{,}63 \text{[N]}$$

According to dynamic tests linear interpolation of front axle braking moment (disc brake) versus braking pressure was made:

$$M_{hpśr} = 342{,}01 * p_{hp} + 28{,}329 \text{[Nm]}$$

$$y = 342{,}01 * x + 28{,}329$$

$$F_{hpśr}=\frac{342{,}01*p_{hp}+28{,}329}{R_{kp}}\ [N]$$

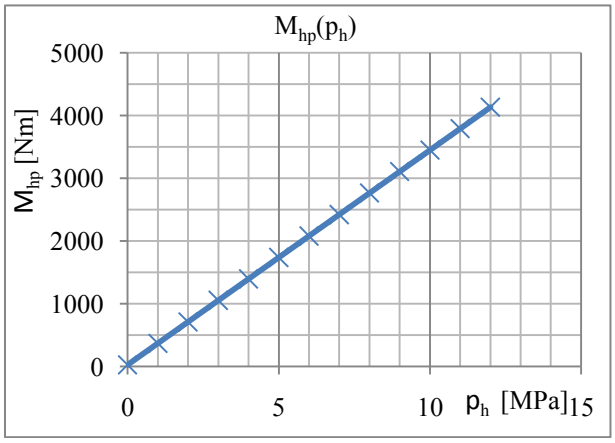


Figure 4. Front Axle Braking Moment Versus Braking Pressure – Linear Interpolation

For rear axle drum brake, two braking moments versus braking pressure two interpolations were made. One is linear as in the case of front brake while the second is non linear. Such analysis was made because of non linear drum brake $M_h(p_h)$ relation caused by drum brake operating principle.

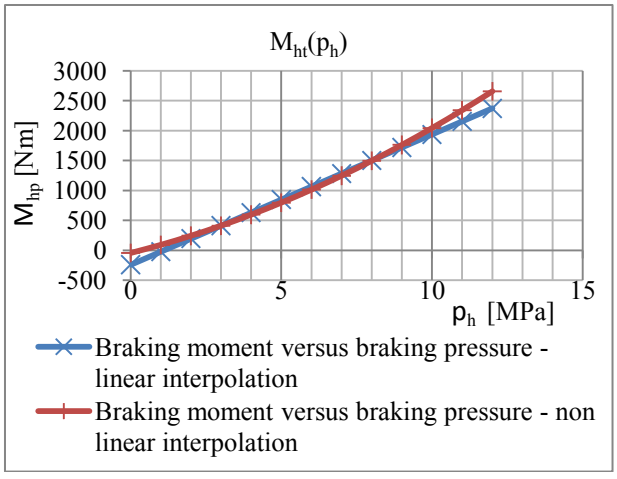


Figure 5. Rear Axle Braking Moment Versus Braking Pressure – Interpolations

Linear interpolation:

$$M_{htśr}=217{,}59*p_{ht}-239{,}2[Nm]$$

$$y=217{,}59*x-239{,}2$$

$$F_{htśr}=\frac{217{,}59*p_{ht}-239{,}2}{R_{kt}}\ [N]$$

Non -linear interpolation:

$$M_{htśr2}=8{,}0975*(p_{ht})^2+127{,}92*p_{ht}-44{,}657[Nm]$$

$$F_{htśr2}=\frac{8{,}0975*(p_{ht})^2+127{,}92*p_{ht}-44{,}657}{R_{kt}}\ [N]$$

Overall braking force is equal to:

For linear interpolation:

$$\frac{2*\left(342{,}01*p_{hp}+28{,}329\right)}{R_{kp}}+\frac{2*(217{,}59*p_{ht}-239{,}2)}{R_{kt}}-F_{hc}=R=0$$

$$R=0{,}731872718$$

$$p_h=11{,}655[MPa]$$

For nonlinear interpolation:

$$\frac{2*\left(342{,}01*p_{hp}+28{,}329\right)}{R_{kp}}+\frac{2*\left(8{,}0975*(p_{ht})^2+127{,}92*p_{ht}-44{,}657\right)}{R_{kt}}-F_{hc}=R=0$$

$$R=1{,}978344$$

$$p_h=11{,}275[MPa]$$

verification

$$F_{hc}=29629{,}62963$$

$$F_{hpśr}*2+F_{htśr}*2=29630{,}36153$$

Summary

Car (mass $m_s=6400$ [kg]) braking from speed $V_s=60\ \left[\frac{km}{h}\right]$, will stop within $s_h=30$[m] when braking pressure is equal to $p_h=11{,}655$[MPa] (linear interpolation) or $p_h=11{,}275$[MPa] (non linear interpolation) what gives us average braking pressure $p_h\cong 11{,}5$[MPa]

2.4. Braking With Constant Tyre-ground Friction Coefficient (μ = 0,8)

Aviation regulations (ex. FAR, JAR) recommend taking μ = 0,8 tyre-ground friction coefficient for permissible side loads.

Where permissible loads are the ones which can be present during standard operation.

Let's take an assumption that analyzed car uses friction tyre-ground coefficient μ = 0,8

Rest of the parameters are the same as in previous chapters (repeated below for better overview).

$m_s=6400$ [kg] car mass

$V_s=60\ \left[\frac{km}{h}\right]=16{,}66667\left[\frac{m}{s}\right]$ car speed

$s_h=30$[m] braking distance

$R_{kp}=0{,}426$[m] front wheel radius

$R_{kt}=0{,}426$[m] rear wheel radius

$E_s=\frac{mV^2}{2}=888888{,}9$ [Nm] car energy in the beginning of braking

$F_{ch}=\frac{E_s}{s_h}=29629{,}63$[N] all wheels braking force

$p_h=11{,}5$[Mpa] average braking pressure from chapter 2.3

Braking forces and moments for one wheel in the actual brakes' configuration.

$M_{hpśr05}$=3961[Nm] $F_{hpśr05}$=9299[N]

E_{hp05}= 278970[Nm]

$M_{htśr05}$=2497[Nm] $F_{htśr05}$=5862[N]

E_{ht05}= 175860[Nm]

E_{hc05}= 909820[Nm] - overall braking energy (4 wheels)

Difference:

$\frac{(E_{hc05} - E_s)}{E_s} * 100 = 2,4$ [%]

Derives from approximations during analysis and calculations. Such a difference is fully acceptable from engineering point of view.

Vertical loads on one axle for friction coefficient $\mu = 0,5$ (chapter **2.1. Preliminary Data**.) are equal to:

F_{p05}=39295[N]

F_{t05}=23636[N]

For friction coefficient $\mu = 0,8$ vertical axis loads will be equal to:

F_{p08}=46211[N]

F_{t08}=16720[N]

Braking forces and moments for one wheel ($\mu = 0,8$) are equal to:

$$F_{hpśr08}=\frac{F_{p08}*\mu}{2}=18485[N]$$

$$F_{htśr08}=\frac{F_{t08}*\mu}{2}=6688[N]$$

$$M_{hpśr08}=F_{hpśr08}*R_{kp}=7874[Nm]$$

$$M_{htśr08}=F_{htśr08}*R_{kt}=2849[Nm]$$

In redesigned brakes pressure will be the same as in previous version, main change will be the area of the brake pistons.

Pistons area coefficient will be:

$$n_p=\frac{A_{pnew}}{A_{pold}}=\frac{F_{hpav08}}{F_{hpav05}}=\frac{M_{hpav08}}{M_{hpav05}}=1,99$$

$$n_t=\frac{A_{tnew}}{A_{told}}=\frac{F_{htav08}}{F_{htav05}}=\frac{M_{htav08}}{M_{htav05}}=1,14$$

Braking distances for tyre-ground friction coefficient $\mu = 0,8$

Overall braking force

$F_{hc}=F_{hpśr08}*2+F_{htśr08}*2=50345[N]$

Braking distance for braking with start speed equal to $V_s=60\ \left[\frac{km}{h}\right]=16,67\left[\frac{m}{s}\right]$:

s_h=17,66[m]

Braking distance for braking with start speed equal to $V_s=100\ \left[\frac{km}{h}\right]=27,78\left[\frac{m}{s}\right]$:

s_h=49,04[m]

Summary

For effective use of tyre-ground friction coefficient equal to $\mu = 0,8$ can be achieved by using p_h=11,5[Mpa] braking pressure for car $m_s = 6400$ [kg] of mass, area of the braking pistons has to be multiplied by:

- Front brake

  n_p=2 $A_{pnew}=n_p*A_{pold}$

- Rear brake

  n_t=1,15 $A_{tnew}=n_t*A_{told}$

Comments

Friction coefficient $\mu = 0,8$ is required by aviation regulations. It is taken for calculations of airplane landing gears and during laboratory tests of landing gears it is proven that friction coefficient is no less than $\mu = 0,8$.

In some cases friction coefficient $\mu > 1$ is achieved during landing gear laboratory tests performed in Institute of Aviation Landing Gear Department. This is also proven by literature, for example by „Budowa samochodów Układy hamulcowe i kierownicze" - A. Reński Oficyna wydawnicza Politechniki Warszawskiej 2004 [1].

3. SUMMARY

Based on dynamic analysis of the car - dissipated energy by the front axle brakes should be equal to E_{kp}=557940[J] and for rear axle brakes should be equal to E_{kt}=351720[J] when tyre-ground friction coefficient is equal to $\mu \approx 0,5$.

During dynamic test of the rear drum brake $M_{ht} \approx 1750$Nm was obtained with braking pressure pht=9[MPa]. For tyre-ground friction coefficient $\mu \approx 0,5$ (or braking distance equal to 30m), braking moment taken from numerical analysis is equal to M_{htA}=2567[Nm] while the same moment taken from test results non linear analysis should be equal to M_{htA}=2497[Nm] for braking pressure equal to pht=11,5[MPa].

Due to 2 % less moment value obtained during dynamic tests compared to calculated during numerical analysis it can be assumed that original brake system was calculated for tyre-ground friction coefficient $\mu \approx 0,5$ (or braking distance equal to 30m). In this case front disc brake should generate braking moment equal to M_{hp}=3961[Nm] what can be achieved by existing brake system with the braking pressure p_{hp}=11,5[MPa].

With constant braking pressure equal to p_{hp}=11,5[MPa] current area of the braking pistons (for one wheel) is equal to:

A_p = 3040 mm^2, A_t = 314 mm^2

When tyre-ground friction coefficient is equal to $\mu \approx 0,5$ braking pistons area will not change and will be equal to:

$A_{p0,5}$ = 3040 mm^2, $A_{t0,5}$ = 314 mm^2

Therefore when tyre-ground friction coefficient is equal to $\mu \approx 0,8$ braking pistons area will change and will be equal to:

$A_{p0,8}$ = 6080 mm^2, $A_{t0,8}$ = 361,1 mm^2

New braking system will be optimized in order to make braking process more effective what can result with reducing braking distance from about 30 [m] to about 18 [m] (according to analytical data) what gives 41% improvement.

4. REFERENCES

Reński A. 2004 “Budowa samochodów Układy hamulcowe i kierownicze”, Oficyna wydawnicza Politechniki Warszawskiej

Institute of Aviation Landing Gear Report, 2009, “ Laboratory tests of armored cars brake lining sectors with termovision made with IL-68 test rig”, 26/LW/2009, Institute of Aviation

Institute of Aviation Landing Gear Report, 2009, “Armored Car Drum Brake Laboratory Tests”, 27/LW/2009, Institute of Aviation

Institute of Aviation Landing Gear Report, 2009, “Car Braking System Dynamics Analysis”, 36/BZ/2009, Institute of Aviation

AMZ-Kutno sp. z o.o. website, http://www.amz,pl

Landing Gear Department website, http://www.cntpolska.pl/index.php/landing-gears-department/about-us

Institute of Aviation website, http://www.ilot.edu.pl

AUTHOR BIOGRAPHIES

WOJCIECH KOWALSKI, born in Warsaw, Poland and went to the Warsaw Academy of Technology, received his M.Sc. in Engineering in 1970, Faculty of Power and Aeronautical Engineering in Applied Mechanics and 2000 PhD from Warsaw University of Technology, Faculty of Automotive and Construction Machinery Engineering. Now he is working at Warsaw Institute of Aviation. His current professional area are: load and stress analysis, construction dynamics, test parameters determination, results analysis and interpretation of tests and calculations

E-mail: wojciech.kowalski@ilot.edu.pl

ZBIGNIEW KONRAD SKORUPKA, born in 1977 in Warsaw, Poland. Finished Warsaw Academy of Technology, where he studied Shaping Machines Design and Steering. He received his M.Sc. in 2004. Currently he is working at Landing Gear Department of Warsaw Institute of Aviation. His current professional area are: use of smart materials in landing gear, general design and testing of landing gear, brakes and test stands.

E-mail: zbigniew.skorupka@ilot.edu.pl

RAFAŁ KAJKA born in 1977, received his M.Sc. in Engineering (2000) from Warsaw University of Technology, (Poland). Directly after graduation he started his doctoral studies and in 2005 he received PhD from Warsaw University of Technology, Faculty of Automotive and Construction Machinery Engineering. Since 2001 he is working at Institute of Aviation, Warsaw initially as a strength engineer and from 2007 as a Manager of Landing Gear Department, Institute of Aviation, Warsaw, Poland.

E-mail: rafal.kajka@ilot.edu.pl

JAN AMBORSKI born in 1972 received his M.Sc. in Engineering in 1999 from Warsaw University of Technology, Faculty of Power and Aeronautical Engineering and 2006 PhD from Warsaw University of Technology, Faculty of Automotive and Construction Machinery Engineering. Now he is working at Warsaw Institute of Aviation.

E-mail: jan.amborski@ilot.edu.pl

BEHAVIORAL MICROSIMULATION OF A DUAL INCOME TAX REFORM: A MIXED-LOGIT APPROACH

Gerhard Wagenhals
Department of Statistics and Econometrics
University of Hohenheim
D-70593 Stuttgart, Germany
Email: G.Wagenhals@uni-hohenheim.de

KEYWORDS

Tax Microsimulation, Discrete Choice, Mixed Logit

ABSTRACT

Simulation studies of the economic impact of dual income taxes are almost always based on general equilibrium models. They assume one representative household. Their results are sensitive to one behavioral parameter, the labor supply elasticity, which is assumed to be given a priori — instead of being estimated.

This paper shows how to model the labor supply incentive effects of a Dual Income Tax reform based on a sample of thousands of households representative for the population in Germany and using flexible mixed logit simulation estimators.

INTRODUCTION

The Dual Income Tax (DIT) remains a hot topic worldwide. Its possible advantages and drawbacks are discussed not only in the European Nordic countries which have introduced them some years ago, but also in the rest of Europe (see e.g. Genser and Reutter, 2007), in Japan (Morinobu, 2004), and in Canada (Sørensen, 2007). Also in Germany, economists and policy makers consider a dual income tax as an option for a fundamental tax reform. Recently, the German Council of Economic Experts (2008) published an expertise commissioned by the German Ministry of Finance. This report strongly favors the introduction of a Dual Income Tax reform.

Previous economic research on the impact of this proposal has concentrated on long-run effects and is mainly based on general equilibrium simulation models. The results of these exercises are largely robust with respect to the choice of the behavioral elasticities, with one important exemption: the labor supply elasticity. Actually, the labor supply elasticity is the only behavioral parameter that is crucial for the long run effects of a DIT. (See e.g. Radulescu (2007) for Germany, or Keuschnigg and Dietz (2007), p. 204, for Switzerland.) General equilibrium simulation studies assume that the household sector can be modelled by a traditional Ramsey model with only *one* single "representative" agent characterized by only *one* labor supply elasticity. Population based microeconometric analyses (pathbreaking: Blundell et al., 1998) show, however, that in the population labor supply elasticities vary widely depending on gender, number of children, regional and other factors. This suggests to supplement existing macroeconomic DIT studies by microeconometric simulation analyses.

The main contribution of the present paper is a simulation analysis of the incentive effects of the most recent DIT proposal for Germany based on a behavioral microeconometric model. It is the only evaluation of the behavioral effects of the income tax amendment EStG-E proposed by the Council of Economic Experts based on a mixed logit approach. This is an improvement to previous studies using a traditional conditional logit model and older data sets (Bach and Steiner, 2007; Wagenhals and Buck, 2009), because the conventional IIA assumption implicit in the traditional model is strongly rejected by the data.

The rest of the paper is organized as follows. The next section describes the data: the generation of the base data set, the definition of the tax base, with special reference to the calculation of capital income and labor income, and the tax schedule used. Then, two sections describe discrete choice models for single persons as well as for cohabiting and married couples. They provide mixed logit estimation and calibration techniques and present empirical results. The last section concludes.

DATA

Base Data Set

My base data set is drawn from the 2005 wave of the German Socio-Economic Panel (GSOEP). I merge some retrospective data from the 2006 wave, such that the base data set refers to 2005, the same fiscal year the German Council of Economic Advisors reform proposal refers to.

Choice alternatives are generated using GMOD, a tax-benefit microsimulation model for Germany developed by the author. GMOD calculates personal income taxes, social security contributions and benefits. It allows for the standard benefits and tax concessions such as housing benefits and child-benefits, allowances for child-raising, child-raising leave and maternity as well as assistance for education or vocational training. Furthermore, it accounts for tax abatements for dependent children and for the education of dependent children, for child-care, tax credits for single parents, maintenance payments and

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

income-splitting for married couples.

Tax Base

A dual income tax differentiates between capital and labor income and taxes these differently. So I have to derive two tax bases, one for capital income, and one for other sources of household income, called "labor income".

Currently, GMOD calculates seven sources of income, because the current German Income Tax Law (Einkommensteuergesetz, EStG) levies one tax schedule on the sum of income from the following exhaustive list of seven sources of income: (1) income from agriculture and forestry (§13 EStG), (2) income from trade or business (§15 EStG), (3) income from independent personal services (§18 EStG), (4) income from dependent personal services, i.e. wages, salaries and retirement benefits of civil servants (§19 EStG), (5) income from investment of capital (§20 EStG), (6) income from rentals and royalties (§21 EStG), and (7) other income designated in §22 EStG, e.g. notational return on investment of a pension from statutory pensions insurance. Gross earnings from all of these sources are calculated by GMOD based on information available in my base data set described above, on the German income tax law and on income tax directives. Net income from the first three sources is calculated on the accrual basis and called "profit-based income". Net income from the other four sources is defined as the excess of total receipts over income-related expenses.

According to the German Income Tax Law (EStG-E) as proposed by the German Council of Economic Experts (2008), there will be four categories of income (see §2 EStG-E): (1) income from business activities (§13, §15 and §18 EStG-E), (2) income from employment (§19 EStG-E), (3) capital income (§20, §21, and §22 EStG-E), and (4) derived income (§23 EStG-E).

To map the traditional seven sources of income to the new categories capital and labor income I proceed as follows: (1) Income from business activities corresponds to traditional "profit based income". (2) Income from employment corresponds to the traditional income from dependent personal services. (3) Income from capital assets is derived from traditional income from capital investments (§20 EStG) and income from rentals and royalties (§21 EStG). (4) Derived income corresponds to traditional "other income" designated in §22 EStG. In my base data set, I do not have information on income from private sale transactions mentioned in §22 EStG-E, so I have to ignore it. I assume that the cash method of accounting is used with respect to income from business activities and decompose profits in a capital and a labor share. Labor income is calculated by adding income from employment, the labor share in profits and derived income. Capital income consists in income from capital investments and the interest share of profits.

The labor income tax base includes wages, salaries (including the employers' calculatory salaries) and civil pensions. The capital income tax base includes business profits, dividends, capital gains, interest and rental income. Taxable labor income and taxable capital income are obtained by subtracting personal allowances and other deductions from the respective tax base. The savings allowance of 750 Euro for the income from capital investments (§20 Section 4 EStG) will be abolished (SVR et al. 2006:109ff.).

The decomposition of profit-based income in a capital and a labor share is the crux of the DIT. The calculatory salary, i.e. the labor income of the self-employed, is hard for an individual to measure and even harder for tax authorities to verify. I use the following trick: First, I estimate a Mincer-type wage function based on observable characteristics on the sub-sample of wage earners. In my data I observe determinants of wages for all individuals. Therefore, I am able to predict the calculatory salary for all self-employed individuals. Finally, I derive their capital income as the residual. (See Wagenhals and Buck, 2009, for details about this decomposition approach.) In my view, this approach improves upon the procedure of using an arbitrary sharing rule (see e.g. Gottfried and Witczak (2009)). In any case, due to data constraints, I did not have the option to compute calculatory salaries for the self-employed as residual profits.

Tax Schedule

The dual income tax combines a progressive tax schedule for labor income with a flat tax rate on capital income.

I assume that labor income is taxed according to the current income tax schedule (§32 a EStG), and that capital income is taxed with a rate of 25 percent (including the solidarity surcharge). To avoid legal concerns and a potential deterioration with respect to the current legal position I follow the Council of Economic Experts and use a stretched tax scale: The taxation of capital income is incorporated in the tax schedule in terms of a proportional zone. The length of this zone depends on the amount of taxable capital income.

Figure 1 compares the marginal tax schedule (based on §32 a EStG) with the DIT schedule (based on §32 a EStG-E) for taxpayers with fixed taxable capital incomes of 10,000, 20,000 and 30,000 Euro.

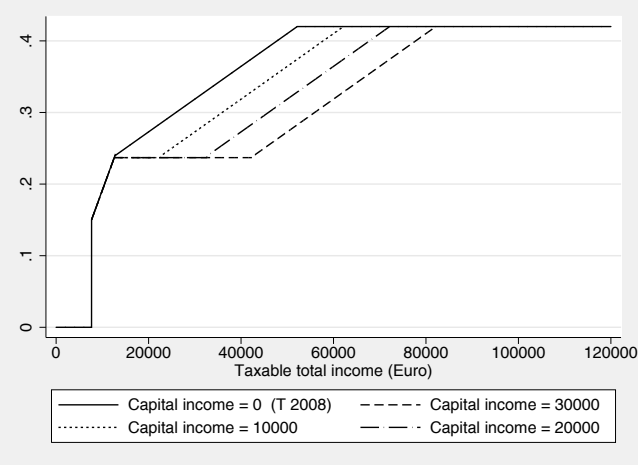
LABOR SUPPLY OF SINGLE PERSONS

To quantify the labor supply incentives of a DIT introduction, I use a discrete choice structural labor supply model. The basic idea is to replace the budget set of a household with a finite number of points, and optimize over this set of points. I first set out the theory, estimation and simulation results for single persons. In the following Section, I turn to persons living in couples.

Theory

I represent any individual's choice set by a six-state labor supply regime and approximate actual hours per week

Figure 1: Marginal Tax Rates



Source: Own calculations.

h^a by hours levels $h \in \mathscr{H} := \{0, 10, 20, 30, 40, 50\}$ applying the following rounding rule

$$\begin{aligned} h &= 0 \text{ if } h^a < 5 \\ &= 10 \text{ if } 5 \leq h^a < 15 \\ &\dots \\ &= 50 \text{ if } h^a \geq 45. \end{aligned}$$

For all elements h in the choice set $\mathscr{H}$ I use GMOD to calculate household net incomes as

$$c(h) = wh + \mu - T(h, w, \mu \mid \mathbf{x}) \tag{1}$$

where w denotes the gross wage rate, μ is income from sources other than employment and $T(\cdot)$ is the tax-benefit function conditional on a vector of observed characteristics $\mathbf{x}$. I assume that preferences can be represented by a utility function U and that individuals act as if to maximize utility

$$\max_{h \in \mathscr{H}} U(c(h), \bar{h} - h \mid \mathbf{x}) \tag{2}$$

subject to the budget constraint

$$c(h) \leq wh + \mu - T(h, w, \mu \mid \mathbf{x}) \tag{3}$$

where $\bar{h}$ denotes total time endowment.

To obtain random utilities, I add state-specific random errors $e(h)$ to utilities for all states $h \in \mathscr{H}$. This gives random utilities

$$U^*(h) := U(c(h), \bar{h} - h \mid \mathbf{x}) + e(h) \tag{4}$$

If the state-specific random errors are i.i.d. Type I extreme value distributed, then the probability P of working h^j hours is

$$P(h = h^j \mid \mathbf{x}) = \frac{\exp[U(c(h), \bar{h} - h \mid \mathbf{x})]}{\sum_{h^k \in \mathscr{H}} \exp[U(c(h^k), \bar{h} - h^k \mid \mathbf{x})]} \tag{5}$$

For the specification of the utility function, I follow the tradition started by Keane and Moffitt (1998) and choose a flexible form quadratic direct utility function. Written in terms of individual consumption $c = c(h)$ and leisure $l := \bar{h} - h$ I obtain

$$U(c, l) = \alpha_{cc} c^2 + \alpha_{ll} l^2 + \alpha_{cl} cl + \beta_c c + \beta_l l$$

where $\alpha_{cc}, \alpha_{ll}, \alpha_{cl}, \beta_c$ and β_l denote unknown parameters. I assume that preferences vary through taste-shifters on income and leisure coefficients:

$$\begin{aligned} \beta_c &= \gamma_{c_0} + \mathbf{x}\boldsymbol{\gamma}_\mathbf{c} \\ \beta_l &= \gamma_{l_0} + \mathbf{x}\boldsymbol{\gamma}_\mathbf{l} \end{aligned}$$

where $\gamma_{c_0}, \boldsymbol{\gamma}_\mathbf{c}, \gamma_{l_0}$ and $\boldsymbol{\gamma}_\mathbf{l}$ denote unknown coefficients and $\mathbf{x}$ is a (row) vector of individual characteristics, including age, region of residence, number of children in different age brackets, and (following van Soest, 1995) dummy variables for part-time categories in order to capture the disutility of inflexible arrangements.

I deal with unobserved wage rates by estimating the expected market wage rates conditional on observed characteristics. I first estimate a reduced form participation equation, get the Mill's rate and use it in a Mincer-type wage equation to correct for sample selection bias. I use the estimates of this sample selection model to draw wage rates for non-workers, conditional on observed characteristics. Due to the nonlinear nature of the labor supply model, replacing wage rates by their predictions leads to inconsistent estimates, even if the wage predictions themselves are unbiased. To account for this, wage rate prediction errors are incorporated as additional unobserved error terms.

Estimation of the unknown preference parameters is based on a mixed logit model proposed by McFadden and Train (2000). Under mild regularity conditions, it can approximate the choice probabilities of any discrete choice model derived from random utility maximization as closely as desired. Under the assumption that the income coefficients are normally distributed and all other coefficients are fixed I proceed by maximum simulated likelihood.

Simulation. Before the tax reform, for all individuals and all states $h \in \mathscr{H}$ I obtain net incomes

$$c_0(h) = wh + \mu - T_0(h, w, \mu \mid \mathbf{x})$$

After the tax reform, I get

$$c_1(h) = wh + \mu - T_1(h, w, \mu \mid \mathbf{x})$$

To simulate labor supply responses, I evaluate utilities $U(c(h), \bar{h} - h \mid \mathbf{x})$ over the net incomes $c_0(h)$ and $c_1(h)$ for all hours alternatives $h \in \mathscr{H}$. I allocate each individual to the utility maximum under each scenario and obtain

$$h_{0(\cdot)} = \arg\max_{h \in \mathscr{H}} U(c_0(h), \bar{h} - h \mid \mathbf{x})$$

$$h_{1(\cdot)} = \arg\max_{h \in \mathscr{H}} U(c_1(h), \bar{h} - h \mid \mathbf{x})$$

Finally, I simulate labor supply responses by comparing $h_{0(\cdot)}$ and $h_{1(\cdot)}$ for each individual. To this end, I follow the calibration procedure described by Creedy and Kalb (2005, page 720 et seq.). For each household I repeatedly draw a vector of unobserved utility components from a Type I extreme value distribution such that utility is maximized at the observed hours category. I then calculate post reform incomes, compute the new utility maximizing choice and allocate each person to the most probable state following each random draw. For each individual, this exercise is repeated 100 times. This generates a distribution of post-reform hours worked, conditional on the observed pre-reform hours. Finally, I build transition tables by allocating each observation to the cell which yields maximum utility for each draw.

Empirical Results

Sample Selection. The starting point for my sample is the base data file described above. First, I concentrate on single adult respondents. I exclude persons younger than 25 or older than 55 years of age, persons in education, pensioners, persons doing compulsory community or military services, persons receiving profit incomes only and civil servants. After dropping persons with missing observations of crucial variables, I receive a sample with 1,116 single men and another sample with 1,312 single women.

Estimation. The main preference parameter estimates for single men and single women are given in Table 1. The estimated parameter values are consistent with economic theory. The marginal utility of net income and of leisure are statistically significant at least at the five percent level, they are positive and declining with income. The interaction effect between leisure and income is practically zero. Not surprisingly, there is less desire to work if an individual is handicapped, or if there is a nursing case in the family. For single mothers, there is less desire to work, the effect being smaller for older children. The main difference between male and female preferences is the role of children: While the number of children in different age groups has the expected sign and magnitude for women, these variables were not significant for men and so were dropped.

In Table 1 I do not report the estimates of the part-time dummies for part-time choice opportunities. For men and women, they all are negative and highly significant. This reflects the fact that low demand for part-time workers requires more effort to find part-time employment. Furthermore, all estimated standard errors of the random coefficients were highly significant. This suggests considerable unobserved heterogeneity of preferences. The traditional conditional logit approach is strongly rejected.

Simulation. Tables 2 and 3 present the simulation results for the labor supply of single persons. The last column gives the distribution of labor supply before the reform, the last row refers to the distribution after the reform. The numbers inside the matrix are row percentages.

My results suggest that — in a short run partial equilibrium view — the DIT reform suggested by the German Council of Economic Experts (2008) will generate only small labor supply reactions. For single persons, on average, they will be slightly positive.

LABOR SUPPLY OF COUPLES

Theory

For married or cohabiting couples I allow for joint decision making. Each partner may account for the decision of the other partner when deciding on hours worked. I assume that each household member selects one of six regimes: non-participation or one of five employment states $\hbar \in \mathscr{H} = \{0, 10, 20, 30, 40, 50\}$ (the elements denoting hours per week). Thus, the choice set for couples is $\mathscr{H} \times \mathscr{H}$. Actual individual working hours observed in the data are rounded (as above) to fit the elements in this set.

I assume that preferences of a couple may be represented by a flexible quadratic utility function

$$\begin{aligned} U(c, l_f, l_m) &= \alpha_{cc}c^2 + \alpha_{mm}l_m^2 + \alpha_{ff}l_f^2 \\ &+ \alpha_{cm}cl_m + \alpha_{cf}cl_f + \alpha_{fm}l_f l_m \\ &+ \beta_c c + \beta_m l_m + \beta_f l_f \end{aligned}$$

Here $l_m := \bar{h} - h_m$, $l_f := \bar{h} - h_f$; l denotes leisure and h hours worked of male (m) or female (f) persons, while c denotes their joint net income. The α and β coefficients are unknown population parameters. The sign of α_{fm} indicates whether male and female leisure are substitutes or complements. Similar to the case of single persons, some preference parameters depend on personal, household and other characteristics. Supplementing representative household utility I add stochastic terms accounting for state specific errors and finally derive the probability of choosing any consumption-leisure combination in the set of feasible household decisions. Estimation proceeds via mixed logit and simulation by calibration as described above. I derive household gross earnings assuming state invariant male and female gross wage rates, and calculate the corresponding state specific net household income for each hours combination in the choice set $\mathscr{H} \times \mathscr{H}$ using GMOD and my base data set described above.

Empirical Results

Sample Selection. Starting point for my analysis is again the base data file described above, now concentrating on couples. I apply the sample selection criteria as described for singles to both partners and obtain a sample of 2,015 couples.

Table 1: Estimated Preference Parameters, Singles

	Single Men	Single Women
Income	0.0680	0.183**
	(0.0363)	(0.0630)
Income^2	−0.000264	−0.00291*
	(0.000403)	(0.00121)
Leisure	0.371***	0.842***
	(0.0806)	(0.123)
Leisure^2	−0.00287***	−0.00469***
	(0.000399)	(0.000496)
Leisure*income	−0.00128	−0.00233**
	(0.000653)	(0.000779)
Leisure*age	−0.00425	−0.0159**
	(0.00361)	(0.00536)
Leisure*age^2	0.0000545	0.000205**
	(0.0000464)	(0.0000682)
Leisure*(East Germany?)	0.0218*	−0.00203
	(0.00872)	(0.00982)
Leisure*(Nursing case in family?)	0.0126	−0.00950
	(0.0240)	(0.0215)
Leisure*foreign?	0.0297**	−0.0209
	(0.0112)	(0.0190)
Leisure*(high education?)	−0.0355**	−0.0289**
	(0.0113)	(0.0104)
Leisure*(low education?)	0.0229**	0.0300*
	(0.00848)	(0.0147)
Leisure*handicapped?	0.0363**	0.00161
	(0.0133)	(0.0226)
Leisure*(no. of kids under 6)		0.0700***
		(0.0122)
Leisure*(no. of kids age 6-16)		0.0358***
		(0.00662)
SD		
Income	0.0902***	0.154***
	(0.0225)	(0.0307)
Observations	1116	1312

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2: Labor Supply Transition Matrix for Single Men

	Post-reform hours						
Pre-reform hours	0	10	20	30	40	50	% (row)
0	16.55	0.00	0.01	0.26	0.56	0.20	17.59
10	0.00	2.12	0.00	0.04	0.05	0.08	2.29
20	0.00	0.00	1.79	0.01	0.02	0.04	1.85
30	0.00	0.00	0.00	17.43	0.20	0.13	17.77
40	0.01	0.00	0.00	0.01	42.05	0.24	42.30
50	0.00	0.00	0.00	0.00	0.03	18.18	18.21
% (column)	16.56	2.12	1.79	17.75	42.91	18.87	100.00

Source: Own calculations. Any summing errors are due to rounding.

Table 3: Labor Supply Transition Matrix for Single Women

	Post-reform hours						
Pre-reform hours	0	10	20	30	40	50	% (row)
0	17.54	0.07	1.27	3.32	2.78	0.21	25.19
10	0.04	4.02	0.18	0.66	0.53	0.08	5.51
20	0.04	0.02	8.49	0.46	0.42	0.08	9.50
30	0.17	0.00	0.02	20.72	0.30	0.05	21.27
40	0.17	0.02	0.03	0.05	31.77	0.02	32.05
50	0.14	0.01	0.10	0.09	0.08	6.07	6.49
% (column)	18.11	4.13	10.09	25.29	35.88	6.50	100.00

Source: Own calculations. Any summing errors are due to rounding.

Estimation. The main preference parameter estimates for married and cohabiting couples are given in Table 4. The estimated parameter values are consistent with economic theory. The marginal utility of both partners' leisures and the marginal utility of net income are highly significant, positive and declining with income. The interaction effect between male and female leisure is statistically not different from zero and practically unimportant. Not surprisingly, there is less desire to work for mothers, the effect being smaller for older children.

Due to space restrictions, in Table 4 I do not report the estimates of the part-time dummies for part-time choice opportunities. For both sexes, they all are negative and highly significant. As in the case of singles, this reflects the fact that low demand for part-time workers requires more effort to find part-time employment. Again, all estimated standard errors of the random coefficients were highly significant. As for singles, this suggests considerable unobserved heterogeneity of preferences of couples. Again, the traditional conditional logit approach is strongly rejected.

Simulation. Tables 5 and 6 show that the partial equilibrium impact of the reform proposal on the labor supply of couples is small. Positive incentives are most likely for women in couples, and here especially married females.

If I finally aggregate over persons living as singles and living in couples we find a positive incentive effect. If for the moment we accept the SVR et al. (2006) finding of an 1.1 percent reform-induced increase in labor demand, annual working time will — on average — increase. This effect, combined with the smaller tax burden on capital income, is reflected in an increasing aggregate net income.

CONCLUSION

My main concern has been to evaluate the incentive effects of the DIT reform proposed by the German Council of Economic Experts (2008). Instead of invoking the assumption of *one* given labor supply elasticity as current general equilibrium simulation models do, I estimate these elasticities. Based on a mixed logit simulation estimation, I find that labor supply incentive effects are small, but — on average — positive. Effects are most pronounced for married women.

Acknowledgements

The data used in this publication were made available by the German Socio-Economic Panel Study (GSOEP) at the German Institute for Economic Research (DIW), Berlin.

REFERENCES

Bach, S. and Steiner, V. (2007). Steuerreformpläne im empirischen vergleich. In Zwick, M. and Merz, J., editors, *MITAX – Mikroanalysen und Steuerpolitik*, Statistik und Wissenschaft, Band 7, pages 54–83, Wiesbaden. Statistisches Bundesamt. Beiträge zur wissenschaftlichen Konferenz am 6. und 7. Oktober 2005 in Lüneburg.

Blundell, R., Duncan, A., and Meghir, C. (1998). Estimating labor supply responses using tax reforms. *Econometrica*, 66(4):827–861.

Creedy, J. and Kalb, G. (2005). Discrete hours labour supply modelling: Specification, estimation and simulation. *Journal of Economic Surveys*, 19(5):697–734.

Genser, B. and Reutter, A. (2007). Moving towards Dual Income Taxation in Europe. *FinanzArchiv: Public Finance Analysis*, 63(3):436–456.

German Council of Economic Experts, Max Planck Institute for Intellectual Property, Competition and Tax Law, C. (2008). *Dual Income Tax. A Proposal for Reforming Corporate and Personal Income Tax in Germany*, volume 39. Physica-Verlag, Heidelberg.

Gottfried, P. and Witczak, D. (2009). Reformoption Duale Einkommensteuer - Aufkommens- und Verteilungseffekte. IAW-Diskussions-Papier 58, Institut für Angewandte Wirtschaftsforschung, Tübingen.

Keane, M. and Moffitt, R. (1998). A structural model of multiple welfare program participation and labor supply. *International Economic Review*, 39(3):553–589.

Table 4: Estimated Preference Parameters, Couples

	Coefficient	Std. Err.
Income	0.0644^{**}	(0.0197)
Income^2	0.0000154	(0.0000644)
Female's leisure	0.486^{***}	(0.108)
(Female's leisure)^2	-0.00364^{***}	(0.000661)
Male's leisure	0.268^{*}	(0.107)
(Male's leisure)^2	-0.00319^{***}	(0.000315)
(Female's leisure)*(male's leisure)	-0.000448	(0.000282)
(Female's leisure)*(female's*age)	-0.00333	(0.00360)
(Female's leisure)*(female's*age)^2	0.0000542	(0.0000450)
(Female's leisure)*(East Germany?)	-0.0434^{***}	(0.00763)
(Female's leisure)*(no. of kids under 6)	0.0701^{***}	(0.0101)
(Female's leisure)*(no. of kids aged 6-16)	0.0301^{***}	(0.00492)
(Female's leisure)*(nursing case in family?)	0.0346	(0.0181)
(Female's leisure)*(married?)	0.0320^{**}	(0.0108)
(Male's leisure)*(male's*age)	0.00514	(0.00470)
(Male's leisure)*(male's*age)^2	-0.0000484	(0.0000555)
(Male's leisure)*(East Germany?)	0.00857	(0.00881)
(Male's leisure)*(no. of kids under 6)	0.00257	(0.00715)
(Male's leisure)*(no. of kids aged 6-16)	0.000312	(0.00439)
(Male's leisure)*(nursing case in family?)	0.0181	(0.0126)
(Male's leisure)*(married?)	-0.0158	(0.0110)
SD		
Income	0.0745^{**}	
	(0.0233)	
Sample Size	2015	
Log-likelihood	-14161577	

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: A question mark means that the variable is binary, coded 1 for a "Yes" and 0 for a "No".

Table 5: Labor Supply Transition Matrix for Men in Couples

	Post-reform hours						
Pre-reform hours	0	10	20	30	40	50	% (row)
0	9.41	0.00	0.01	0.03	0.08	0.06	9.59
10	0.00	0.48	0.00	0.00	0.00	0.00	0.49
20	0.00	0.00	1.00	0.00	0.01	0.00	1.02
30	0.03	0.00	0.00	16.38	0.10	0.10	16.61
40	0.08	0.00	0.02	0.12	50.31	0.13	50.66
50	0.04	0.00	0.01	0.05	0.18	21.35	21.62
% (column)	9.55	0.49	1.05	16.58	50.68	21.65	100.00

Source: Own calculations. Any summing errors are due to rounding.

Table 6: Labor Supply Transition Matrix for Women in Couples

	Post-reform hours						
Pre-reform hours	0	10	20	30	40	50	% (row)
0	33.78	0.14	0.26	0.19	0.24	0.02	34.63
10	0.02	10.88	0.05	0.07	0.02	0.00	11.05
20	0.01	0.00	15.71	0.02	0.06	0.00	15.82
30	0.01	0.02	0.03	16.63	0.07	0.04	16.81
40	0.00	0.00	0.01	0.02	17.96	0.04	18.04
50	0.00	0.00	0.01	0.01	0.01	3.62	3.66
% (column)	33.83	11.05	16.07	16.95	18.37	3.73	100.00

Source: Own calculations. Any summing errors are due to rounding.

Keuschnigg, C. and Dietz, M. (2007). A growth oriented dual income tax. *International Tax and Public Finance*, 14:191–221.

McFadden, D. and Train, K. (2000). Mixed MNL model for discrete response. *Journal of Applied Econometrics*, 15:447–470.

Morinobu, S. (2004). Capital income taxation and the Dual Income Tax. Technical report, Policy Research Institute, Ministry of Finance of Japan. PRI Discussion Paper Series (No.04A-17).

Radulescu, D. M. (2007). *CGE Models and Capital Income Tax Reforms: The Case of a Dual Income Tax for Germany*. Springer, Berlin-Heidelberg, Lecture Notes in Economics and Mathematical Systems, Vol. 601.

Sørensen, P. B. (2007). The nordic dual income tax: Principles, practices, and relevance for Canada. *Canadian Tax Journal*, 55(3):557 – 602.

van Soest, A. (1995). Structural models of family labor supply: A discrete choice approach. *Journal of Human Resources*, 30(1):63–88.

Wagenhals, G. and Buck, J. (2009). Implementing a Dual Income Tax in Germany. effects on labor supply and income distribution. *Journal of Economics and Statistics*, 229(1):84–102.

AUTHOR BIOGRAPHY

Prof. Dr. rer. pol. habil. GERHARD WAGENHALS is Full Professor of Statistics and Econometrics at the University of Hohenheim in Stuttgart, Germany, and Research Fellow at the Institute for the study of Labor (IZA) in Bonn. He received his diploma in economics from the University of Tübingen, Germany, in 1976; his doctoral and habilitation degrees were received from the University of Heidelberg, in 1980 and 1984. He worked as a postdoctoral research fellow at the Department of Economics, University of Pennsylvania (1980–1982), as an Associate Professor at the University of Heidelberg (1986–1988), as a Visiting Professor at the Department of Economics, University of Bern (Switzerland) (1989–1990), as a Professor of Computational Economics at the Department of Economics, University of Paderborn (1990–1992), and since then as Full Professor of Statistics and Econometrics at the University of Hohenheim. His key research interests are microsimulation modelling and microeconometrics. His e-mail address is `G.Wagenhals@uni-hohenheim.de` and his web page can be found at `http://www.statistik.uni-hohenheim.de`.

A FRAMEWORK FOR EMERGENCY DEPARTMENT CAPACITY PLANNING USING SYSTEM DYNAMICS APPROACH AND THE THEORY OF CONSTRAINTS PHILOSOPHIES

Norazura Ahmad
Noraida Abdul Ghani
Anton Abdulbasah Kamil
School of Distance Education, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia

Razman Mat Tahar
Faculty of Technology Management, Universiti Malaysia Pahang, 26300 Gambang, Kuantan, Pahang, Malaysia

KEYWORDS:
Capacity planning, Emergency Department, System Dynamics, Theory of Constraints

ABSTRACT

Patient waiting time and service delivery problems characterize health care services and are more acute in an emergency department (ED). With more patients needing care and fewer resources to care for them, ED that operates at or above capacity is inevitable. This paper is a review of work-in-progress of a study being conducted in a government hospital in Penang, Malaysia. This paper proposes a hybrid of System Dynamics (SD) approach and the Theory of Constraints (TOC) in solving health capacity planning. The potential combination of these methods will be reviewed that is hoped to reveal the synergy between the established methods in addressing the health capacity planning options for the long-term future.

INTRODUCTION

Capacity planning is a strategic element in designing a system. It involves many decisions which may have long-term consequences for an organization. In capacity planning, all resources should be planned in a manner that can facilitate work flow across every process in the system. Ideally, capacity has to be matched with demand. However, in reality, capacity planning is a complex problem due to demand uncertainty, particularly in the service sector because services often cannot be stored (Barnes 2008). Hospital is a part of the service sector. With growing population and health care costs mounting, hospital is facing growing demand with lack of quantifiable data regarding its capacity (Ballard and Kuhl 2006). Lack of resource capacity, such as beds, doctors and nurses will cause overcrowding that can reduce health care quality by increasing the potential of medical errors and long waiting times, which leads to customer dissatisfaction with services (DeLia 2005). This scenario is inherently difficult in emergency care, because capacity decision is a major determinant for providing services over a specified time interval.

Acting as a critical entry to hospitals, the emergency department (ED) receives emergency patients at random and must be admitted with minimum delay and with priority over elective patients. A common perception of an efficient ED services usually relates to the availability of spaces, staff and equipments for emergency care in a timely manner (Kolb et al. 2007). Patients not only look at the result on the treatment provided, but they also consider the process of giving treatment. The smooth running processes in the ED are closely related to capacity planning (Baesler et al. 2003).

Much research in ED has focused on solving the operational problems that solves 'snapshot' problems and are not capable of drawing attention to the interdependency of the problem understudy. However, as many units integrate in a hospital, the ED service delivery also depends on other unit like the labs and wards. This problem is addressed here by proposing that simulation could be used to understand the structure and behavior of the system and is capable to create some strategic options for the future. It is worth noting that problems in the ED sometimes are originated from factors external to ED (Kolb et al. 2007).

In order to provide a better service and able to treat the surge in demand of different case mixes, this study attempts to give a 'big picture' view of the delivery process in ED. From a broader perspective, we hope to create a better understanding of the system that can be useful for policymaking. This paper starts with problems faced by a government's ED, followed by pertinent literature and the proposed methodology for the problem. Finally, the paper ends with a brief conclusion and the direction of future research.

MOTIVATION OF THE STUDY

In Malaysia, lack of medical staff has led to poor response time at most EDs around the country (News Strait Times 2008). The doctor-patient ratio in the country is recorded as 1: 1,105 (Health Informatics Centre 2009) which is lagged behind the developed

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

countries. The shortage of doctors is predicted to continue and be more serious in government hospitals (Mohamed 2003). As a result, the public begins to advocate that government hospitals are not achieving the standard of efficiency displayed by private hospitals (Pillay 2009).

With the increased level of population, demand for health services particularly in the ED will also be increased (Hong and Ghani 2006). For instance, in the four-year period in Penang from 2005 to 2008, the number of attendances to an ED increased from 101,841 to 113,111. It represents an overall increase of 11.1%, which is higher than the 5.3% increase in Penang's estimated total population in the same period (SERI 2009). This situation has positioned the ED under increasing pressure to treat its patients within the expected timeline. With more patients needing care and fewer resources to care for them, ED that operates at or above capacity is inevitable. If this trend persists, what will happen to ED should there be a surge in demand in the future?

ED surge can be defined as a sudden increase in the demand for ED services while surge capacity is the resource availability in the hospital (McCarthy et al. 2006). Daily ED surge causes ED overcrowding in facilities that have inadequate resources to serve during the peak of demand. In other words, ED overcrowding is the result of a mismatch between ED surge and surge capacity (McCarthy et al. 2006). Surge capacity of an ED includes physical and personnel resources such as beds, doctors, nurses and pharmaceutical equipments.

Generally, in government hospitals, emergency patients are classified using three-color triage zone. Yellow and Red medical triage cards (MTC) are given to patients with semi-critical to critical illnesses, while MTC Green is assigned to non-critical patients. The surge in demand for emergency care associated with continually changing demand case mixes makes the ED an ever-changing system. This situation is tougher with the introduction of new target time for patient to be attended in the ED by Ministry of Health (MOH) in 2009 (refer Table 1). The new target time has put more pressure on ED staff as it must be achieved without additional resources.

Table 1: MOH's New Target Time

Medical Triage Card	Time to be attended	
	Prior	New
Red	immediately	immediately
Yellow	30 minutes	15 minutes
Green	2 hours	90 minutes

An ED with many critical patients usually will be out of control, pushing staff beyond their capacity to provide adequate care. This will prolong and breach the target times for patients who have been assigned with Yellow and Green medical triage card (MTC), as most of the resources are directed to treat patients with Red MTC. The absence of secondary triage also tends to increase patients waiting time because doctors have to examine the patients first before ordering for clinical tests if they were needed. Moreover, outgoing staff responding to ambulance calls exacerbates this situation as ED capacity will be depleted and thus increase the workloads among doctors and nurses to meet the incoming demand.

As ED generates new inpatients, whenever beds are fully occupied, they are placed in treatment area or hallway until an inpatient bed is available. Often times, pre-hospital care is also administered in the treatment area. These scenarios affect patients' flow, tie up space, equipment, and personnel that would otherwise be available to meet the needs of incoming patients. Will extra staff and new building be able to mitigate this turmoil? To make such capacity decision requires an understanding of the system behavior and the flow of the patients through the ED. The ED administrations believe that operation research/management science (OR/MS) approaches could furnish more information before they could commit to a particular course of action.

In the light of the above reasons, there is a need to develop a model representing the real environment of ED to examine and understand resource capacity and as well as to provide a framework that will consult hospital managers do the right decision to improve the service. In other words, a study should be done to determine if the current structure, staffing and physical resources are adequate to cope with a surge in demand. The question to be highlighted is whether ED is able to handle a sudden, unexpected surge in demand by using only the daily operating resources of the hospital. If sudden surges in ED exist, what are the relative effects to the labs and wards? Moreover, how quickly does ED recover from the unexpected surge in demand and how is this reflected in ED census?

This study aims to develop a system dynamics (SD) model that incorporates:

1. the service and flow of patients through the ED
2. the concept of Theory of Constraints (TOC) in capacity limitations
3. decision-making delays in regard to capacity limitations of the ED

This study would also investigate whether process bottlenecks in the system are related to low capacity in the department itself or stemming from factors external

to ED. Though many units in the hospital have direct interaction with the ED, this study will only focus on the causality of various changes proposed in the ED to the labs and wards. In addition, this study examines surge capacity primarily in the context of responses within the hospital's environment or on its grounds that are managed and staffed by the hospital and does not rely on outside assistance.

A REVIEW OF RELATED RESEARCH WORK

Studies in health care capacity planning can be divided into three areas, i.e. bed planning, room planning and staff planning (Jun et al. 1999). However, forming as the main elements in hospital's capacity, these three areas should be studied in a single model to represent the complexity of resource planning in the realm.

A significant amount of research in health care has been conducted in EDs (Jun et al. 1999). This trend is probably attributed to the many problems encountered in the unit. Overcrowding, long waiting time are among the common problems faced by worldwide EDs. Several studies highlight that these problems are related to limited resources, which require proper planning and allocation (Kolb et al. 2007; Gunal and Pidd 2006).

Many researchers have been using OR/MS methods to solve problems in health care domains (Garg et al. 2008). However, simulation outnumbered other OR/MS approaches in health care (Davies and Davies, 1994) for at least three reasons: Firstly, healthcare systems require stochastic approach as there are many uncertainties and variability that are involved in the systems. Secondly, the complex nature of healthcare systems requires a modeling approach that can deal with complexity. Finally, human involvement in healthcare systems needs proper approach for interactions and communications between modeler and user (Brailsford 2007).

There are a few simulation approaches in healthcare such as discrete-event simulation (DES), system dynamics (SD), Monte Carlo simulation and agent based approach (Brailsford 2007). However, the most widely used simulation approach is DES, which has been used to analyze operations in many areas of healthcare like surgery department, clinics and pharmacies and as well as EDs (Spry and Lawley 2005; Gunal and Pidd 2006). This approach focuses more on operational characteristic without portraying the 'big picture' of the system. Unlike DES, SD can offer a better understanding of the 'big picture' by displaying the interactions and causality effects between elements in the system. There are quite a number of literatures on the application of SD modeling in the health care sector. (Koelling and Schwandt 2005; Dangerfield 1999; Royston et al. 1999) reviewed and discussed all previous studies conducted in the area of health care using system dynamics. These studies applied SD approach to solve problems revolved in healthcare such as patients' long waiting time, increasing demand on healthcare and resources allocation, infectious epidemic and community care.

However, thus far, the number of SD models in the context of healthcare planning is rather limited. Among them, the significant models are the works from (Brailsford et al. 2004; Lane et al. 2000 and Wolstenholme 1999). (Brailsford et al. 2004) developed a system dynamics model of the entire health care system in the city of Nottingham, England to simulate patients' flows and identifying the causes of bottlenecks in the health care system. Lane and his team, on the other hand, developed a model to investigate the relationship between waiting times in ED and bed closures. In other words, the study was conducted to see whether reduction in number of bed contributes to the high waiting time in ED. Study by Wolstenholme demonstrated the applications of SD in resource allocation issues. He studied the growing waiting list of elective surgery that was caused by emergency admission and non-elective admission of older people in winter.

Though many models of ED were constructed, those models mainly represent hospitals in the UK and Europe. The different factors and structures applied in these countries are not applicable in Malaysia e.g. technology development and triaging system. Besides, in Malaysia there is a disinclination to acknowledge policy transfer from both the west and neighboring countries, due to federal government control on the national political structure (Common 2004). Therefore, there is a need to find a specific model that can reflect ED in the country.

PROPOSED METHODOLOGY

At present, there is a growing concern on applying manufacturing philosophies to improve the health care system. Despite healthcare differs in many ways from manufacturing, there are also unanticipated commonalities. Whether producing a vehicle or providing healthcare for a patient, staff must rely on multiple, complex processes to accomplish their tasks and provide value to the customer or patient. In this regard, several hospitals in the United Kingdom have been implementing lean principles, six sigma methodology and theory of constraints (TOC) to improve operational performance, reduce operating costs, and improve the quality of service to patients (Young 2005).

Though many simulation studies in manufacturing have tested the combination of DES approach with manufacturing philosophies (Vollman et al. 2005), we could find little in the literature describing the combination of SD and manufacturing philosophies to improve the capacity planning in health care. Therefore, the combination of these approaches could help to understand the dynamic complexity of health system interactions and develop effective solution to improve health capacity planning.

This research is mainly concerned with the response and feedback actions among elements, delays and capacity constraints in the system being study. Feedback actions and delays are well established components of the SD discipline, whereas the concern of this study on capacity constraints is relative to the fact that any system will have at least one constraint as propounded in the TOC (Dettmer, 1998). The concept of managing constraints in TOC may help to identify the root of causes to constrained performance. Nevertheless, TOC may complement SD not only in generating an understanding of the problem, but also in discovering ways of solving them (Davies et al. 2004). TOC encourages further actions to be taken using the thinking process tools in identifying hurdles to implementation and conceive proper plans to get over the hurdles (Dettmer 1998). TOC thinking process consists of five logical tools that can be used individually or can be integrated. There are *Current Reality Tree (CRT), Conflict Resolution Diagram (CRD), Future Reality Tree (FRT), Prerequisite Tree(PT)* and *Transition Tree (TT).* For elaboration on these tools, readers are referred to other TOC text such as (Dettmer 1998) and (Lepore and Cohen 1999).

There are several stages usually involved in approaching a problem thinking using SD. Although different authors introduce different steps or phases in SD modeling process, the goal of each step is generally the same. In this study the modeling process suggested by Sterman (2000) and Maani & Cavana (2000) is modified and will be applied. The modeling process can be divided into six stages:

1. Problem Definition (Boundary Selection)
2. Formulation of Dynamic Hypothesis
3. Formulation of a Simulation Model
4. Model Testing
5. Policy Design and Evaluation
6. Strategy Implementation

In order to mix SD and TOC, this study is based on Mingers' framework for characterizing the philosophical assumptions underlying SD approach (refer Mingers 2003) and the framework by Davies et al. (2004) for characterizing the philosophical assumptions underlying TOC methods. By mapping both frameworks, we identify appropriate TOC thinking process to complement steps in conducting SD modeling as shown in Figure 1.

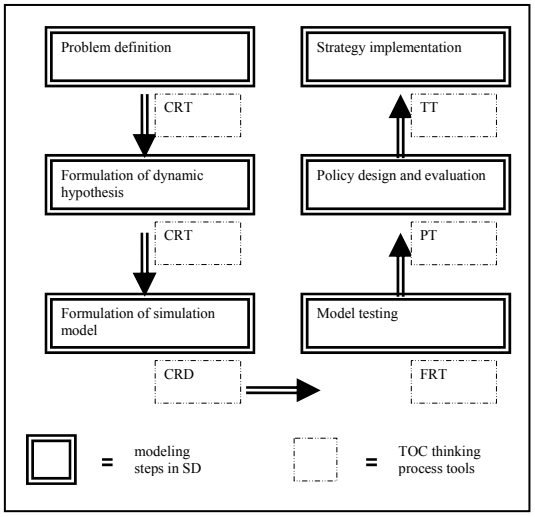


Figure 1: Steps in Conducting a System Dynamics Study and Applicable TOC Thinking Tools

At the initial phase, interviews with the ED administration have helped to identify the main actors of a potential model and the boundaries to the problem. The high level map of factors affecting ED services is summarized in Figure 2.

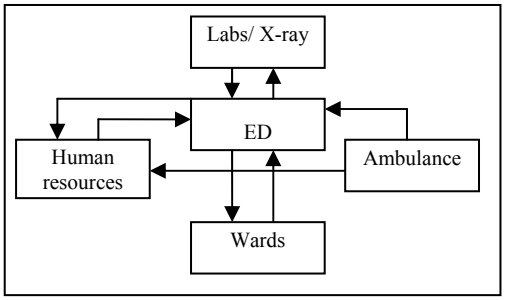


Figure 2: High-level map of factors affecting ED services

At this phase, the CRT will be used to list symptoms that indicate what is inappropriate within the system. During second phase, dynamic hypothesis will be developed. Causal loop diagrams will be constructed to reveal the dynamic hypothesis of problem understudy. The dynamic hypothesis is obtained from the patterns of behavior that arise from the relationship among elements in the system. In minor contrast, the CRT will also be used to capture the systemic nature of relationships within the system being modeled using

undesirable effects to trace back the root of the problem. The different protocol offers by CRT will serve as complementary tool to causal loop diagrams in identifying the cause and effect chains of the problem.

On the next phase, with the assist of the CRD, a simulation model will be able to display any conflicting views among stakeholders and find ways to resolve the conflict. It can also show the real level at which conflict usually occurs. In model testing, the FRT will be applied to visually unfold the cause and effect relationship between changes we make to the existing systems and their resulting outcomes. During policy design and evaluation phase, the PT will be used to identify obstacles preventing achievement of a desired course of action. Finally, the TT will be used to support strategy implementation by providing a step-by-step method for implementation.

CONCLUSION AND FUTURE RESEARCH

This study is currently in a preliminary stage and thus much more work is needed before constructing the computer simulation model. Causal loop diagrams for each sector in Figure 2 must be developed to become the basis for model development. Next, model assumptions underlying the development of the model must be detailed for each sector. Finally, once the initial model is built up, experimentation and analysis will be performed. Analysis from the model is hoped to help determining the effective intervention for ED capacity planning and capture the concept of system improvement as espoused in TOC.

REFERENCES

Baesler, F. F., H. E. Jahnsen and M. DaCosta. 2003. "The use of simulation and design of experiments for estimating maximum capacity in an emergency room." In *Proceedings of the 2003 Winter Simulation Conference*, USA. 1903-1906.

Ballard, S. M. and M. E. Kuhl. 2006. "The use of simulation to determine maximum capacity in the surgical suite operating room." In *Proceedings of the 2006 Winter Simulation Conference*, USA. 433-438.

Barnes, D. 2008. *Operations management: An international perspective*. Thomson Learning: London.

Brailsford, S. C. 2007. "Tutorial: Advances and challenges in healthcare simulation modeling." In *Proceedings of the 2007 Winter Simulation Conference*, at USA. 1436-1448.

Brailsford, S. C., V. A. Lattimer, P. Tarnaras, and J. C. Turnbull. 2004. "Emergency and On-Demand Health Care: Modelling a Large Complex System." *Journal of the Operational Research Society* 55 (1):34-42.

Common, R. 2004. "Public management and policy transfer in South-east Asia." In *Policy transfer in global perspective*, edited by M. Evans. Hants, England: Ashgate Publishing Limited.

Dangerfield, B. C. 1999. "System Dynamics Applications to European Health Care Issues." *The Journal of the Operational Research Society* 50 (4):345-353.

Davies, J, V. J. Mabin, and J. F. Cox. 2004. "The Theory of constraints and system dynamics: A suitable case for multi-methodology." In *Proceedings of the International Conference of the System Dynamics Society*, Oxford, England.

Davies, R, and H. T. O. Davies. 1994. "Modelling patient flows and resource provision in health systems." *Omega: The International Journal of Management Science* 22:123-131.

DeLia, D. 2005. "Emergency department utilization and surge capacity in New Jersey, 1998-2003." A report to the New Jersey Department of Health and Senior Services.

Dettmer, H. W. 1998. *Breaking the constraints to world-class performance*. ASQ Quality Press, Milwaukee, Wisconsin.

Garg, L., S. McClean, and M. Barton. 2008. "Is management science doing enough to improve healthcare?" In *Proceedings of World Academy of Science, Engineering and Technology* 30: 76-80

Gunal, M. M., and M. Pidd. 2006. "Understanding accident and emergency department performance using simulation." In *Proceedings of the 2006 Winter Simulation Conference*, USA. 446-452.

Health Informatics Centre, Planning and Development Division. 2009. *Health Facts 2008*: Ministry of Health, Malaysia.

Hong, Ng Cheow and Noraida Abdul Ghani. 2006. "A Model for Predicting Average Ambulance Service Travel Times in Penang Island." In *Proceedings of the 2nd IMT-GT Regional Conference on Mathematics, Statistics and Applications*, Universiti Sains Malaysia.

Jun, J. B., S. H. Jacobson and J. R. Swisher. 1999. " Application of discrete event simulation in health care clinics: A survey." *Journal of the Operational Research Society* 50, 109-123.

Koelling, P. and M. J. Schwandt. 2005. "Health systems: A dynamic system- benefits from System Dynamics." *In Proceedings of The 2005 Winter Simulation Conference*, USA. 1321-1327.

Kolb, Erik M.W., T. Lee, and J. Peck. 2007. "Effect of coupling between emergency department and inpatient unit on the overcrowding in emergency department." In *Proceedings of the 2007 Winter Simulation Conference*, USA, 1586-1593.

Lane, D. C., C. Monefeldt, and J. V. Rosenhead. 2000. "Looking in the Wrong Place for Healthcare improvements: A System Dynamics Study of an Accident and Emergency Department." *Journal of the Operational Research Society* 51 (5):581-531.

Lepore, D. and O. Cohen. 1999. *Deming and Goldratt: The Theory of Constraints and the system of profound knowledge*. The North River Press, USA.

Maani, K. E. and R. Y. Cavana. 2000. *System thinking and modeling: Understanding change and complexity.* Pearson Education New Zealand Limited, New Zealand.

McCarthy, M. L., D. Aronsky, and G. D. Kelen. 2006. 'The measurement of daily surge and its relevance to disaster preparedness.." *Academic Emergency Medicine* 13:1138-1141.

Mingers, J. 2003. "A classification of the philosophical assumptions of management science methods."

Journal of the Operational Research Society 54:559-570.
Mohamed, Mafauzy. 2003. "Medical Schools: The supply and availability of qualified human resources-challenges and opportunities." *Malaysian Journal of Medical Sciences* 10 (1):1-3.
News Strait Times. 2008. (23/12/2008). Simply short of doctors at A&E. p.1.
Pillay, S. 2009. "Inside the emergency room?" *News Strait Times*, 24/03/2009, 8.
Royston, G, A. Dost, J. Townshend, and H. Turner. 1999. "Using system dynamics to help develop and implement policies and programmes in health care in England." *System Dynamics Review* 15:293-313.
SERI, Socio-Economic and Evironmental Research Institute. 2009. *Penang Statistics*. A Document submitted to Penang State.
Spry, C. W. and M. A. Lawley. 2005. "Evaluating hospital pharmacy staffing and work scheduling using simulation." In Proceedings of the 2005 Winter Simulation Conference, USA. 2256-2263.
Sterman, J. 2000. Business dynamics: System thinking and modeling for a complex world. The McGraw Hill, Boston, USA.
Vollman, T. E., W.L. Berry and D.C. Whybark. 2005. *Manufacturing Planning and Control for Supply Chain Management.* 5 ed. The McGraw-Hill: Boston.
Wolstenholme, Eric. 1999. "A patient flow perspective of U.K. Health Services: Exploring the case for new "intermediate care" initiatives." *System Dynamics Review* 15 (3):253-271.
Young, Terry. 2005. "An agenda for healthcare and information simulation." *Health Care Management Science* 8:189-196.

AUTHOR BIOGRAPHIES

NORAZURA AHMAD is a Ph.D. candidate in the Department of Mathematics in School of Distance Education in Universiti Sains Malaysia. She received a B.Sc (Statistics) from Universiti Putra Malaysia , her M.Sc (Decision Sciences) from Universiti Utara Malaysia (UUM). She is currently a lecturer in College of Arts and Sciences, UUM. She can be contacted by email address norazura@uum.edu.my

NORAIDA ABDUL GHANI is an Associate Professor in Mathematics at Universiti Sains Malaysia . She received her B.A (Mathematics/Statistics) from California State University , USA, her M. Sc (Statistics) from San Diego State University, USA and her D.Sc (Operations Research) from The George Washington University, USA. Her areas of interests include the location/allocation problems and stochastic modeling. She can be reached at noraida@usm.my.

ANTON ABDULBASAH KAMIL is an Associate Professor in Mathematics at University Sains Malaysia . He received his B.Sc (Statistics) from Padjadjaran University, Bandung, Indonesia, his M. Sc (Statistics) from Bogor Agricultural University, Indonesia and Ph. D (Econometric) from University of Economics, Prague, Czech Republic. He specializes in Econometrics and Financial Mathematics. And has published numerous publications in journals and proceedings. He can be reached at anton@usm.my.

RAZMAN MAT TAHAR is a professor at Faculty of Technology Management in Universiti Malaysia Pahang. He received his B.Sc. (Mathematics/Statistics) from Carleton University, Canada, his M.Sc. (Computer Based Modeling/Simulation) from Sunderland University and his Ph.D. (Computer Simulation) from University of Bradford. His areas of research include production/ manufacturing modeling, logistics/ supply chain modeling, discrete simulation and system dynamics. His email address is razman779@ump.edu.my

A PROTOTYPE SIMULATOR OF POLICE OPERATIONS IN CRISIS SITUATIONS

Andrzej Urban and Mariusz Nepelski and Grzegorz Gudzbeler

Higher Police School in Szczytno

12-100 Szczytno POLAND

tailer@wspol.edu.pl

Abstract

The paper presents modern perception of crisis management and its importance for security in the context of non-military threats. The paper shows the role of the police as a part of a governmental system, performing their tasks in all stages of crisis management. The article is about simulator crisis situation. The Minister of Knowledge and Higher Education gives money for a project for 2009 – 2011, for consortium Higher Police School and ETC – PZL Aerospace Industries Ltd.

Introduction

Necessity reasons of conducting a research – a problem situation

To guarantee the security is regarded as realization of a specific mission fulfilled by the state's executives, autonomies, community and citizens[1]. This mission refers to the two basic areas related to the security. The first contains the elements connected with protection and defense of 'values and national interests against the existing and potential threats', the second is related to the creation of 'internal and external conditions for free development and copying with challenges, which are carried for a nation by inconstancy, unpredictable conditions and civilization progresses[2]. The importance of education for security was noticed by the European Union. One of the aims of the United Europe for the years 2007-2013 is the development of technology and knowledge due to the creation of necessary abilities to ensure the security of citizens in the scope of such threats as terrorism, disasters and delinquency, simultaneously respecting the basic human rights including the right to privacy; assurance of optimum use of the available technologies in favor of the European civil security, stimulation of cooperation between providers and users of solutions in the scope of civil security, improvement of competition in the European security sector and the delivery of research results directed to the realization of mission due to the reduction of existing shortages in the field of security. Undoubtedly the education in this range notably inscribes into the realization of the conducted research. It is not possible to provide security without proper educated personnel in that way[3]. Thus the state's interest is to guarantee security of its citizens and one of the subjects responsible for this issue is the Police, which is equipped with proper forces and means to eliminate the threats. Undoubtedly a necessary element of a successful action is an adequately prepared command – personnel and gaining the knowledge and abilities in commanding itself should take place during a professional training and physical participation in police operations thus in the whole decision process. The commanding issue itself has a complicated structure, presented in the further part of this study.

Commanding[4] police operations in the crisis situations[5] is based on three organizational forms of

[1] More to this topic: M. Nepelski, A. Tyburska, *Ochrona infrastruktury krytycznej*, [*The protection of critical infrastructure]*, Szczytno 2008, p. 8-9. Compare: R. Jakubczak (ed.), *Obrona narodowa w tworzeniu bezpieczeństwa III RP. Podręcznik dla studentek i studentów*, [*The national defense in creation of security in the III Republic of Poland. A handbook for students]*, Warsaw 2003, p. 61-62; R. Jakubczak, J. Flis (ed.), *Bezpieczeństwo narodowe Polski w XXI wieku. Wyzwania i strategie. [Poland's national security in the XXI century. Challenges and strategies]*, Warsaw 2006, p. 21; J. Marczak, R. Jakubczak, K. Gąsiorek, *Obrona terytorialna w obronie powszechnej RP, [The territorial defense in the Poland's common defense]*, Warsaw 2003, p. 23-45; W. Kitler, Z. Piątek (ed.), *Realizacja zadań bezpieczeństwa przez samorząd terytorialny, [Realization of the security tasks by the local autonomy]*, Sandomierz 2006.

[2] *R. Jakubczak, J. Flis (ed.), Bezpieczeństwo narodowe Polski w XXI wieku. Wyzwania i strategie, [Poland's national security in the XXI century. Challenges and strategies]*, Warsaw 2006, p. 22

[3] A. Urban, *Dytaktyczne wyzwania dla bezpieczeństwa w XXI wieku. Wyzwania bezpieczeństwa cywilnego XXI wieku – inżynieria działań w obszarach nauki, dydaktyki i praktyki, [Didactic challenges for security in the XXI century. Challenges of civil security in the XXI century – operations' engineering in the field of science, didactic and practice],* Warsaw 2007, p.202, 203. More to this topic: Decision No. 1982/2006/We of the European Parliament I Cabinet from the 18th December 2006, concerning the seventh framework programme of the European Commonwealth in the field of researches, technological development and demonstration (2007 0 2013).

[4] §2 passage 1 Decree No. 213 of the Police Commander-in-Chief from the 28th February 2007 in the

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

police operations, that are interventions, actions and operations, pointing at the process within the decisions are taken, that is commanding. The presented elements of the taking decision process are closely attributed to the people taking decisions, thus commanders of interventions, actions and police operations. From the point of importance of the threats and the degree of complication of the police actions, the police operations are that kind of police actions, which are formed in case of an event dangerous for life and health of the people or their property, caused by unlawful assaults violating these goods or a natural disaster, characterized by the possibility of losing control over the events or escalation of a danger, in which it is necessary for security and public order protection to use a greater number of police officers organized in units or squads, including armed teams. The following catalogue of crisis situations is defined in police depiction:

- mass events of increased risk,
- assemblies and public ceremonies of high risk,
- blockades of the roads and buildings' occupations,
- organized pursuit operations,
- terror acts,
- collective violation of security and public order, in particular for the social, economical, political and religious reasons,
- natural disasters, which results may lead to social riots,
- other situations which may cause a danger to peoples' life and health or property and also to security and public order, characterized by the possibility of losing control over the events or escalation of a danger, where to counterattack or to eliminate them, it is necessary for security and public order protection, to use a greater number of police officers organized in units or police squads, including armed teams.

In connection with such serious events the police operations are ordered. Firstly, the ordering person will be the Police commander-in-chief, when the range of an event is bigger than the area of one Police provincial headquarter or there is more likely that such an event is possible to happen or also during operations taking place on the actual site of a Police provincial headquarter, which extend in time, and in case of a need to support the forces and means including the logistics going beyond the potential of an actual Police provincial headquarter. Secondly, the Police commander-in-chief is managing a police operation, when the range of an event is bigger than the area of one Police district headquarter or there is more likely that such an event is possible to happen or also during operations taking place on the actual site of a Police district headquarter, which extend in time, and under condition of an increasing need to support the forces and means including the logistics beyond the potential of an actual Police district headquarter. A police operation may be ordered by the district/municipal/regional Police commander, when the range of an event is bigger than the area of one subordinated Police unit or there is more likely that such a threat is possible to happen or during operations extending in time, under conditions of an increasing need to support the forces and means not being at his or her disposal including the logistics beyond the potential of an actual Police district headquarter. The sub operations connected with different events can be lead within the confines of one operation, if they are simultaneously aimed at the realization of the main operation's target. Within the confines of sub operations the division into such sections is allowed, which are aimed at the realization of the sub operation's target and into subsections, which are aimed at the realization of the section's target. The Police commander-in-chief, the provincial Police commander and the district Police commander or their deputies and also a Police officer appointed by the one of the commanders, with a proper training, having predispositions to and experience in commanding are entitled to command an operation. In case of a division of an operation into sub operations, a police officer with predispositions and experience in commanding is being entitled to command a sub operation. The operation is realized according to a commander's operation plan. Such plan is being prepared by the chief of staff of the operation's commander and it is being approved by the operation's commander. Such operations are called planned operations. There are also operations realized in case of a sudden event, which makes it impossible to prepare a plan. The realization of such operations relies on starting an operation on the basis of an oral decision of the actual provincial or district Police commander and the immediate written confirmation as well as leading the operation according to the commander's operation plan, adapted to the nature of the event. The realization of a police operation ensues on the basis of three levels: strategic, tactical and executive. The first is attributed to the police operation commander, the second to the headquarter of the police operation commander, which is responsible for working out the commander's decision, that is supporting the commanding process, and the third to executors of the operation, inter alia to the sub operation commanders.

The constructed prototype simulator of police operations in crisis situations will be based on the presented way of operating of the police forces. The people functioning in the process of commanding the police operations will undergo a training, that are the police operation commanders, the police sub operation commanders and the people working at the headquarter of the police operation commander. The training, reflecting the stages of the commanding process, will consist of three parts: preparation, realization and reconstruction stage.

matter of methods and forms of preparation and realization of Police tasks in case of a threat to people's life, health, their property or public order and security.

[5] Ibidem, §2 passage 6.

The functioning essence of the prototype will be a simulation of processes not only connected with taking decisions (tactical and strategic) but also reactions of the virtual environment to an arising situation in the police and non-police context. The construction of the detailed behavior algorithms (decision trees) and their implementation in the system on the basis of the existing legislative solutions as well as experts experience in the field of Police operations in crisis situations and also methodology of teaching, connected with the existing technical solutions at the market as well as creation of the new more functional ones, will enable the complete functioning of the simulator.

The description of the innovative solutions of the prototype simulator of Police operations in crisis situations

Operations associated with the implementation of the project will, above all, relate to thematic areas concerning software engineering, knowledge and decision support, designing of specialized systems, computational sciences and creating scientific databases. The designed simulator will be a specialized system using the resources of interdisciplinary character. The created solution will realize all tasks in the field of simulating and presenting incidents in virtual environment. The modelled objects will be built according to the real objects' parameters. Each object will have predefined features concerning the appearance and the possibility of interaction with other objects. Such environment will enable simulating of most of the incidents occurring in the natural environment with the possibility of observing their consequences. Furthermore, modern solutions will also be applied in the field of artificial intelligence and its quality and efficiency will be defined in the real time simulation of objects' reactions of different characteristics. Additionally, the solutions concerning the presentation of the incidents will allow to analyse and define the principles of implementation of information acquisition using such solutions as CCTV and correlation between the quality of such solutions and the quality of decision processes in crisis management.

The simulation models used in the system will be created using statistical data gathered by the Polish Police as well as the experts' knowledge in the fields of police sciences who will prepare analysis for each of the realised stages. The gathered information resources will be used in the teaching process of implemented module of artificial intelligence. The sets of standard scenarios and role-playing which are being prepared, will be supervised by an instructor – a teacher conducting the training. The instructor will additionally have the possibility of modifying the scenario or generating new incidents. It will give a possibility of creating different decisional situations basing on one scenario. In addition, using the possibility of modification of the whole environment the leader will be able to make changes in the weather conditions. The scenarios can be joined and role-played in the real time, which allows to fully calculate the level of difficulty of the task implementation. The people being trained will have all the possible information sources and appropriate resources (forces and means) necessary during the preliminary activities in the training. Decisions taken by the trained persons will be based on more or less complete information according to the scenario and dispositions. The system will be able to simulate municipal CCTV network, industrial CCTV network, mobile CCTV centres and other information sources. In the given situations the trained person who is making the decision, thanks to the system, will be able to go to the scene of the incident in the virtual environment to command directly from the scene of the incident. The possibility of using information sources will depend on many different factors (weather conditions, simulated breakdowns, catastrophes) modified by the instructor conducting the training.

Many participants will be able to take part in the training at the same time and the number will only be limited by the number of operating stations designed in the system. The participants of the training will realise the tasks on two levels of commanding – strategic and tactical. According to the level the participants will receive appropriate information from the system and they will also manage the means and forces appropriate for their position in the structure. The solution will be fully flexible – changes in the structure, organisation of commanding and responsibilities of services will not affect the functioning of the system, and simulations will be instantly adapted to the new requirements. The structure of the conducted training will determine specific roles for all of the participants. The organizer of the training will have appropriate personnel resources, such as instructors responsible for the whole training process, supported by the administrator of the system and its operators, who will play the part of the persons fulfilling dispositions during the practice session. Communication will be executed using the appropriate means of communication. The exchange of information between the trained (cooperating forces) will be executed using advanced means of digital (trunking) communication system TETRA. The number of operators in the training will depend on the number of trained persons and the level of difficulty of the implemented scenario. Such solution will cause that the people trained will not have to get familiar with the applications not present in the unit they represent. The trained persons will get the knowledge strictly concerning the training programme in the field of crisis management and will develop the ability of its implementation.

The system will allow training in the situations very similar to the real ones and also provide an instant assessment of the ways of managing by reactions of the virtual environment on the decisions taken. It will cause that the trained persons will feel the pressure of the environment and observe the results of their accurate and wrong decisions. The level of the training

conducted using such methods will be very high and will include all the aspects of operations in real situations.

Creating of the simulator prototype will allow to illustrate the operations implemented by the police in terms of crisis events from the point of view of the police, which strictly relates to Order 213/2007 of Chief Commander of Polish Police from 28 February 2007 concerning methods and forms of preparation and implementation of police tasks and public order, which describes the commanding system in the situations mentioned above. During the training the simulator will also allow to check the knowledge and skills which are based on and crisis reaction procedures, dispositions concerning coordinating and cooperating activities scattered in different legal acts.

Students of the Higher Police School in Szczytno as well as training participants will be able to use the police operations simulator in crisis situations, within the framework of the curriculum. Training of the decision making processes will be supported by multimedia which will allow identification of errors and gaps in the plans of the police operational commanders and crisis reaction plans. They will enable education of managing (especially commanding) forces and means by the police and other forces and institutions of paramilitary character and also promote activities' self-control and immediate reaction to mistakes made in the process and also how to work in a team. Properly prepared didactic decision game on the tactical and strategic level perfectly simulates operations in real conditions and allows to develop skills of creative solution seeking and making quick decisions, comprehension and joining cause and effect relations, and also stress resistance.

Improvement of the managing (commanding) process in crisis situations by the police will be executed by implementation of decision games on the tactical-strategic level supported by multimedia decision training. The aim of the training will be improvement of police officers' practical skills (commanders of police operations and their staff) in the field of executing their responsibilities as commanders, and above all, organizing of cooperation and coordination of the activities of all functional subjects (teams) of staff, in order to calculate the data and different variants of activities necessary for the commander to make a decision. The essence of the decision training will be practicing of particular staff activities according to the valid procedures concerning selected crisis situations in terms of police activities on the level of commander of the police operation, in order to rationally execute the tasks of their staff, to achieve desired level of captured knowledge demanded by the particular tactical or strategic situation. The substance of the decision training will be gathering, studying and analyzing of data (information) by the commander of the police operation, making operational calculations (tactical and strategic), varianting of different ways of using force and means, developing planning documents, executing operational tasks in the virtual terrain in the field of organizing and realization of commanding, activities synchronization and improvement of skills concerning the use of technical means of commanding and communication. Furthermore, on the basis of the created system, there is a possibility of creating innovating organizational and technical solutions with the object of educating forces other than the police.

Conclusions

The final result of the scientific research will be a police operations simulator in crisis situations functionally directed at improving of managing (commanding) process in crisis situations by the police, especially by implementation of decision games on the tactical-strategic level, supported by multimedia decision training. The created system will be designed for simulating of physical phenomena in artificially created environment. The proposed solution will be an computer system which executes tasks both in the field of simulating and presenting incidents in virtual environment. The modelled objects in the simulator will be built using the parameters of the real objects. Each object will have its own predefined features concerning its appearance and possibility of interaction with other objects. Such environment will allow to simulate most of real incidents with the possibility of observing the consequences of their occurrence. All adjacent objects will react in the way determined by algorithms in each particular incident occurring in the system. The modelled objects will include trees, buildings, vehicles and people. The system will be able to simulate behaviour of both services on the scene of incident and incidental witnesses, as well as injured persons. The picture generated by the system will be based on the most recent solutions and will utilize the possibilities offered by modern graphic cards. This will cause that all of the presented objects and situations seen by the users of the system will be very accurate and can be recognized as the view from the real picture source, as in case of industrial or municipal CCTV. The simulator will be an ideal place to execute incidents' simulations, which can be specially difficult in real environment, due to organizational reasons.

In the future the designed technical and organizational solution can be used by other services responsible for public security. The action plan includes preparing models concerning characteristics of most of the services responsible for public security in Poland. It means, that implementation of this solution in the simulation of activities of other services will be much less time consuming. In practice, it comes to ready-made modifications or developing new scenarios. Additionally, the solution will be fully modifiable and scalable, which will make it easier to accommodate for the further users.

TIMER EMBEDDED FINITE STATE MACHINE MODELING AND ITS APPLICATION

Duckwoong Lee, Byoung K. Choi and Joohoe Kong
Department of Industrial and Systems Engineering
KAIST
335 Gwahak-ro, Yuseong-gu, Daejeon, 305-701, Republic of Korea
E-mail: ldw721@kaist.ac.kr

KEYWORDS

Timer Embedded, Finite State Machine, Synchronization Manager, TEFSM Toolkit

ABSTRACT

In this paper, we propose an extension of classical finite state machine as it called a timer embedded finite state machine (TEFSM) with its formal modeling methods. In the proposed state-based approach, a discrete event system is modeled as a coupled TEFSM. Also presented is a systematic procedure and architecture of developing a simulation executor with a synchronization manager for the coupled TEFSM model. A TEFSM toolkit for modeling and simulation of the proposed TEFSM model has been implemented and a ping pong system was developed as an illustrative example.

INTRODUCTION

A **finite state machine (FSM)** is the oldest known formal model for modeling the sequential behavior of a discrete event system (Wagner 2004). FSM is also called *finite state automata*, finite state transducer, state machine, etc. A ***FSM*** is defined as a model of computations consisting of a set of *states*, a *start state*, an *input alphabet*, and a *transition function* that maps inputs and current states to a *next state*. Computation begins in the start state with an input string and changes to new states depending on the transition function (Paul 2009).

There exist various *formal definitions* of FSM. In a *classical definition* (Peterson 1981), a FSM is defined as a structure (S, X, Y, δ, λ). In *computer science* where the term "finite state automata (FSA)" is mostly used for FSM (Hopcroft 2006), a FSM is defined as a structure (S, X, δ, s_0, F). Also, in *engineering*, a FSM generating outputs is referred to as a *finite state transducer* which is a structure (S, X, Y, δ, s_0, λ), where:

- S is a finite set of states
- X is a finite set of symbols (input alphabet)
- Y is a finite set of symbols (output alphabet)
- δ is the state transition function
- λ is the output function
- s_0 is the start state
- F is the set of final states

Outputs of FSM are generated by *actions*. There are three types of actions associated with a state: (1) ***entry actions*** performed when entering the state, (2) ***exit actions*** performed when exiting the state, and (3) ***input actions*** performed depending on the present state and input conditions. And, an action associated with a transition is referred to as ***transition action*** (Wagner 2003).

Figure 1 shows the execution flow of FSM (Wagner 1992). It waits for an input at the current state, and when the input is received the *input action condition* is tested. If the condition is met, the input action is executed and transition condition is checked. If the transition condition is met, the FSM exits from the current state after executing the *exit action*, moves to the next state while executing the *transition action*, and enters into the next state while executing the *entry action*.

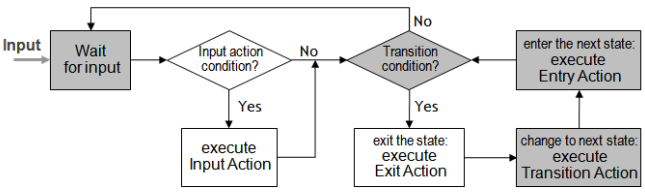


Figure 1: Execution Flow of FSM with Actions

In this paper, we propose an extension of the FSM which a ***timer*** is embedded as it called the **timer embedded finite state machine (TEFSM)**. The proposed TEFSM is executed with only shading flow in Figure 1 due to the timer. This paper presents a formal modeling definition of the proposed TEFSM and a systematic procedure and architecture of developing a simulation executor with an

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

illustrative example.

TIMER EMBEDDED FSM: TEFSM

We propose a timer embedded FSM (TEFSM) to be used for state-based modeling of discrete event systems. A **timer** means delay time (i.e., timeout value) of a state and then, the state transition occurs automatically after timeout unless it receives further inputs before timeout. An *automatic transition* of states is referred to as an **internal transition**. The proposed TEFSM is *deterministic* (i.e., it has to be in one and only one state at a time), and it has *transition actions* as well as *entry actions*. Also incorporated are *condition* and *probability* associated with transitions, making it a *probabilistic* FSM. In addition, it is allowed to have ***state variables*** which are updated by actions and are used in defining conditions. In summary, the TEFSM has the following extended features:

(1) *Timers* (time-delays) and *entry actions* (E-Action) associated with states.
(2) *Conditions* (Boolean or *probability*) and *actions* (T-Action) associated with transitions.
(3) *Internal transitions* enabled by timers/conditions without external inputs.
(4) *Time* is implicitly associated with inputs and outputs.
(5) *State variables* are used to reduce the state space.

Specifying a FSM as an *algebraic structure* with a detailed description of the transition function δ is both tedious and hard to read, thus there are two preferred notations for describing FSM (Hopcroft 2006): *FSM diagram* and *state transition table*. Therefore, a **TEFSM diagram** and/or a **state transition table (STT)** which is a tabular listing of the transition function are used in the proposed TEFSM model. The conventions for constructing the proposed **TEFSM diagram** are summarized in Table 1. A number of symbols are used in order to increase readability: '?' symbol for input; '!' for output; '%' for probability; '~' for condition and probability, and 'Δ' for time delay.

In the proposed state-based approach, a discrete event system is modeled as a ***coupled*** TEFSM model consisting of a number of ***atomic*** TEFSM models at modeling phase and the modeling methods with an illustrative example are described in next Section. Also, a *TEFSM executor* may easily be written from the STT of the TEFSM at execution phase and the details of execution of TEFSM are presented in Section 4.

Table 1: Conventions for Constructing a TEFSM Diagram

Primitives	Notations
External Transition Edge	Condition %(probability) ?(Input) T-Action
Internal Transition Edge	Condition %(probability) T-Action
State Node	Initial State; State: Δ(t), E-Action; Final State

T-Action= {!(output), State variable updates}

SYSTEM MODELING EXAMPLE WITH TEFSM

In order to understand our proposed TEFSM easily, a ping pong game, as a well acquainted example, is modeled. Two players, Player-A and Player-B, and an Umpire are involved the game. The rules of the ping pong game with an expedite system (ITTF 2001) are presented in Appendix-A and the reference model is depicted in Figure 2. The expedite system is applied in which an umpire sends an 'Expedite' output if the game is unfinished after 10 minutes, while a game is finished in which any player who wins the game sends an 'Over' output. During a rally, the player sends 'Ball' or 'Out' output.

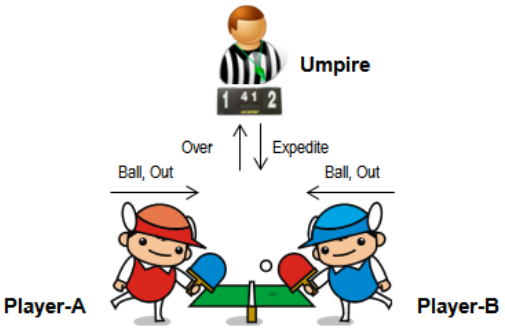


Figure 2: Reference Model of Ping Pong Game

In accordance with conventions in Table 1, we build each atomic TEFSM diagram for two players and an umpire (i.e., atomic TEFSM diagrams for both Player-A and Player-B are identically same except for message name of "Ball" and "Out"). Then put the three diagrams together, a coupled TEFSM model of the ping pong game may be obtained as shown in Figure 3. Since the TEFSM model is self-explanatory, additional explanations are omitted. The internal transition conditions associated with the states and functions are defined in Appendix-B, where t_o = offense delay time, t_w

= wait delay time, t_e = expedited delay time (in Player), t_e = expedited time (in Umpire), P_{IN} = probability of ball-in (success), and P_{OUT} = probability of out (failure).

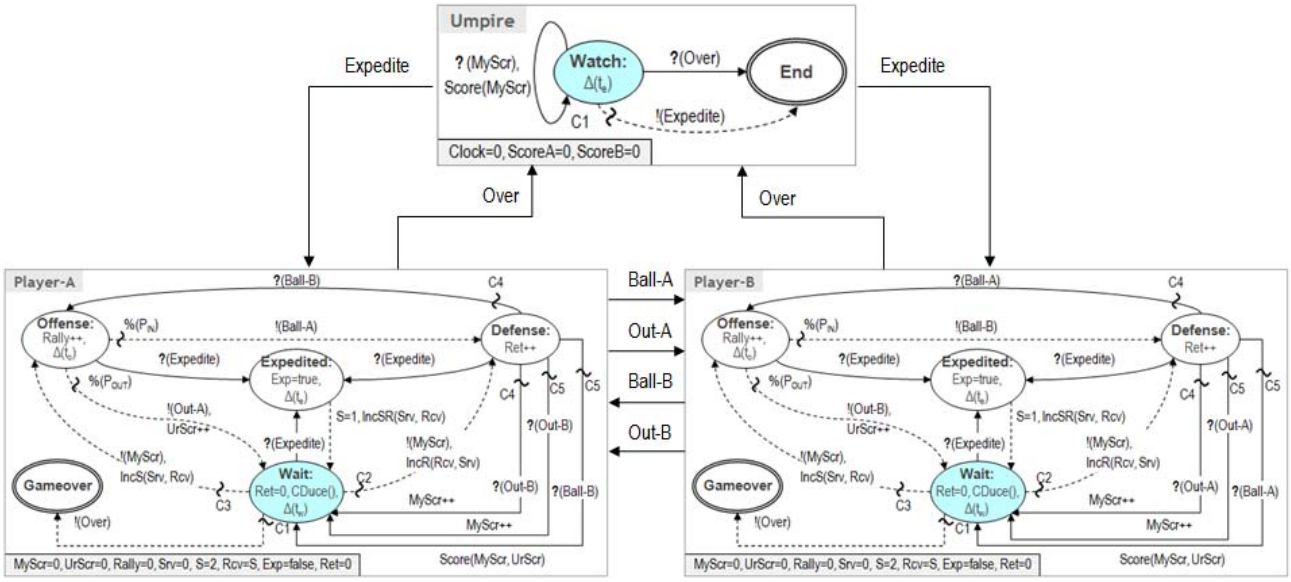


Figure 3: TEFSM Coupled Model of Ping Pong Game

Table 2 is a STT of an atomic TEFSM model for the Player-A. Both TEFSM diagram and their corresponding STT are contained identical model of TEFSM, and therefore the STT of Umpire is omitted.

Table 2: State Transition Table of Player-A

State	Entry Action	Input/Delay	Condition	Transition Action	Next State
Wait	Ret=0, CDuce()	$\Delta(t_w)$	C1	!(Over)	Gameover
			C2	IncR(Rcv, Srv)	Defense
			C3	IncS(Srv, Rcv)	Offense
		?(Expedite)			Expedited
Offense	Rally++	$\Delta(t_o)$	$\%(P_{IN})$	!(Ball-A)	Defense
			$\%(P_{OUT})$	!(Out-A), UrScr++	Wait
		?(Expedite)			Expedited
Defense	Ret++	?(Ball-B)	C4		Offense
			C5	Score(MyScr, UrScr)	Wait
		?(Out-B)	C4	MyScr++	Wait
			C5	MyScr++	Wait
Expedited	Exp=true	$\Delta(t_e)$		S=1, IncSR(Srv, Rcv)	Wait

EXECUTION OF COUPLED TEFSM MODEL

This section presents methods of constructing a TEFSM simulator, which is also a coupled TEFSM, and of building a TEFSM executor of the coupled TEFSM model.

The key issue in the simulation of a coupled TEFSM model is how to synchronize simulation times of individual atomic TEFSM models. The *time-synchronization* (Fujimoto 2000) method we use in this paper is based on the concept of the ***synchronization manager*** (Lee 2010). The overall structure of the TEFSM simulator of Figure 3 is shown in Figure 4.

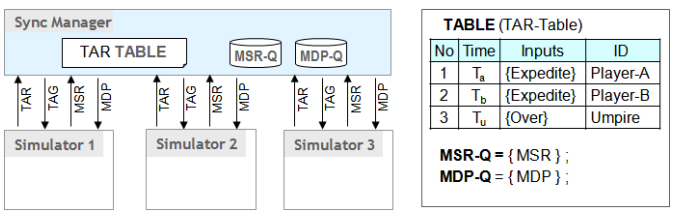


Figure 4: Overall Structure of TEFSM Simulator for the Coupled TEFSM Model in Figure 3

In the TEFSM simulator, all the interactions among the

atomic TEFSM models are made through ***Sync Manager***. At the beginning, each atomic model sends a TAR (*time advance request*) message to Sync Manager at its *start state*. Then, Sync Manager builds a *TAR table* named TABLE, and sends a TAG (*time advance grant*) message to an atomic model whose request-time is smaller than those of others. An instance of the TAR table is depicted in Figure 4. Upon receiving the TAG, the atomic model advances its simulation time and moves into a new state. During and/or after this *state transition*, the atomic model may send MSR (*message send request*) messages to Sync Manager who will store the received MSR messages in a queue named MSR-Q. Since a MSR in which an "output" of the atomic model is contained may be an "input" to multiple atomic models, the messages to be sent out are temporally stored in another queue named MDP-Q where MDP stands for *massage delivery packet*. The data objects introduced so far have the following structures:

- **TAR = (Time, Inputs, ID)** // time advance request (ID= atomic model ID)
- **TAG = (Now, ID)** // time advance grant (Now= current simulation time)
- **MSR = (Msg, ID)** // message send request (Msg= *input/output* message)
- **MDP = (Msg, Now, ID)** // message delivery packet
- **MSR-Q = {MSR}** // simple list of MSR
- **MDP-Q = {MDP}** // simple list of MDP

Shown in Figure 5 are details of Sync Manager which is also a TEFSM. The synchronization manager presented in Figure 5 is a general one that can be used for any coupled TEFSM model. The functions Select(TAG) is for making a TAG with an atomic model whose TAR is smaller than those of others and Get-MDP(m) is for getting MDPs from MSR-Q in order to deliver them to corresponding atomic models. Also, the model is self-explanatory, the additional explanations are omitted.

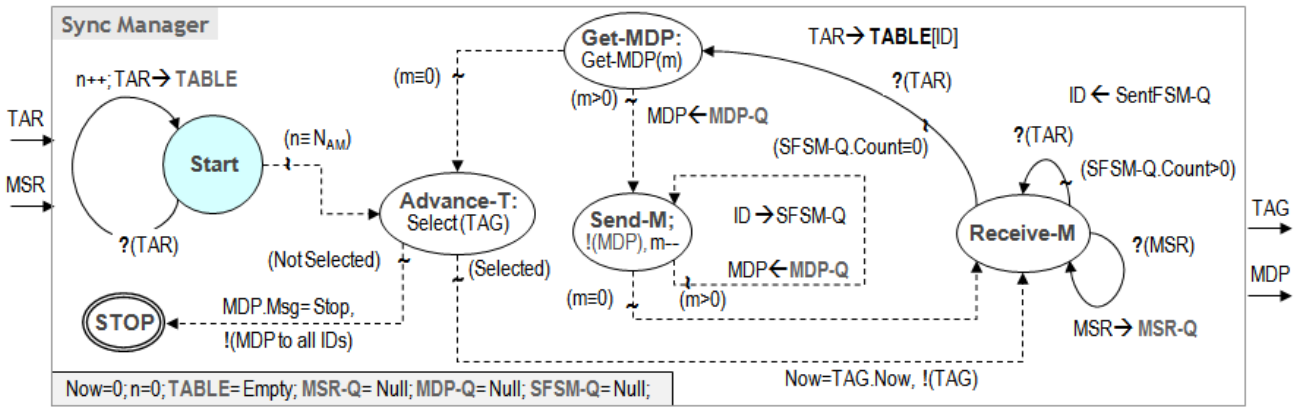


Figure 5: TEFSM Model of the Synchronization Manager in Figure 4

An "atomic model" of a *coupled TEFSM model* (Figure 3) is different from that in a *TEFSM simulator* (Figure 4). The former may be called an *atomic system model* and the latter an *atomic simulation model* or ***atomic simulator*** in short. Therefore, an atomic simulator obtained by converting atomic TEFSM model at execution phase without internal transitions (i.e., this conversion is performed automatically in our TEFSM toolkit which is presented in next Section and makes the TEFSM model into the classical FSM model). In general, the ***rules for converting*** an atomic TEFSM model having final states to an atomic simulator are:

- If the atomic model has *final states*, a STOP state is added as a *final state*.
- Each of the original *final state* is converted to a *regular state*, and a *transition edge* having "?(MDP[Stop])" is defined from each *converted state* to state STOP.
- For all *states* except state STOP, an *entry action* "Clock= Now" is added.
- For a *timed state* with time-out value of t_o, (1) an entry action "!(TAR[t_o, I])" is added where I denotes a set of inputs of the state and (2) each *internal transition* edge is replaced by an *external transition* edge having "?(TAG)".
- For a *state without timer*, an entry action "!(TAR[∞, I])" is added.
- Each output "!(Out)" is replaced by "!(MSR[Out])" and input "?(In)" by "?(MDP[In])."

IMPLEMENTATION AND APPLICATION

Modeler and simulator for the proposed TEFSM model

have been implemented as a TEFSM toolkit under a Microsoft .NET Framework 3.5 environment using the C# programming language. The software architecture of TEFSM toolkit is shown in Figure 6.

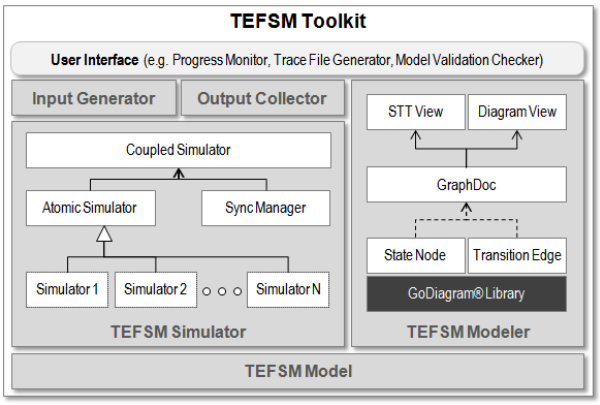


Figure 6: Software Architecture of TEFSM Toolkit

The toolkit consists of a user interface, a TEFSM modeler that provides a diagram view or a STT view of TEFSM model, a TEFSM simulator that converts the TEFSM model into TEFSM simulator and generates C# codes for simulation, and TEFSM model that the fundamental model objects are contained. The installation and tutorial files of the toolkit may be downloaded from http://vms.kaist.ac.kr.

Figure 7 shows a GUI for modeler in the toolkit including the atomic TEFSM model STT (top of Figure 7: same to Table 2) and diagram (bottom of Figure 7: same to one of Figure 3) for Player-A. In this toolkit, the basic C# codes for simulation are generated from the modeler so that the developer could implement the simulator easily.

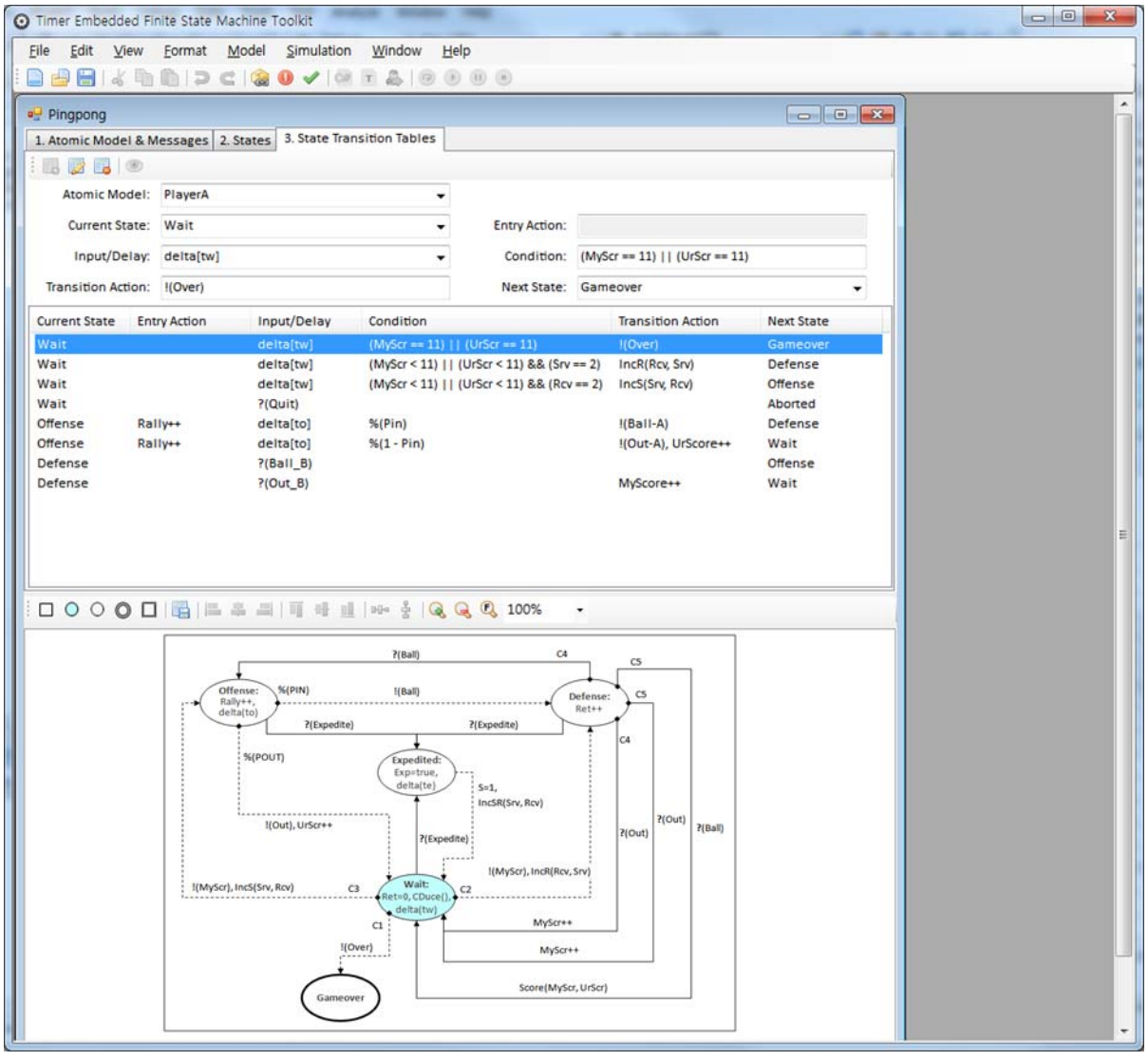


Figure 7: GUI of Modeler in TEFSM Toolkit

In order to animate progress of the simulation after or while the simulation is run, the toolkit generates the trace file of the simulation result for Proof® Animation (See http://www.wolverinesoftware.com). Figure 8 shows a screen capture of the ping pong game animation of which a trace file generated from the TEFSM toolkit.

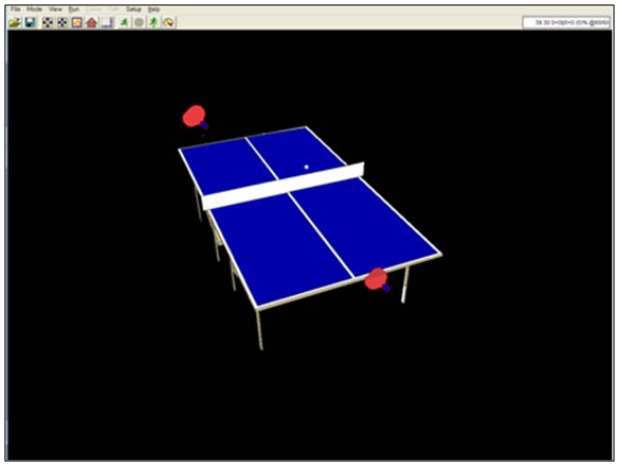


Figure 8: Proof® Animation with TEFSM Trace File

CONCLUSION

Presented in this paper is an extension of FSM, timer embedded FSM (TEFSM). In the proposed TEFSM, a discrete event system is modeled as a coupled TEFSM and well simulated as a simulation executor with synchronization manager. Also presented is a TEFSM toolkit for modeling and simulation with a ping pong game as an illustrative example. Further developments of TEFSM toolkit might be needed in order to test rigorously and undergo a refinement of a real-life environment.

Comparing to the class *discrete event system specification* (DEVS) which is a well known state-based modeling formalism (Zeigler 2000), the proposed TEFSM has generality and modeling power as follows: (1) the output function λ is defined as a *transition action* of an *external transition*, while the output of DEVS is a transition action of an *internal transition*, (2) state variables that are used in defining conditions for transitions are allowed, and (3) supports the entry actions while DEVS does not.

APPENDIX

A. Modeling assumptions (considering rules) of a ping pong game

- Consider only one match with two player and one umpire.
- A game is over without duce if a player scores 11 points (each player servers twice before changing its turn).
- If both players reach 10 points, then service alternates after each point, until one player gains a two point lead.
- Expedite system is as follows:
  - Except where both players have scored at least 9 points, the expedite system shall come into operation if a game is unfinished after 10 minutes' play or at any earlier time at the request of both players.
  - If the ball is in play when the time limit is reached, play shall be interrupted by the umpire and shall resume with service by the player who served in the rally that was interrupted.
  - Thereafter, each player shall serve for 1 point in turn until the end of the game and if the receiving player 13 returns the receiver shall score a point.
  - Once introduced, the expedite system shall remain in operation until the end of the match.

B. Internal transition conditions and details of functions used in Figure 3

```
C1 = ((MyScr ≡ 11) || (UrScr ≡ 11)) & (abs(MyScr, UrScr) ≡ 2) //Game-over
C2 = ((MyScr < 11) & (UrScr < 11) || (abs(MyScr, UrScr) < 2)) & (Srv ≡ S) //Receive
C3 = ((MyScr < 11) & (UrScr < 11) || (abs(MyScr, UrScr) < 2)) & (Rcv ≡ S) //Serve
C4 = (Exp ≡ false) || ((Exp ≡ true) & (Ret <13)) //Normal
C5 = (Exp ≡ true) & (Ret ≡ 13) //Expedited & score

IncS(Srv, Rcv): Srv += 1; if(Srv ≡ S) then Rcv = 0
IncR(Rcv, Srv): Rcv += 1 ; if(Rcv ≡ S) then Srv = 0
IncSR(Srv, Rcv): if(Srv ≤ S) then Srv = 0, Rcv = S
                 else then Srv = S, Rcv = 0
Score(MyScr, UrScr): if(Srv ≡ S) then MyScr++;
                     else then UrScr++;
CDuce(): if(MyScr ≡ 9 & UrScr ≡ 9) then S = 1
```

REFERENCES

Fujimoto, R.M., 2000, *Parallel and Distributed Simulation Systems*, John Wiley & Sons

Hopcroft, J.E. *et al.*, 2006, *Introduction to Automata Theory, Languages, and Computation*, 3rd Ed., Addison Wesley

ITTF (International Table Tennis Federation), 2001, http://www.allabouttabletennis.com/basic-table-tennis-rule.html

Lee, D., Shin, H., and Choi, B.K., 2010, Mediator Approach to Direct Workflow Simulation, *Simulation Modeling Practice and Theory*, Volume 18, Issue 5, pp. 650-662

Peterson, J.L., 1981, *Petri Net Theory and the Modeling of Systems*, Prentice-Hall

Paul E. Black, 2009, http://www.itl.nist.gov/div897/sqg/dads/HTML/finiteStateMachine.html

Wagner, F. and Wolstenholme, P., 1992, VFSM Executable Specification, *IEEE CompEuro 1992 Proceedings*

Wagner, F. and Wolstenholme, P., 2003, Modeling and Building Reliable, Re-usable Software, *IEEE ECBS'03 Proceedings*

Wagner, F. and Wolstenholme, P., 2004, Misunderstandings about State Machines, *Computing and Control Eng.* Volume 15, Issue 4, pp.40–45

Zeigler, B., Praehofer, H. and Kim, T., 2000, *Theory of Modeling and Simulation*, Academic Press: Boston

AUTHOR BIOGRAPHICS

Duckwoong Lee is a post doctor in the Department of Industrial and Systems Engineering at KAIST. He received a BS from Ajou University in 2002, a MS from KAIST in 2004, and a Ph.D. from KAIST in 2010, all in Industrial Engineering. His research interests are in the area of business process management system (BPMS), system modeling and simulation, and parallel and distributed simulation.

Byoung K. Choi is a professor of the Department of Industrial and Systems Engineering at KAIST since 1983. He received a BS from Seoul National University in 1973, a MS from KAIST in 1975, and a Ph.D. from Purdue University in 1982, all in Industrial Engineering. His current research interests are system modeling and simulation, BPMS, simulation-based scheduling, and virtual manufacturing. He can be reached at bkchoi@kaist.ac.kr

Joohoe Kong is a graduate student in the Department of Industrial and Systems Engineering at KAIST. She received a BS from Arizona State University in 2005, a MS from KAIST in 2007 in Industrial Engineering. Her research interests are in the area of BPMS, systems modeling and simulation. She can be reached at joohoe@vmslab.kaist.ac.kr

ADOPTION OF SIMULATION TECHNIQUES FOR MASTERING LOGISTIC COMPLEXITY OF MAJOR CONSTRUCTION AND ENGINEERING PROJECTS

Dr.-Ing. Katja Klingebiel
Chair of Factory Organisation
Technische Universität Dortmund
Leonhard-Euler-Str. 5
44227 Dortmund, Germany
katja.klingebiel@tu-dortmund.de

Yuriy Gavrylenko
Fraunhofer Institute
for Material Flow and Logistics
Department of
Supply Chain Engineering
Joseph-von-Fraunhofer Straße 2-4
44227 Dortmund, Germany
yuriy.gavrylenko@iml.fraunhofer.de

Dr.-Ing. Axel Wagenitz
Fraunhofer Institute
for Material Flow and Logistics
Department of
Supply Chain Engineering
Joseph-von-Fraunhofer Straße 2-4
44227 Dortmund, Germany
axel.wagenitz@iml.fraunhofer.de

KEYWORDS

Logistics, construction projects, public events, supply chain management, supply chain simulation.

ABSTRACT

Major construction and engineering projects pose high challenges at logistic planning, including numerous restrictions, risks and dynamic effects. Consequently many tasks cannot be effectively solved using conventional tools and methods. We propose an integrated planning procedure that adopts simulation techniques in order to consider the whole supply chain and incorporate operative processes as well as stochastic factors. These integrated planning principles have been applied successfully in the project Sochi 2014 and will be outlined here.

INTRODUCTION

Large public projects like construction of facilities for major sport events (e.g. Olympic Games, Soccer World Championship) are often characterized by specific sets of logistic problems that result from the complexity of the projects: often more than hundred constructions sites have to be managed simultaneously and enormous amounts of materials have to be supplied while infrastructure restrictions and physical locations (central location in the host-cities and regions) have to be taken into account. Furthermore, these projects are characterized by a heterogeneous partner landscape and dynamical and partly non-transparent interdependencies between project management and logistics.

However, planning activities in major construction and engineering projects focus primarily on project management and operative construction site logistics. These approaches often lack the necessary interfaces between project management and logistic layer. Additionally, there are a number of factors that make difficult the trade-off between logistic and planning layer: extensive planning tasks and specific challenges require new approaches and tools for solving the problems of strategic design and mid-term planning of the major projects; thereby considering dynamic restrictions and risks as well as later implementation and operative processes.

Though simulation is a proven approach for assessing logistics concepts, the subproject "simulation" is often badly integrated in the overall logistics planning. Up to the present time applications of simulation techniques in construction logistics have been limited to material flow problems of the on-site logistics. Typically conventional material flow simulation tools are applied here. Thus all information flow is neglected.

In contrast we propose an integrated logistics planning procedure, where we apply simulation techniques. This allows the holistic design of dynamic logistics networks including material and information flow processes.

In this paper we will first specify the logistic challenges in major construction and engineering projects in detail. Against this background we will outline our integrated approach. An application of this technique for the major construction and engineering project is consequently introduced in this paper.

LOGISTICS CHALLENGES IN MAJOR CONSTRUCTION AND ENGINEERING PROJECTS

Major public projects like Olympic Games, European Championship or World Exhibitions take place in well-established host cities or other well-known locations. This strategy poses the advantage of exploiting already existing infrastructures: both during the event itself (public transport, gastronomy, hotels etc.) and during construction period (existing storages, roads, transshipment facilities etc.). However, there is a downside to this strategy: most host cities and regions have grown historically and have not been planned to handle such amounts of additional logistics activity. Hence, extensive planning is needed to integrate new logistics concepts into existing and evolving infrastructure. Here a number of challenges arise.

Off-site logistic restrictions

Regarding off-site transportation capacities (inbound transportation channels), four main types of transportation modes can be distinguished: railway, ship

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

(sea, inland water transports or combination), air cargo and road.
Road transport can be used for delivering construction materials and goods directly to material consumption points, thus combining off-site and on-site transports. All other transport modes require transshipment facilities in order to forward materials to consumption points.
Railway transports pose a number of restrictions for logistics operations: track capacities, train and railcar capacities, rail yard operational capacities and transshipment capacities.
Air-cargo transports are typically restricted by operational and transshipment capacities of the airport. Furthermore, due to the high costs of air-cargo operations, only a small part of material supplies can be carried by air.
Sea and inland water transports are cost-effective and capable of transporting high volumes of materials. However, ship transports are restricted by ship capacities, port yard capacities and port transshipment capacities. Both ship transports and railways have a lot of restrictions regarding infrastructure and bindings to particular supply chains in contrast to road transports that generally show high grade of supply chain flexibility and infrastructure availability.

Local logistic restrictions

The infrastructure of host cities is generally not suited for handling additional amounts of transports and thus has to be adapted or expanded in order to provide adequate logistic capacities. Furthermore, flawless operability of the city's public services during the construction period has to be assured in order to minimize negative effects for citizens.
Some logistic concepts target these specific restrictions, including strategies like strict separation of on-site and off-site logistics, outsourcing of off-site activities and storages from the host-city to its outskirts, providing high performance transportation channels (e.g. railways) from off-site areas to main construction sites (e.g. London 2012).

Risk and uncertainties

Franke points to the following structure of risks for industrial plant engineering projects that are also valid for public construction projects (Franke 1987, pages 32-33):

- Risks resulting from quantities and efficiencies (processes, engineering, procurement, erection, tests and inspections),
- Risks resulting from dependencies (customer, suppliers, etc.),
- Scheduling risks,
- External influences (authorities, politics, market situation, etc.),
- Uncertainties as to payment and concerning liabilities (delay, penalty, etc.),
- Warranty risks.

Furthermore, there are two external influences that are especially important for major public projects: social-acceptability and sovereign risks. Social-acceptability risks refer to the likelihood that sponsors will meet opposition from local groups, economic-development agencies, and influential pressure groups. Sovereign risks in turn involve the likelihood that a government will decide to renegotiate contracts, concessions, or property rights (Miller 2001, page 439).
Other external, payment, liability and warranty risks are mostly out of sphere of the influence of logistics. Yet risks resulting from quantities, efficiencies, dependencies and scheduling risks are pivotal for logistics planning.

Ramp-up of facilities

It is obvious that many infrastructure facilities are not available for logistics operations from the first day of the project. They have to be built up or extended either before project's start or during construction time. Due to the extensive construction volumes and mostly tight time schedules it is essential to carry out sophisticated planning, taking into account overlaps between schedules of infrastructure projects and main construction objects. In other words, construction plans must match the ramp-up of transportation, transshipment and storage capacities.
Furthermore, especially in the beginning of the project's lifespan material flow focuses on bulk goods (crushed stones, gravel, cement etc.) changing to general cargo with overall project's advancement. Accordingly transshipment and storage capacities must be adjusted.

Operative restrictions and conditions

Strategic design and planning for major event and construction projects depend strongly on the later operative logistic concept. Processes included in the operative layer influence strategic design and mid-term planning and vice versa. However, these interdependencies are complex. A center for logistics coordination is crucial. Its main task is to integrate strategic, tactical and operative planning processes under consideration of scarce infrastructure capacities.
So, this covers a wide range of operational tasks, including organizational issues like fleet accreditation, supporting workflows or even coordination of centralized supplies (e.g. gravel, sand or concrete). Nevertheless strategic logistic design and planning need to focus on those operative processes which can be influenced to a great extend:

- order consolidation rules and lead times for inbound transport, differentiated both by material type and transport mode;
- dispatching rules for local transport;
- transport restrictions;
- material storage rules (which material groups are centrally stored and which are delivered directly to construction sites);
- storage call-off rules;

Waste disposal and environmental concerns

Major projects environmental issues come to the fore due to the strong public attention. Consequently the logistics design has to be analyzed in-depth to prevent or minimize negative implications on environment and populations. In context of construction or engineering projects waste and pollution are the most important environmental issues to be taken into account.

Sochi region is located in the very sensitive environmental area and many construction activities take place near nature protection areas. Therefore it is essential to evaluate waste flows and development of the waste disposal yards in order to assess influence of the construction activities n the environment.

AN INTEGRATED PLANNING AND EVALUATION APPROACH FOR LOGISTIC CONCEPTS

Logistics design and planning addresses the configuration of the logistics structure as well as the planning and allocation of processes and resources (Arnold 2008, page 3). It has to determine the cost-optimal and service demand satisfying structure and design of logistics networks (Kuhn and Hellingrath 2002, page 88). The objective is to identify, specify and validate recommended logistics structures including an optimal process and resource allocation. Typically all three aspects are strongly interconnected. Processes must be designed to comply with resources and network structure, resource concepts must be supported by reliable network structures.

Nevertheless effective logistics design requires the evaluation and validation of logistics performance and efficiency. Yet explication and interpretation of system behavior is difficult for complex logistics systems (Kuhn et al. 2010, page 1), especially as given in context of major construction and engineering projects. Logistics scenarios have become too complex to be handled without tools of advanced information technology providing transparency. The conclusion on the question which methodology to apply for assessing logistics concepts has to be drawn against the background that strategic network design is a task with high impact on subsequent planning and execution phases (Seidel et al. 2005, page 55). Considering the size and complexity of construction and engineering projects, mathematical models have been assessed as rather unsuitable to master all named challenges. And evaluation cannot stop at a highly aggregated level, but has to consider detailed logistics characteristics.

Nevertheless the possibility to rapidly model and evaluate, large scale networks represents a strong argument for the application of static analysis. Yet, in order to allow for thorough assessment in a dynamic environment the application of simulation techniques provides valuable capabilities. Therefore we strongly support the application of both methodologies to be prepared for the challenges at hand in major construction and engineering projects (see also Klingebiel and Seidel 2007):

In an early phase of logistics planning static analysis provides the possibility to rapidly design a large number of alternative models and to evaluate these by help of low granularity key performance indicators (KPI). The number of feasible and economically efficient models is being reduced drastically and a small group of alternatives remains. In the next step these alternatives can be assessed in detail by help of simulation techniques which then provides high granularity KPIs. Thus, the combination of both methodologies provides the potential for consistent and thorough assessment of logistics concepts.

With this integrated planning approach, simulation has gained significance besides analytical and optimizing methods. It is applied to validate pre-selected planning, control and material flow processes as well as structure and resource oriented concepts.

Nevertheless, simulation as a method needs to be strongly integrated in the well-known procedure of design and planning of logistics systems (Kuhn and Hellingrath 2002; Scheer 2002; Klingebiel 2009). This procedure starts with identification of strategic project objectives and related measurable key performance indicator and continues with a detailed analysis of current state processes and structures before identifying fields of actions and developing and evaluating to-be scenarios (see Figure 1). We note that three essential conditions need to be met (Kuhn et al. 2010):

- Assignment of simulation experts with know-how in the field of logistics planning
- Full integration of these simulation experts into the planning team
- Avoidance of budget distribution between planning and simulation in order to decide about the degree of detail in simulation without bias.

The general method for proceeding within the simulation step is separated into the phases of conceptual design, data preparation, modeling and verification, scenario evaluation, analysis and documentation (Klingebiel and Seidel 2006). This proven methodological approach is based on German industrial guidelines for simulation (VDI 2000; Rabe et al. 2008; Wenzel et al. 2007) and depicted in the Figure 1. In parallel to these steps the models, scenarios and results need to be continuously validated. This initiates back loops to previous steps to refine the model, model concept, or even the scope and objectives.

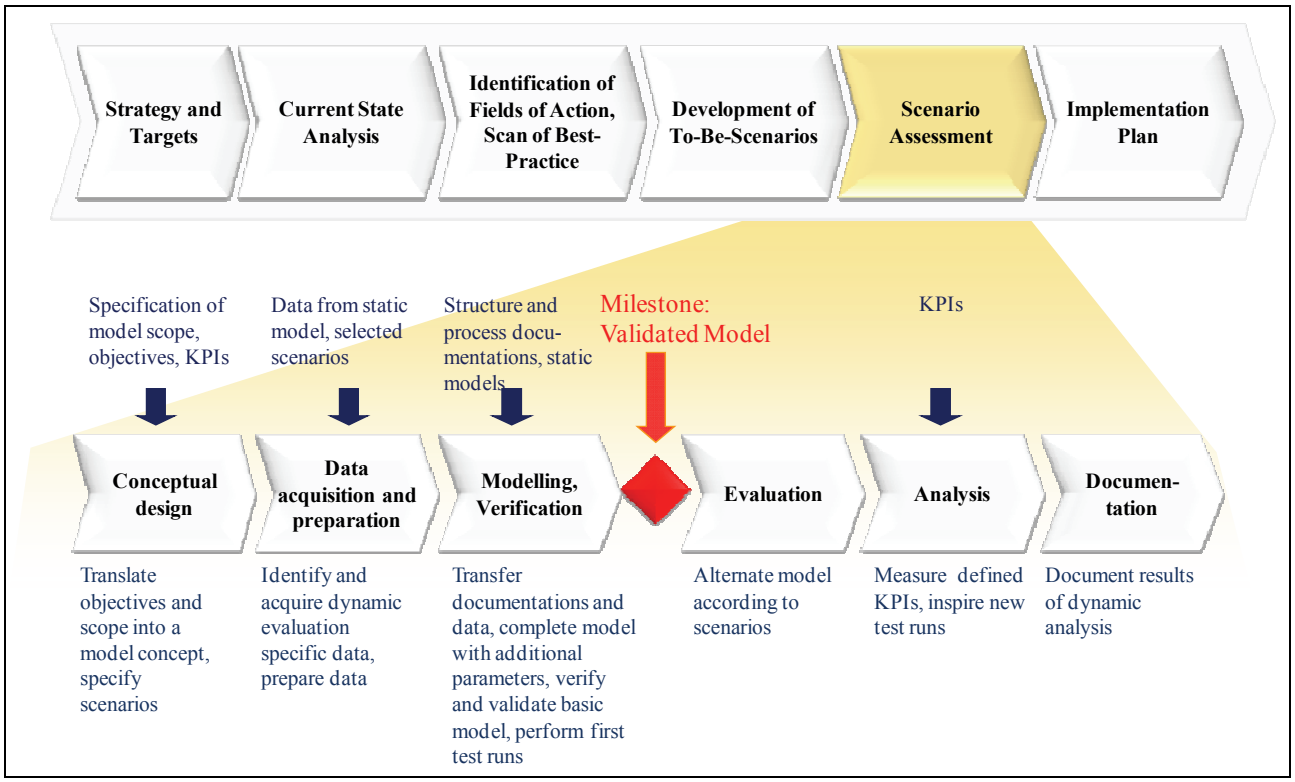


Figure 1: Methodological approach for simulation (Klingebiel 2006)

Conceptual design

During this phase a general overview of the evaluation model is developed which is in most parts related to the to-be model, but might vary in details - depending on the specifics of the applied evaluation tool.

Data acquisition and preparation

Usually evaluation of concept feasibility requires more complex model data; hence the to-be model must be completed with additional data. So the next step comprises the identification of this data as well as the acquisition and preparation of it for application.

Modeling and verification

In the step 'modeling and verification' an evaluation model is built. The model provides the basis for the analysis of specified alternatives and scenarios of the to-be model. Thus it has to first be verified against semantic and syntactic failures. This comprises a check of whether the to-be model concept has been transferred correctly into the evaluation model concept and the model environment, i.e. whether the functionalities of the evaluation methods and supporting tool environment were applied as defined by the model concept.
In addition to this formal verification the validation of input data, evaluation model and results is an important step. Validation comprises the examination of the correspondence between the model, results and reality. By iterative validation we need to assure that the behavior of the model replicates reality in a sufficient way.

Because the model abstracts and idealizes reality, usually not every behavior can be studied in the model itself. Hence continuous validation requires the specification of expected results and associated accuracy beforehand. The model is verified against this tolerance framework. In many cases the evaluation models replicate stochastic effects parameterized by distributions. The characteristics of a distribution are replicated best when many random samples are executed. Consequently it is often necessary to execute a high number of test runs within the validation phase to gain reliable insight into the model's behavior. The outcomes of these test runs have to be discussed with industry experts.

Computation and simulation

In the next phase alternative scenarios are defined. The evaluation model is altered as specified and outcomes computed or simulated. Often the outcomes of one run cause new questions and thus scenarios. Thus evaluation is often executed in systematic trials where previous results inspire new test runs.

Analysis

The results of the test runs are measured by the previously defined KPIs. Thus the resultant data, which often comprises of basic measures, is aggregated to key performance indicators. Most evaluation methods and tool environments provide interfaces to external analysis tools that allow for specific and individual processing of result data.

Documentation

In the documentation phase the evaluation results are summarized and prepared for the interpretation and discussion. As this documentation constitutes the basic information source for the later decision which concepts to apply, in reality a clean and accurate analysis of results is necessary here.

Consequently, this integrated logistics planning approach and the underlying, clear procedure model give us the guideline for the logistics planning process within the context of major construction and engineering projects. Within several projects the application of this basic procedure has proven its efficiency. In the following we outline the benefits of this integrated approach as well as the valuable insights given by simulation-based evaluation for the project Sochi 2014.

THE APPLICATION OF A SIMULATION-BASED LOGISTICS DESIGN PROCESS FOR CONSTRUCTION LOGISTICS

The city of Sochi holds the Olympic Winter Games 2014. Sochi is located in the remote area of Krasnodar Krai (federal subject of Russian Federation) and is badly connected to the main road network. Though Great Sochi area has a length of 145 km, many of the 180 Olympic construction objects are distributed in the densely populated city of Sochi, in its proximity or in the mountain area Krasnaya Pollyanna. The existing logistics infrastructure is scarce and even not satisfactory for the current city's needs, especially in summer tourists additionally strain the infrastructure.

Consequently the project focus was to support the development of a logistics strategy in form of an integrated package of activities aimed to ensure construction of transport infrastructure, terminals and storage facilities from 2008 till 2012. As corresponding logistics activities would heavily affect local population a sophisticated transport strategy during construction had highest priority. Subtasks of our work have been the review and analysis of the given infrastructure under estimated goods movement within Sochi region from 2008 till 2012, the identification of existing risks and possible bottlenecks as well as the development of a concept for the freight logistics control.

Based on our experience in other major construction projects, specifications obtained from our local partners and visit in Sochi, we carried out the detailed analysis of given structures and processes so that major fields of action could be identified. In a next step the scenarios for inbound logistics have been developed. Based on geographical and logistics data a network model was designed that allowed the identification of optimal distribution channels and site locations.

It was one of the most important aspects within this project to evaluate the impact on the regional infrastructure. Consequently, logistics processes and information flows have been specified by help of an holistic simulation model of Sochi's logistics system for the construction of the Olympic facilities.

During all project steps the simulation experts shadowed the logistics planner and have been integrated in the planning team. With this concept the benefits of adopting simulation techniques could be fully exploited. It was possible to choose the right level of abstraction according to the beforehand named challenges and expectations.

We applied the simulation environment OTD-NET which has been developed by Fraunhofer Institute of Material Flow and Logistics in Dortmund, Germany (Wagenitz 2007). OTD-NET introduces a holistic approach for modelling and simulation of complex production and logistics networks. It delivers in-depth insights into information and material flows, stock levels, stability of the network, boundary conditions and restrictions.

Heart of OTD-NET is the discrete event simulation of business processes. The developed hybrid approach visualizes processes by aid of UML (Unified Modeling Language) and implements these in an object-oriented programming language (C# and partly C++). The analysis module provides reporting functionalities by an adequate processing of simulation data into multidimensional data structures including condensed data for online analytical processing (OLAP). Thus OTD-NET supports the diversity as well as the complexity of factors inherent in this type of project. However, specific challenges that have been described above have influenced the individual project steps immensely. In the following paragraphs a detailed insight into the simulation model is given.

Model Conception

Material inbound channels serve as source of material deliveries for the about 180 Olympic construction projects within Sochi region. We incorporated 33 basic types of goods and materials into the model (e.g. crushed stone, building steel or electrical equipment). The sources, i.e. inbound channels, can be distinguished into railway freight yards, sea ports and truck delivery channels. The model concept incorporates the essential information for inbound channels: connection to the transportation network of Sochi region, throughput and transshipment capacities. Especially the last two have been characterized as scenario data as they are subject to planning.

It was identified that for the aspired evaluations it was possible to cluster the sinks, i.e. the 180 Olympic construction sites, into 26 sub-clusters. Each sub-cluster incorporates several construction objects (stadiums, hotels etc.) as planned in the Olympic construction program which may be consolidated due to geographical aspects and logistic needs. Furthermore, we introduced waste yards as additional material consumption points to mirror the ecological challenges at hand.

Besides sinks and sources, distribution channels are the fundamental objects of the model concept. They are described by a number of parameters: transportation times, transport constraints, scheduling behavior. Within the transport network we also designed central storage points which are capable of buffering materials. This is necessary as material demand and material supply may not always be optimally synchronized due to the infrastructure and transport restrictions named before.

Figure 2 illustrates the resulting model concept showing construction site sub-clusters (yellow), distribution channels (arrows), inbound channels (green) and storage capacities (orange).

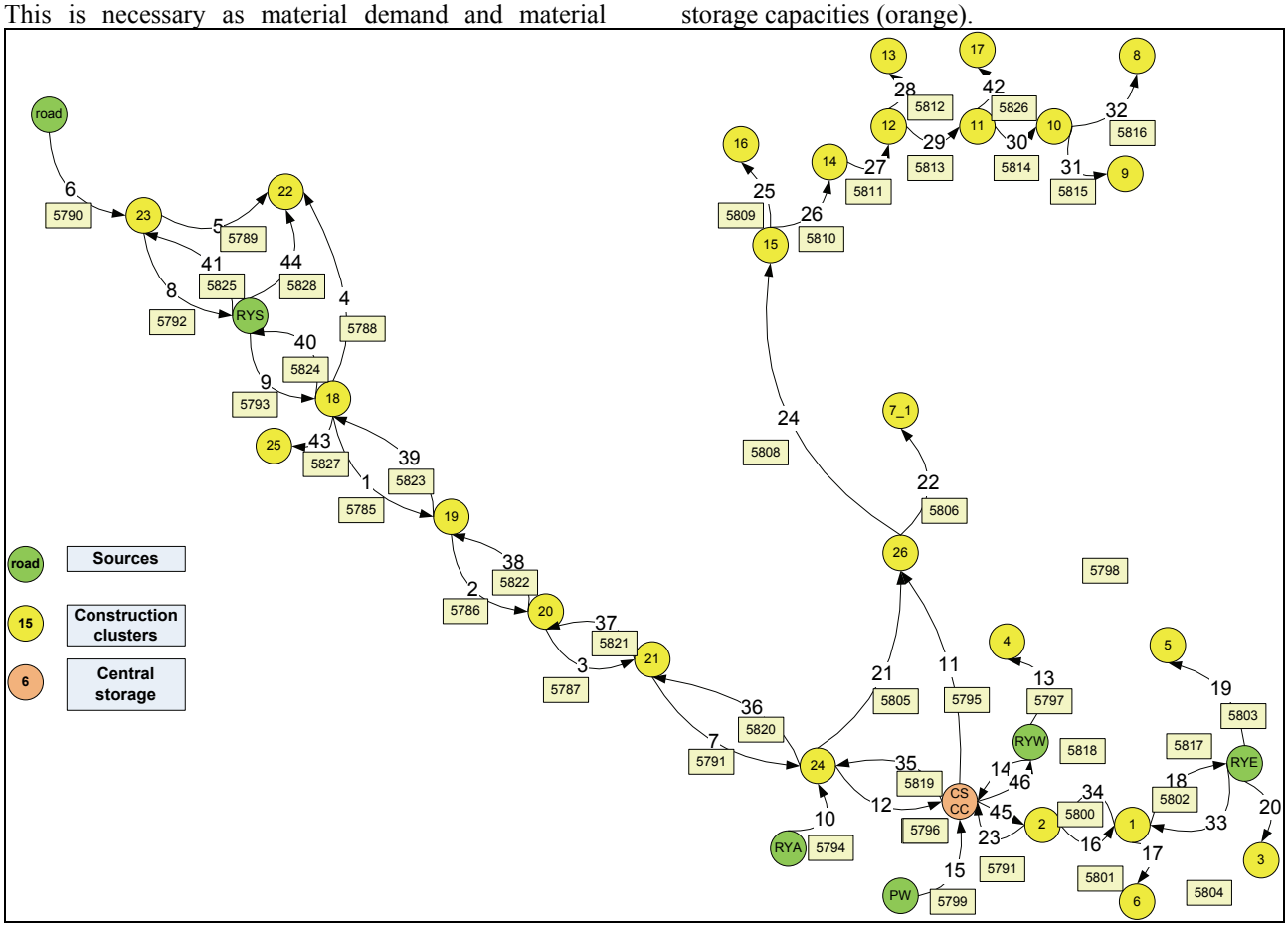


Figure 2: Graphical representation of the simulation model for Sochi 2014

Data acquisition

Due to the specific conditions of this project, data acquisition has been divided into two steps – preparation of a basic model and scenario-dependent input data for the simulation.

The first dataset comprises mainly the structural data, i.e. location of sinks and sources, transportation network specification, material types. The transportation network data has been specified by help of a geo-data database which was provided by the project partners. As seen in Figure 2, transportation channels have been implemented either in both directions or in one direction (for one-way roads).

The second dataset is much more dynamic und individually configurable. It comprises comprising material needs by material type for every sub-cluster (sinks), sourcing restrictions and dispatching rules.

In order to generate data for the second dataset, truck loads based on the material needs of the construction sites have been specified and allocated to specific inbound channels. The assumption was made that material delivery into Sochi region will correspond to the needs of the construction sites. However, a lot of restrictions and other conditions had to be taken into account during generation of truck loads, especially concerning truck profiles: 30 tons for bulk materials within main construction sites, 20 tons trucks for bulk materials outside construction sites, and 16 tons for general cargo.

Ramping-up of transshipment capacities is a crucial restriction in the beginning phase of construction works. As already mentioned, many infrastructure objects, like ports and rail freight yards are being brought into service over a specific period of time: for example in the first year only two port piers and 30% of rail freight yard are available for operations. Several other capacities are also being taken gradually into service. These limitations needed to be individually specified and integrated into the simulation model for ports, and rail freight yards. Additionally to the transshipment capacities there are also limitations concerning track capacities of the railways that must be taken into account during generation of the loads (separately for

bulk and general cargo). The user can individually define scenarios for extending such capacities.
The introduction of material groups allowed it to easily evaluate the simulation results. In this context scenario parameters comprise the allocation of individual material types to material groups (bulk, general cargo or cement).
Furthermore, priority rules for material sourcing of materials for individual construction sub-clusters need to be specified. For example, these rules define which construction materials are supplied by use of which transport modes and/or transshipment point.
In order to integrate realistic material flow behavior it is necessary to consolidate transport and supply orders. In this context lead time restrictions had to be introduced into the model: For example, these lead time restrictions specify to what extent materials may be delivered earlier into the demand region (In other words: if a transport mode has free capacities, a certain amount of material might be delivered a given time period before demand). These restrictions also apply to construction site supplies. Of course, all lead time restrictions influence inventory levels: relaxing inbound lead time restrictions can lead to increasing inventories at central storage points; in contrast, relaxing lead time restrictions for construction site supplies may lead to a reduction of these inventories, but inventory levels at the construction sites will build up.

Model Configuration

Dispatching rules are the main instrument that allows replication of operative processes in the OTD-NET environment but it depends on demand data or patterns. Yet, in the early stage of the project information about later material demands is given in low level of detail (monthly and quarterly data after 2009). Consequently, it is hard to reproduce realistic demand behavior of the construction sites.
One option to process this demand data is to uniformly distribute the given, rough data to gain daily demands. This results in steady demand profiles that don't correspond to any real demand situation, which is typically characterized by heterogeneous patterns. One solution for this problem is the introduction of stochastic variance for demand profiles if more precise data concerning operative level is not yet available. Such variance (e.g. normal distribution) would imitate realistic demand, delivery and transportation data patterns. But all these approximations of simulation parameters for sources, distributions channels and sinks must be chosen carefully and agreed with all partners.
Additionally to these demand patterns, custom dispatching rules defined how material call-offs from the central storages occur and which materials are delivered directly to the construction sites, i.e. do not pass central storage points.
Additionally, detailed information for all external transport modes is needed, though this iss mostly not available within strategic planning. A typical mistake is to assume that external transport modes and transshipment facilities are used uniformly and operated at full capacity: These assumptions do not correspond to realistic behavior. Especially in case of maritime transports a closer examination of general cargo supplies is necessary. Unlike bulk materials, general cargo is a highly specialized good and cannot be sourced flexibly from other countries using standard supply chains.

Model Evaluation

The main questions that were effectively answered by application of this simulation model focused on the utilization of transportation channels, utilization of transshipment points, development of inventory levels (including waste yards), number of trucks that are required for stable operations in the region and the influence of the operative concepts on the overall logistics concept. These KPIs have been critical for both dimensioning of transshipment infrastructure and shaping of transport strategy.
The utilization of all inbound transshipment points (ports, rail yards) is essential for ensuring a continuous material flow into the region during the construction phase. Our conceptual work focused here on the dimensioning of capacities (differentiated by different types of material handling) as well as on optimizing of multimodal transshipment facilities. Simulation could provide an accurate evaluation of the inbound material flow under consideration of different restrictions, e.g. track limitation for rail transports. Figure 3 illustrates an exemplary resulting utilization profile of one of the rail yards (in number of transports per month).

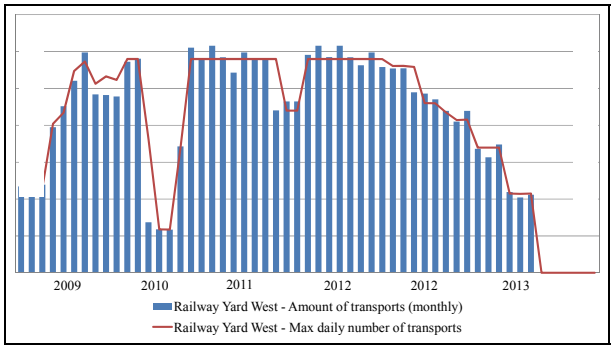


Figure 3: Utilization of the Rail Yard West

However, inbound transportation channels must not necessarily be transshipment points. Figure 4 shows the utilization profile of one of the inbound channels that was specified as a standby channel in case other inbound channels (rail and sea transports) are fully utilized. This channel represents a windy and dangerous coastal road. Any transports on this road should be avoided, so that utilization of this transportation channel was considered as a lack of main transportation or transshipment capacities.

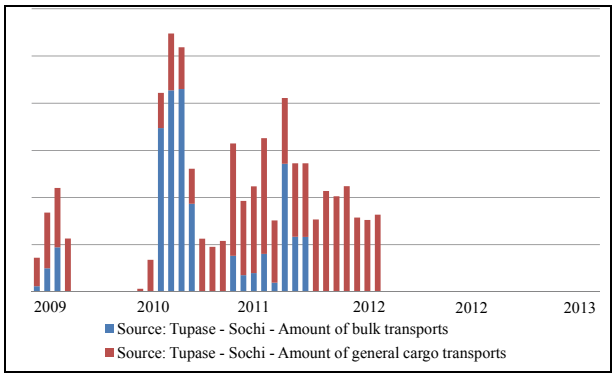


Figure 4: Number of transports on the road Tuapse – Sochi

Simulation has revealed that in several scenarios the road would have been heavily used for bulk cargo during the first half of the construction period (2010-2011) and furthermore for considerable amounts of general cargo in the year 2009 (see Figure 4). It was concluded that main transshipment points are not ready to handle sufficient amount of incoming materials (due to incremental ramp-up of the facilities). Consequently, alternative scenarios for bringing transshipment facilities into service had to be taken in consideration and corresponding measures for ensuring secure transports on that specific road were introduced in parallel.

Figure 5: Road Tuapse – Sochi

One could state that an evaluation of the total number of transports during a certain time period can easily be carried by static assessment. Yet, calculating the number of trucks simultaneously used in the system, i.e. average number of trucks at the same time on one road during a specific day) is a challenging tasks that demands consideration of many factors, e.g. operational handling parameters at the transshipment facilities. Only dynamic assessment, i.e. simulation in this case, can evaluate measures like

- maximum amount of trucks on the road;
- the maximum and average travel time in one transport channel per month;

Especially, maximum amount of trucks on the road is an important indicator for the feasibility of an operative concept that also allows the dimensioning of transportation fleets.

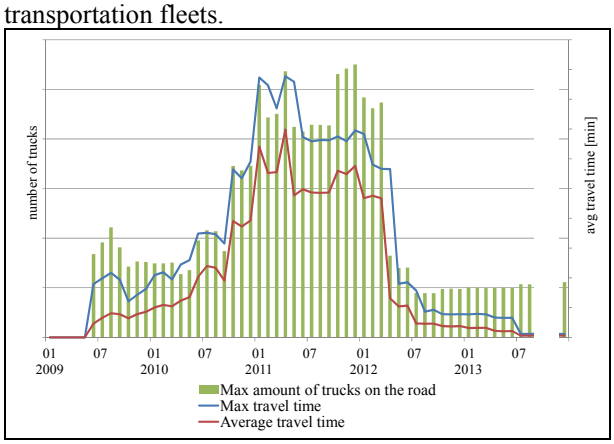


Figure 6: Key indicators for transport channels

Another key question focused on the the utilization of storage capacities. A storage concept comprises many operative concepts, i.e. order consolidation rules, dispatching rules, direct/indirect transports ratio. Applying simulation techniques made it easy to evaluate for example call-offs from the central storage, differentiated by trucks weights (see Figure 7). In combination with other measures like goods receipts, inventory levels and turnover we gained a detailed insight into the anticipatable utilization of a specific storage capacity under a given scenario. For example material call-offs illustrates the intensity of stock usage and are consulted for dimensioning of queuing areas. Furthermore, only simulation allowed assessing the implications of dynamic risk factors like delivery time fluctuations. Consequently, adequate strategies for preventing critical situations have been developed.

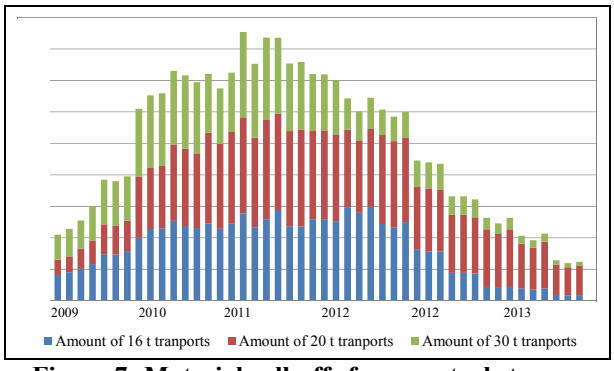


Figure 7: Material call-offs from central storage

CONCLUSION AND PROSPECTS

In this paper we presented an integrated logistics planning approach which focuses on dynamic evaluation of logistic concepts for construction and engineering projects. Key success factor is the integration of simulation experts into the logistics planning team and a close contact to the project management which allows for the avoidance of any friction losses. Any improper coordination between planning and evaluation can bee averted in advance.

Furthermore, by integration of simulation experts and their tools early in the planning process valuable insights can be revealed for the planning team.
The simulation expert may perceive early the restrictions and necessities of simulation. The simulation model itself is always designed around this knowledge: All available planning information is integrated but the level of detail is chosen carefully. So first simulations on a rough level typically abstract from stochastic influences and detailed dispatching rules, but already allow assessing dynamic influences.
This integrated approach has been applied successfully for a number of projects. For the presented project Sochi-2014 it delivered valuable insights concerning the practical application of simulation techniques for logistic planning of major construction and engineering projects. It has been agreed that methods and tools conventionally applied within these types of projects could not provide the adequate support for the challenges at hand, discussed above in this paper. The integrated simulation study provided guidance for the development of a sustainable logistic concept for the Sochi region and brought substantial value for dimensioning of both infrastructure facilities and operational processes.
Considering the high complexity, multiple restrictions and interdependencies in this logistics network, it is essential to continue this support for the operative work within the planned logistics centers. The prospective task is to apply similar simulation techniques for controlling of later material flow during the construction period itself. The benefits that arise by continuously applying the same procedures, methods, tools and experts from strategic to operative planning is highly mispriced today. To prove this is within focus of our future work.

REFERENCES

Franke, A. 1987. “Risk analysis in project management.” International Journal of Project Management, Jg. 5, H. 1, pages 29–34.

Klingebiel, K. 2009. “Entwurf eines Referenzmodells für Built-to-order-Konzepte in Logistiknetzwerken der Automobilindustrie.” Verlag Praxiswissen, Dortmund.

Klingebiel, K. and T.Seidel 2007. “Transforming the Automotive Industry by Rapid Supply Chain Design.” In New technologies for the intelligent design and operation of manufacturing networks 2007. M.Rabe and P.Mihók. IRB Verlag, Stuttgart, 53–70.

Klingebiel, K. and T.Seidel. 2006. “Methodology for the dynamic interenterprise network process design and evaluation”. Report D8.4.2. Information Societies Technology (IST) And (NNP) Joint Programme, Project Intelligent Logistics for Innovative Product Technologies (ILIPT).

Kuhn, A. and B.Hellingrath. 2002. “Supply Chain Management. Optimierte Zusammenarbeit in der Wertschöpfungskette.” Springer, Berlin.

Kuhn, A.; A. Wagenitz, A.; and K.Klingebiel. 2010 “Praxis Materialflusssimulation: Antworten zu oft zu spät.” Jahrbuch der Logistik. free beratung Gesellschaft für Kommunikation im Marketing mbH, Korschenbroich, a-b.

Miller, R. and D.Lessard 2001. „Understanding and managing risks in large engineering projects.“ In International Journal of Project Management, No.19, pages 437–443.

Rabe, M.; S.Spieckermann; and S.Wenzel. 2008. “Verifikation und Validierung für die Simulation in Produktion und Logistik.” Springer, Berlin.

Scheer, A.-W. 2002. “ARIS – vom Geschäftsprozess zum Anwendungssystem.” Springer, Berlin.

Seidel, T.; S.Wolff; K.Klingebiel; and M.Toth. 2005. “BTO Network Design and Evaluation Methodologies and Tools.” Report D7/8.1.2. Information Societies Technology (IST) And (NNP) Joint Programme, Project Intelligent Logistics for Innovative Product Technologies (ILIPT).

Verein Deutscher Ingenieure (VDI). 2000. “Richtlinie 3633, Blatt 1: Simulation von Logistik-, Materialfluss und Produktionssystemen – Grundlagen.” Düsseldorf.

Wagenitz, A. 2007. “Modellierungsmethode zur Auftragsabwicklung in der Automobilindustrie.” University of Dortmund.

Wenzel, S.; M.Weiss; S.Collisi-Böhmer; H.Pitsch; and O.Rose. 2007. “Qualitätskriterien für die Simulation in Produktion und Logistik.”Springer, Berlin.

VISUAL MODELING AND SIMULATION TOOLKIT FOR ACTIVITY CYCLE DIAGRAM

Donghun Kang and Byoung K. Choi
Department of Industrial & Systems Engineering
KAIST
335 Gwahak-ro, Yuseong-gu, Daejeon 305-701, Republic of Korea
E-mail: donghun.kang@vmslab.kaist.ac.kr

KEYWORDS
Activity Cycle Diagram, Activity Transition Table, Three-phase Rule.

ABSTRACT

A direct use of activity cycle diagram (ACD) model into a simulation execution has a limitation that it does not maximize the power of the widely adopted three-phase rule in the simulation execution of ACD models. This paper presents a key model specification for the simulation execution of the ACD model, named, activity transition table (ATT). The proposed ATT reduces the gap between the ACD (the flow of state change) and the three-phase rule (activity transition) and maximizes the modularity of the three-phase rule. The presented ATT model and ACD model can be implemented and executed with the help of the visual modeling and simulation toolkit.

INTRODUCTION

The activity cycle diagram (ACD) is a method to describe the interactions of objects in a system. It uses the common graphical modeling notation to explain series of activities in real-life diverse circumstances.

The core idea of the ACD was conceived by Tocher to describe the congestion problem at the steel plant in a general framework, called *flow diagram* (Tocher 1960) with the *three-phase rule* (Tocher 1963).

The objects in a system can be classified into two classes: 1) transient object or *entity* that receives the services and leaves the system, 2) resident object or *resource* that serves the entities.

In the ACD, the behavior or lifecycle of an entity or resource in the system is represented by an activity cycle, which alternates the active states with the passive states. The passive state of an entity or resource is called a *queue* in a circle, and the active state is called an *activity* in a rectangle as shown in Figure 1. The arc is used to connect the activity and queue.

The activity represents the interaction between an entity and resource(s), which usually takes a *time delay* to finish it. The token is used to represent the state of the queue and activity. All activity cycles are closed on itself (Carrie 1988).

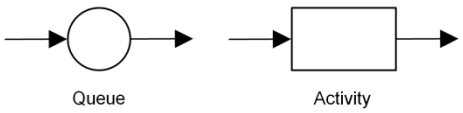


Figure 1: Basic Graphical Notation for the ACD

An ACD model for a single machine system with a setup operator is shown in Figure 2. This model consists of four activity cycles: three for resources of "generator", "machine" and "operator" and one for an entity of "jobs". A job is generated at the interval of t_a time unit by the generator and stored in a queue "B" waiting its processing on a machine. A ready-to-process machine serves a job for t_p time unit if a queue "B" has at least one job and it holds for a moment until the operator is available. The operator sets up the machine for t_s time unit as soon as it is available. Other resources also perform one or more different activities in any sequence or are idle. Here, all activity cycles are closed.

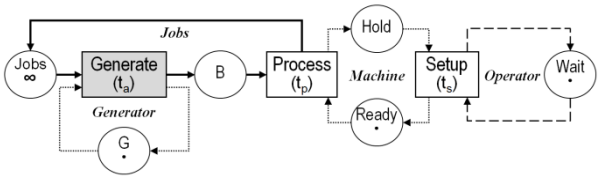


Figure 2: An Example of the ACD - Single Machine System with a Setup Operator

The three-phase rule is also proposed by Tocher (Tocher 1963) to handle the flow of time in the discrete event simulation:

Phase A: Advance the clock to the time of the next (bound-to-occur) event.

Phase B: Terminate any activity bound to end at this time.

Phase C: Initiate any activity whose condition now permits.

The ACD represents the state flow of an entity or resource in a system, while the three-phase rule is based on the event that denotes the change in the state of the model. In phase B, the activities bound to occur at a time are terminated with the release of resources and entities (into output queues), which is called bound or *bound-to-occur (BTO) event*.

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

In phase C, the *conditional events*, which satisfies the beginning condition of the availability of entities and resources, are initiated by acquiring of them (Crookes 1986).

The gap between the ACD and three-phase rule makes difficult to use these well-structured methods for the modeling of the complex system or detailed modeling. For this reason, a key model specification for the execution of ACD models, *activity transition table*, is needed. This describes the dynamics of a system in the viewpoint of the activity transition (of BTO event and conditional event). The activity transition table (ATT) has the one-to-one relationship with the atomistic structure of the three-phase rule. Therefore, it makes the three-phase rule more efficient.

The paper is organized as follows. The second section presents a model specification for the simulation execution of the ACD models, named the activity transition table. The third section presents the three-phase activity scanning algorithm simulating the activity transition table. In the fourth section, the proposed activity transition table and its execution method are realized by the visual modeling and simulation toolkit. At last, conclusions and discussions are provided in the last section.

ACTIVITY TRANSITION TABLE

The three-phase rule has the atomistic structure of advancing time and executing BTO and conditional events. In the simulation execution, the BTO event is handled by the event routine and the conditional event is executed by the activity routine.

The activity routine firstly checks the at-begin condition of an activity, whether all input queues of that activity has at least one token or not. If it is true, the at-begin state-update is fulfilled, which takes one token out of each input queue. Then it schedules a BTO event to occur in a time delay or time duration. The event routine executes the at-end state-update, which adds one token to each output queue.

The phase C of the three-phase rule has an inefficiency of scanning all activity in the ACD model, even though the BTO event has an effect only on the succeeding activities.

The activity transition table (ATT) as a model specification for the simulation execution of the ACD models is a set of activity transitions. Each activity transition has at-begin condition, at-begin state-update, BTO event with the time delay, at-end state-update and influenced activities.

The ATT model can be derived from the following formal definition of the ACD model:

$M = <A, Q, I, O, T, \mu>$, where
A is the finite set of activities,
Q is the finite set of queues,
I: $A \rightarrow \{Q\}$ is the input queues of an activity a_i,
O: $A \rightarrow \{Q\}$ is the output queues of an activity a_i,
T is the time delay function T: $A \rightarrow R+$,
μ is the finite set of tokens for each queue.

The at-begin condition specifies the condition of that every input queue qj of an activity ai should have at least one token, in short, $\mu_j > 0$, for all $q_j \in I(a_i)$. The at-begin state-update is defined as $\mu_j' = \mu_j - 1$, for all $q_j \in I(a_i)$, which decreases the token value of every input queue q_j of an activity a_i by one. The BTO event is scheduled to occur in a time delay of an activity a_i, $T(a_i)$. The at-end state-update is derived by $\mu_j' = \mu_j + 1$, for all $q_j \in O(a_i)$. The influenced activities are defined as a set of activities, $\{a_k \mid q_j \in O(a_i), q_j \in I(a_k)\}$, whose input queue is one of output queues of an activity a_i.

Table 1 shows the ATT model for the single machine system with a setup operator in Figure 2. The queue "Jobs" does not show up in the ATT model, because it is a dummy node used for making the activity cycle closed.

Table 1: ATT Model for the Single Machine System with a Setup Operator

Name	At-begin		BTO Event		At-end	
	Condition	State Update	Time	Name	State Update	Influenced Activities
Generate	G>0	G--	t_a	Generated	G++, B++	Generate, Process
Process	Ready>0 && B>0	Ready--, B--	t_p	Processes	Hold++	Setup
Setup	Hold>0 && Wait>0	Hold--, Wait--	t_s	Setup	Wait++, Ready++	Setup, Process

It is one of the advantages of the ATT model that it can handle some extensions of the ACD model (e.g. arc condition and arc multiplicity) without further extensions of it. This minimizes the modification of the simulation toolkit to cover the extended ACD.

The other advantage of the ATT model is its atomistic structure inherited by the three-phase rule. This enables the automatic code generation with ease.

In addition, the influenced activities of the activity transition make the simulation execution efficient so that the three-phase rule becomes more powerful.

THREE-PHASE ACTIVITY SCANNING ALGORITHM

The three-phase rule for the simulation execution of the ACD models is formally expressed in a three-phase activity scanning algorithm as shown in Figure 3. In this algorithm, two lists are maintained: CAL (candidate activity list) for storing the influenced activities of current activity and FEL (future event list) for storing the bound-to-occur events. CAL is a FIFO (First-In First-Out) queue, while FEL is a priority queue in ascending order of scheduled time.

The three-phase activity scanning algorithm (in short, activity scanning algorithm) starts with the initialization of system state and putting initially ready activities into

CAL. Let us make an example of the single machine system with a setup operator in Figure 2. Tokens for each queue are set to the initial token values. The "Generate" activity is the only activity whose at-begin condition is satisfied at this time. Therefore, it is put into CAL.

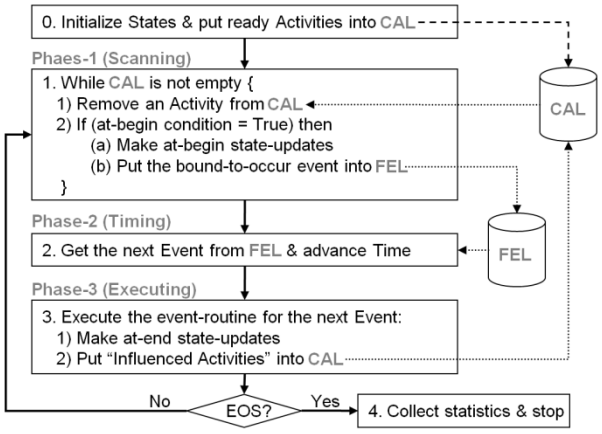


Figure 3: Three-phase Activity Scanning Algorithm

Now at the activity scanning phase of Phase 1, 1) "Generate" activity is removed from CAL, of which the at-begin condition (G>0) is satisfied, 2-a) the token for its input queue "G" are updated, and 2-b) the bound-to-occur event "Generated" is scheduled to occur at time $t_a + 0$ (i.e., stored in FEL). The current system state is:

Current State =
{G=0, B=0, Hold=0, Ready=1, Wait=1}

At the timing phase of Phase 2, the BTO event "Generated" is retrieved from the FEL (the one that has the lowest scheduled time) and the simulation clock is advanced to its scheduled time ($t_a + 0$). Then at the executing phase of Phase 3, the current marking is updated again according to the at-end state-update of the activity "Generate" and the influenced activities ("Generate", "Process") are stored in CAL. At this point, the current marking is:

Current State =
{G=1, B=1, Hold=0, Ready=1, Wait=1}

From now, the above procedure is repeated until it reaches the end-of-simulation condition.

The principle of the three-phase rule has not been changed. The difference between the traditional three-phase rule and proposed activity scanning algorithm is the order of the execution phases. The phase C starts firstly, the phase A is followed, and then phase B is executed. This is because of the influenced activities, which reduces the scanning time of activities in the phase C. The ATT has a more atomistic structure than the three-phase rule does, which is well integrated into the three-phase activity scanning algorithm with more modularity.

VISUAL MODELING AND SIMULATION TOOLKIT

Since Tocher introduced the concept of the ACD, many ACD simulation software tools have been developed in various types of ACD model implementation: 1) automatic code generating simulation software tools such as DRAFT (Mathewson 1985), CAPS (Clementson 1986), and AUTOSIM (Paul and Chew 1987), 2) simulation software tools using simulation language such as CYCLONE (Halpin 1977) and STROBOSCOPE (Martinez and Ioannou 1994), and 3) visual interactive modeling software tools using graphical modeling notations such as EZSTROBE (Martinez 2001) and GroupSim (Araújo et al. 2004).

The automatic code generation and simulation language approaches are not easy to learn and they lack the ability of simple modeling of the complex system. The visual interactive modeling approach, however, provides the graphical modeling notations so that the modeler who is not familiar with the programming can focus on its own role (Pidd and Carvalho 2006).

The modeling and analysis of the complex system still require the customization of the implemented model in the programming language, because the visual interactive modeling software tools only provide simplified output data collection and analysis.

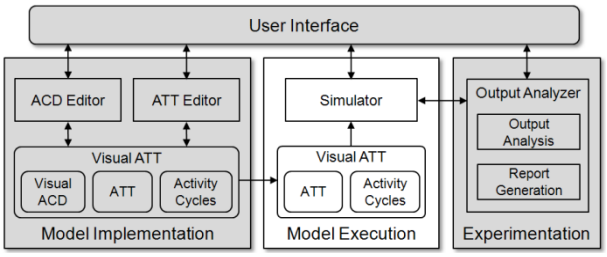


Figure 4: The Visual Modeling and Simulation Toolkit

For this reason, the simulation software tool presented in this paper, as shown in Figure 4, consists of the visual modeling toolkit and simulation toolkit for the support of both visual interactive modeling and custom model implementation. The visual modeling toolkit (in gray) provides graphical modeling notation to implement the ACD model, it also has the ability of modeling the ATT model at the same time and supports output analysis and report generation for the experimentation. The simulation toolkit (in white) is used to execute the simulation and collect the output data during the simulation execution.

Visual Modeling Toolkit

The visual modeling toolkit is developed for the model implementation and experimentation in the simulation model lifecycle. The model implementation can be done with ACD editor to create the ACD model using graphical modeling notation and ATT editor to construct the ATT model simultaneously. Two editors

take roles of view and controller in model-view-controller (MVC) pattern (Buschmann et al. 1996). The visual activity transition table (in short, visual ATT) is the model in the MVC pattern. This maintains the visual ACD (information on queue and activity nodes in a graph), ATT, and activity cycles for entities and resources.

The user can use only one of two editors or both of them. The controller of each editor receives the input from the user and notifies the visual ATT model of the user input, resulting in a change in the model, and then each view of both editors is automatically notified of the change of the visual ATT model.

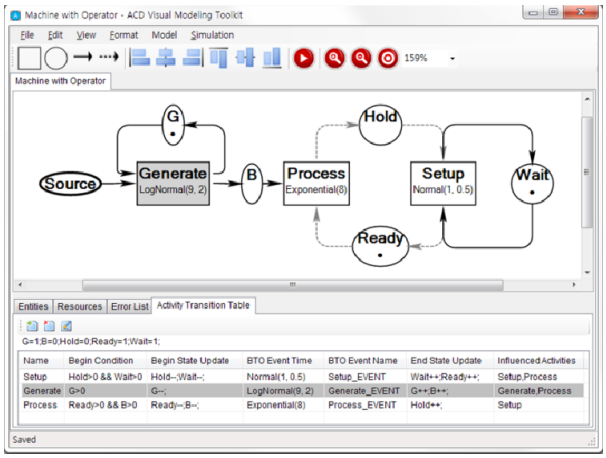


Figure 5: Visual Modeling Toolkit

In Figure 5, the visual modeling toolkit shows the dual view of the single machine system with a setup operator.

The ACD editor located at the middle shows the ACD model using graphical modeling notation: queues in a circle, activities in a rectangle. Inside of the activity node, the name and time delay of an activity is displayed. The activity node in gray represents the initial activity ("Generate" activity), which is ready to begin at the initial state. The queue node shows the name and the initial state of a queue. If the initial state of a queue is more than zero, it is displayed on the queue node with "dot (•)".

The ATT editor located at the last tab of the bottom shows the initial states of all queues and activity transitions. The activity transition can be automatically derived from the ACD model in the ACD editor, or directly inserted or edited by the user using the dialog box for an activity transition.

Prior to the experimentation, the experimental frame should be set to collect the output data and calculate the performance measures by defining the activity cycles of entities and resources. Here, one activity cycle for the "jobs" entity and two activity cycles for the "machine" xand "operator" resource can be made.

Figure 6 shows the output report for resources generated just after the experiment: the utilization of each resource is calculated. The utilization of a machine of single machine system is divided into four states of its activity cycle ("Process" activity, "Hold" queue, "Setup' activity and "Ready" queue).

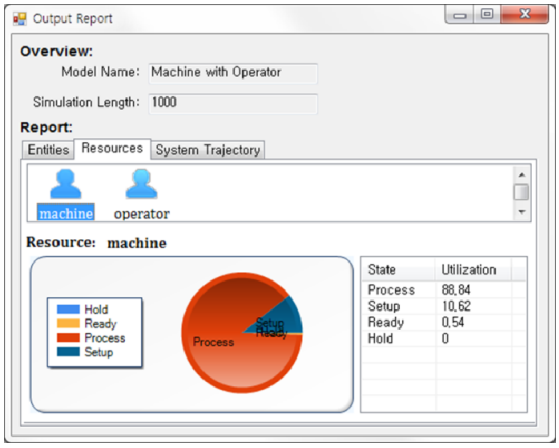


Figure 6: Output Report - Resource View

The visual modeling toolkit stores the visual ATT model into XML document so that it can be exchanged between different simulation toolkits.

Simulation Toolkit

As shown in Figure 7, the simulation toolkit is a set of libraries (model library, simulation library and output data library) for the simulation execution of the ATT model and the output data collection.

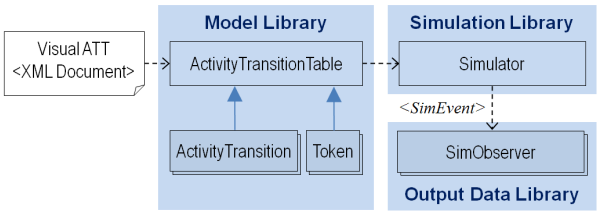


Figure 7: Simulation Toolkit

The model library converts the visual ATT in XML document into the core parts of ATT model in the *ActivityTransitionTable* class with the *ActivityTransition* and *Token* classes. It also supports the modeling of the behavior of the complex system with the inheritance of these classes.

The simulation library consists of the simulator and supporting classes by implementing the three-phase activity scanning algorithm in Figure 3. According to the activity scanning algorithm, 1) in the activity scanning phase, the simulator retrieves an activity transition by invoking the *get-activity* () method on CAL and calls the activity routine of the current activity transition. The activity routine evaluates the at-begin condition of current activity. If it is true, then the at-begin state-update is made to update the token values of

its input queues and the BTO event is scheduled by invoking *schedule-next-event* () method on FEL. This is repeated until CAL becomes empty. 2) In the timing phase, the simulator retrieves the next event by invoking *get-next-event* () method on FEL and advances the simulation clock to the scheduled time of the next event. At last, 3) in the executing phase, the simulator calls the event routine of the next event: the at-end state-update is made to update the token values of output queues and the influenced activities are stored into CAL by invoking *store-activity* () method on CAL.

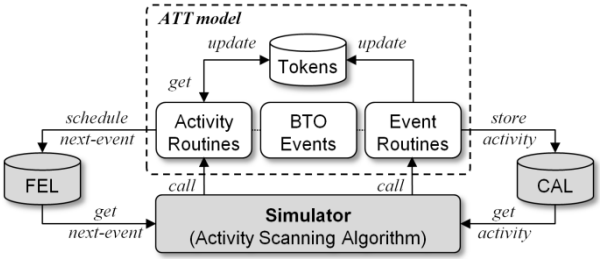


Figure 8: Simulation Library

The output data library is a set of classes and interfaces to support the output data collection, which uses publish-subscribe mechanism of the observer software design pattern (Gamma 1994). Whenever changes are made on the simulation objects such as tokens, CAL and FEL, the simulation events (SimEvent class) are published to the simulator, then these simulation events are distributed to the their subscribers, simulation observers (*SimObserver* class). The simulation observer collects these simulation events to calculate the performance measures.

In our implementation, the simulation events are fired when a BTO event is scheduled or fired or when the token value is changed. Currently, three types of the simulation observer are available: 1) event counter that records the occurrence of a specific simulation event and 2) entity/resource observer that collects any simulation event related to the activity cycle of an entity/resource.

The visual modeling toolkit and simulation toolkit work on Microsoft's .NET Framework (3.5 version) and are implemented with C# programming language. They are available at authors' web site (http://vms.kaist.ac.kr).

CONCLUSION

Proposed in the paper, the ATT is a model specification for the simulation execution of the ACD model. Distinctive features of the proposed ATT includes (1) it reduces the gap between the ACD model and its simulation execution, (2) it covers some extensions of the ACD model without further modification of the simulation toolkit, and (3) it makes the simulation execution more efficient enabling the three-phase rule to become more powerful.

The visual modeling toolkit supports both ACD and ATT modeling views. It helps the ease of ACD modeling with the graphical modeling notation and the automatic generation of ATT model of the XML document. Then, the simulation toolkit uses the ATT model to execute the simulation using the three-phase activity scanning algorithm and supports the output data collection with the publish-subscribe mechanism.

While two symbols of the ACD makes easy to learn and to model, it lacks the modeling power to describe the complex system in details (Hlupic and Paul 1994). For this reason, some researchers (Araúro and Hirata 2004; Kienbaum and Paul 1994; Martinez and Ioannou 1994; Halpin 1977) have proposed many extensions. As an academic research, the generality of the extensions has to be proved with the formal definition and it also needs to prove its real modeling power with its applications to the complex systems, such as automatic material handling system in the factory and general job shop system.

REFERENCES

Araújo, W.F.; C.M. Hirata; and E.T. Yano. 2004. "GroupSim: A Collaborative Environment for Discrete Event Simulation Software Development for the World Wide Web." *SIMULATION*, Vol. 80, No. 6, 257-272.

Araújo, W.F. and C.M. Hirata. 2004. "Translating Activity Cycle Diagrams to Java simulation Programs". In *Proceedings of the 37th Annual Simulation Symposium*. IEEE, Piscataway, N.J., 157-164.

Buschmann F.; R. Meunier; H. Rohnert; P. Sommerlad; M. Stal. 1996, "Model-View-Controller". In *Pattern-oriented Software Architecture: a System of Patterns*. John Wiley & Sons, Chichester, N.Y., 125-144.

Carrie, A. 1988. *Simulation of Manufacturing Systems*. John Wiley & Sons.

Crookes, J.G.; D.W. Balmer; S.T. Chew; and R.J. Paul. 1986. "A Three-Phase Simulation System Written in Pascal." *Journal of the Operational Research Society*, Vol. 37, No. 6, 603-618.

Clementson, A.T. 1986. "Simulating with Activities Using C.A.P.S./E.C.S.L.". In *Proceedings of the 1986 Winter Simulation Conference*, 113-122.

Gamma, E.; R. Johnson; R. Helm; and J. Vlissides. 1994. *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley.

Halpin, D.W. 1977. "CYCLONE: Method for Modeling Job Site Processes." *Journal of the Construction Division*. ASCE, Vol. 103, No. 3, 489-499.

Hlupic, V. and R.J. Paul. 1994. "Simulation modelling of flexible manufacturing systems using activity cycle diagrams." *Journal of the Operational Research Society*, Vol. 45, 1011-1023.

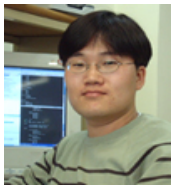
Kienbaum, G. and R.J. Paul. 1994. "H-ACD: Hierarchical Activity Cycle Diagrams for Object-oriented Simulation Modelling". In *Proceedings of the 1994 Winter Simulation Conference*. 600-610.

Martinez, J.C. and P.G. Ioannou. 1994. "General Purpose Simulation with STROBOSCOPE", In *Proceedings of the 1994 Winter Simulation Conference*. 1159-1166.

Martinez, J.C. 2001. "EZStrobe: General-purpose simulation system based on activity cycle diagrams". In *Proceedings of the 2001 Winter Simulation Conference*. 1556-1564.

Mathewson, S.C. 1985. “Simulation Program Generators: Code and Animation on a P.C.” *Journal of the Operational Research Society*, Vol. 36, No. 7, 583-589.
Paul, R.J. and S.T. Chew. 1987. “Simulation Modelling Using an Interactive Simulation Program Generator.” *Journal of the Operational Research Society*, Vol. 38, No. 8, 735-752.
Pidd, M. and A. Carvalho. 2006. “Simulation software: not the same yesterday, today or forever”, *Journal of Simulation*, Vol. 1, No. 1, 7-20.
Tocher, K.D. 1960. “An integrated project for the design and appraisal of mechanical decision-making control systems.” *Operational Research Quarterly*, Vol. 11, No. 1/2, 50-65.
Tocher, K.D. 1963, *The art of simulation*, English Universities Press.

AUTHOR BIOGRAPHIES



Donghun Kang is a PhD candidate student in the Department of Industrial & Systems Engineering at KAIST. He received as BS from KAIST in 2003 in Computer Science, a MS from KAIST in 2005 in Industrial Engineering. His research interests are in the area of system modeling and simulation. His e-mail address is donghun.kang@vmslab.kaist.ac.kr and his web site can be found at http://www.dhkang.org/.

Byoung K. Choi is a professor of the Department of Industrial Engineering at KAIST since 1983. He received a BS from Seoul National University in 1973, a MS from KAIST in 1975, and a Ph.D. from Purdue University in 1982, all in Industrial Engineering. His current research interests are system modeling and simulation, BPMS, simulation-based scheduling, and virtual manufacturing.

VIRTUAL COMMISSIONING OF MANUFACTURING SYSTEMS A REVIEW AND NEW APPROACHES FOR SIMPLIFICATION

Peter Hoffmann
Reimar Schumann
Fakultät II – Abt. Maschinenbau
University of Appl. Sc. and Arts Hannover
Ricklinger Stadtweg 120
30459 Hannover, Germany
E-mail: peter.hoffmann@fh-hannover.de

Talal M.A. Maksoud
Giuliano C. Premier
Faculty of Advanced Technology
University of Glamorgan
8 Forest Grove
Pontypridd, Wales, UK, CF37 1DL
E-mail: tmaksoud@glam.ac.uk

KEYWORDS
Modelling, virtual commissioning, manufacturing systems.

ABSTRACT

Virtual commissioning (VC) of manufacturing systems has been researched for more than 10 years. Its intention is to test manufacturing systems and associated control programs through simulation conducted before the real systems are realised. The expected benefits in reducing debugging and correction efforts expended during real commissioning, however, can only be achieved if sufficiently detailed manufacturing system models are available for simulation. To date, the design of such models has certainly required a high level of expertise and considerable effort, which makes virtual commissioning unattractive, especially for small and medium-sized enterprises (SME). After reviewing the current status of VC, this paper describes some new concepts for the systematic and simplified design of manufacturing system models for VC based on model libraries and standardized recipes for the design of component models from CAD data. This work is carried out as part of the research cooperation between the University of Glamorgan and the University of Applied Sciences and Arts Hannover.

INTRODUCTION

Today the design of manufacturing systems takes place in an industrial environment characterized by significant cost constraints, shortening of product life-cycles and strategies for rapid time-to-market. For these reasons, the timeframe for manufacturing system engineering is progressively tightening whereas the demands on planning accuracy and planning quality are growing.

Manufacturing systems consist of different elements such as storage, magazines, conveyors, handling and transportation systems, machining and assembling tools, robots, control and HMI systems, often in a combination of a large number of standard parts and some purpose-built parts or sub-systems. The development of a manufacturing system, in general, comprises several phases: facility design, mechanical engineering, electrical engineering and automation engineering (programming of robots, PLCs and HMI), which are often sequentially executed.

There are many different powerful and specialized tools for design and engineering, often with integrated simulation, however, there are great problems regarding data exchange between the different engineering phases and the different associated tools. A typical problem is the repeated data entry generating random errors. One principal problem is the missing generally accepted data exchange format which might be solved by AutomationML® (Drath et al. 2008a).

Up to now, after completion of engineering, procurement and assembly, the real commissioning is finally done. Conventionally an integrated test of the planned manufacturing system cannot be done before it has been built; consequently a considerable number of design problems and faults often remains undetected before the first system start-up. This leads in general to time and money consuming corrective measures being required during commissioning and the early production phases resulting in time delays and increased costs to all parties involved.

According to (Zäh and Wünsch 2005), referencing a study of the VDW (German Association of machine tool builders), the commissioning time consumes up to 25% of the time available for plant engineering and construction; and up to 15% is expended on correcting errors in the control software alone. As a possible solution to these problems the authors propose virtual commissioning (VC).

During VC, a simulation model of the manufacturing system is used to allow commissioning through simulation, before building the system. The goal is the early detection and correction of errors generated during planning, design and programming. VC may be conducted during all engineering phases, but at the latest after engineering has been completed.

An experimental study in virtual commissioning (Zäh et al. 2006) shows the positive effects of VC on the error

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

rate during real commissioning. The study was conducted with two groups of control programmers. Each group had 30 individuals. One group applied VC to the software development for a machine. The results were compared to those from the second group of programmers that did not use VC. A tin can moulding press was used as a test bed, with a Siemens S7-300 PLC which uses 10 actuator outputs and 17 sensor inputs. One group programmed the PLC and tested the program afterwards in a real world commissioning on a real machine. The other group programmed by using a virtual machine model. They did not execute the real commissioning before achieving successful VC. The results showed a reduction of real commissioning time by 75%, resulting from enhanced software quality at the start of real commissioning. This emphasizes the advantages of running a VC, but the virtual machine model had already been developed in the run-up to this study, and this effort was not taken into consideration. In their conclusion the authors point out the need for simplified and accelerated model building.

APPROACHES TO REALISING VC

Investigating the feasibility of VC has been an academic and industrial research objective for several years. For separate verification of geometry, kinematics and mechanical design, a 3D simulation of the expected and specified mechanical behaviour is sufficient. A VC based on such simulations is able to detect mechanical resp. geometrical planning errors. For separate verification of the control programs, a simulation reflecting the specified behaviour of the manufacturing system mechanics at I/O level is needed. A VC based on such simulations is able to detect deviations from the specified control functions. If the impact of control programs on the 3D mechanical behaviour of the manufacturing system is to be tested in detail in an integrated manner, modelling and simulation of the complete functional chains from control programs through sensors, actuators and drives onto the mechanical movements, is necessary which includes both, simulation of mechanical behaviour and of control programs .

An early approach to VC was presented by Auinger et al. (1999) and termed as “soft-commissioning”. The authors propose VC because of the time consuming and expensive testing and debugging required by PLC based control software. As a test bed for their VC approach, the authors used a PLC controlled pallet transfer system. The modelling of the transfer system, based on the simulation tool ARENA®, was done for the design and optimization of the manufacturing process. This model contained all of the control logic. Later they reduced the model, and the model used for the soft-commissioning no longer contained the control logic. The model provided only realistic I/O signals for the coupled PLC and visual feedback. In their experiments, they found malfunctions of stoppers caused by overlooked inverse logic and problems in the material flow. Normally such errors would only be detected during real commissioning. The modelling effort was not (probably with good cause) discussed.

In accordance with Auinger et al. (1999) the verification of control can be arranged in four basic system configurations:

1. Real plant and real control system: The traditional way of testing during real commissioning.
2. Simulated plant and real control system: “Soft-commissioning” often called “hardware in the loop” (HIL). The hardware controller is necessary in advance, but a VC before building the plant is possible.
3. Real plant and simulated control system: termed “Reality in the loop” by the authors.
4. Simulated plant and simulated control system: This offers a complete VC.

A VC in the second or fourth configuration requires the coupling between real or simulated controller and the mechanical plant simulator which can be realized with e.g. OPC. In the fourth configuration the simulation of controller and plant can run in one tool. This is for example possible for several robot controller and Siemens S7 PLC inside the plant simulation tool CIROS® (RIF 2010).

Currently, two research domains are linked to VC. The first domain deals with test and verification of control programs. This is possible by applying many different techniques, from testing on the real plant, to formal logic analysis. A comprehensive survey of 18 different methods can be found in (Danielsson et al. 2003). In addition to the 4 basic configurations mentioned above, the authors refer to hardware test panels for process simulation, which allow only a rough plant simulation, or simulation with process models blocks inside the control software tool. A useful survey of the research area related to formalization of existing PLC code was given by (Bani Younis and Frey 2003). For VC the verification of formalized control code against an abstract model of the plant, in the language of the analytical tool, would be necessary. The generation of such models is a theoretical and practical problem and a usable industrial implementation for VC is not foreseeable today.

In (Thapa et al. 2006) an approach for offline verification and validation of the control logic is defined and justified by the lack of tools to provide completely integrated solutions for the verification and validation of control logics. The authors propose a III-phase method and classify these phases as: Manual testing, Model checking and VC. Manual testing means checking the code on a softPLC (or simulated PLC like S7-PLCSIM) by user inputs; this is only useful for small programs or parts of programs. Model checking uses formalization as described by (Bani Younis and Frey 2003). The

standard IEC 61131-3 code is converted to an intermediate language, transformed to timed automata and the model checked. For the virtual commissioning they coupled a virtual plant model and a softPLC running the checked standard code. The III-phase method requires considerable effort and expertise to build the virtual system, which is costly. The authors referred to the market situation where customers want to validate the process using VC or simulation using 3-D models and not textual verification only.

The second domain addresses the scope of the "Digital Factory" (Kühn 2006, Schlögl 2007). By means of digital models, computer-aided planning and design, computer-aided engineering, associated software tools and with the aid of integrated data management, the "Digital Factory" would permit integrated planning, simulation and validation of manufacturing processes and systems (VDI 2008). The use of the "Digital Factory" should extend to all phases of manufacturing system development. Nowadays simulation is already used intensively in some phases, e.g. material flow simulation for facility design, 3D simulation of kinematics for mechanical engineering and possibly interaction with product data at Digital Mock-Up (DMU). Off-line programming and 3D robot simulation have been used for many years, sometimes 3D human models are used today to evaluate ergonomics in design. Up to now the focus of simulation has been primarily in design and mechanical engineering, where fit and specified behaviour of mechanical plant components must be ensured.

In the context of the "Digital Factory" it is in principle possible to use the complex off the shelf engineering tools from major vendors for a VC, but this usually requires a high level of training or in-house secondment of specialists from the vendor so that only large enterprises (e.g. in the automotive industry) selectively choose to conduct a VC. These "Digital Factory" solutions cannot solve the problems of small and medium-sized enterprises, because they normally do not have the resources to use these techniques (Westkämper et al. 2003), and there is therefore limited use of the "Digital Factory" today (Drath et al. 2008b). The generation of simulation models with consideration of mechanical and electrical planning data and control programs is especially associated with high effort.

In (Wischnewski 2007) the author described the VC procedure when using the simulation tool COSIMIR® and its extension module Cosimir Transport (Wischnewski and Freund 2004) for carrier based transport systems. The author justifies the VC by the exceedingly error-prone control design of transport systems originating from the use of many sensors and actuators and the complex program sequences required for the routing strategies implemented. Rossmann et al. (2007) described a VC with COSIMIR® using an example of the iCIM manufacturing systems from Festo®. The authors specified a reduction of engineering effort by up to 50% and an up to 50% faster start of production (SOP) when all component models were available in the internal simulation model library of COSIMIR®. Additionally, a library of reusable controller programs was developed. The effort for the first system is specified to have been 30% greater, with a 20% delayed SOP. This emphasizes the importance of model libraries with standardised automation components.

This short review already shows the current options and limitations of VC for manufacturing systems. On the one hand beneficial effects such as reduced real commissioning time or improved planning quality are emphasized (Auinger et al. 1999, Zäh et al. 2006, Reinhart and Wünsch 2007, Wischnewski 2007, Rossmann et al. 2007) but on the other hand the modelling required for the virtual manufacturing system (if not just neglected) is judged by many authors to be difficult and associated with large effort (Moore et al. 2003, Park et al. 2006, Zäh et al. 2006), regardless of which simulation tool is used. The review shows moreover, a lack of accessible and 'easy to use' engineering and simulation environments which could assist the engineer to set up and conduct VC.

Thus this review indicates especially the need for improved model building methods to minimize the effort and expertise required to build a virtual manufacturing system, which can be used for checking control code and planned physical setup in a VC. This must comprise especially the set up and utilization of model libraries with standardised mechanical, electrical and control components in an engineering and simulation environment which is accessible even by SMEs.

GENERATION OF PLANT MODELS FOR VC

The investigation of possibilities for simplifying the generation and use of simulation models in a VC is a collaborative research project of the University of Applied Sciences and Arts (UASA) Hannover and the University of Glamorgan. The starting point for this project is the industrial 3D plant simulation tool CIROS®, originally developed at the institute for robot research (University of Dortmund), as a robot simulation tool COSIMIR® (Freund et al. 1994, Freund and Rossmann 1995). CIROS® allows the integrated execution of 3D mechanics with robot and control programs using either internally emulated controllers or external real or virtual controllers via OPC. Besides this, CIROS® provides features such as sensor and actuator simulation, collision detection, transport simulation for carrier based systems or AGVs, and also an XML model interface. The basic model generation concept for mechanical components within CIROS® is organized in two levels, here called high-level modelling and low-level modelling.

High-level Modelling

CIROS® allows the composition and simulation of virtual manufacturing systems based on an internal component model library containing several mechanical components including robot models. These models already contain the functional interaction of mechanical behaviour with actuators and sensors. If it is possible to compose the virtual manufacturing system from such library components – high-level modelling - it is relatively easy to set up and conduct VC, where some additional effort arises when composing the plant model within the 3D editor, configuring I/O connections and transferring controller programs.

Low-level Modelling

If there are no appropriate simulation models available in the library, the effort is far greater because it is necessary to build new models based on CAD data, which is not only a problem in CIROS® but similarly for other simulation tools such as e.g. Delmia Automation®. This low-level modelling comprises the whole functional chain, and is a non-trivial task requiring considerable modelling expertise (Park et al. 2006). Here it is necessary to carry out the geometrical, functional and electrical modelling to create a structured mechatronic component model.

Geometrical Modelling

The generation of a structured mechatronic model starts with the import of the geometry data from a CAD system. For this purpose CIROS® provides import filters for e.g. STEP and IGES. Manual simplification of overly complex geometry data may become necessary. The import of standard CAD data often results in an unstructured geometrical model as shown in figure 1 (①), illustrating an example component for a transportation system.

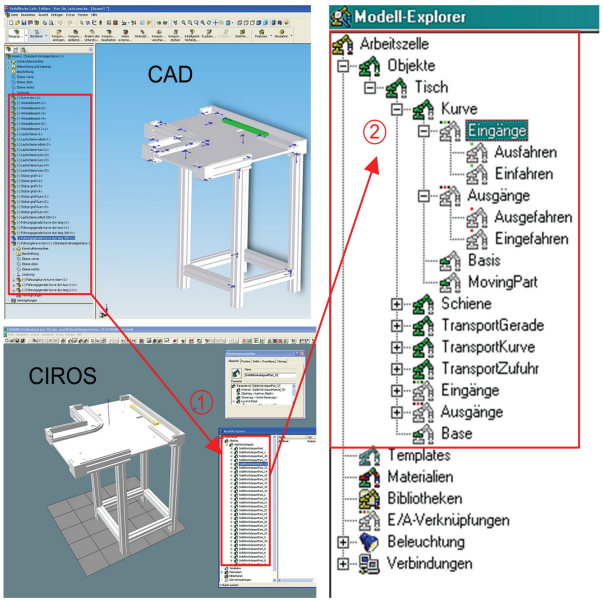


Figure 1: From CAD data to simulation object

In this case, manual hierarchical structuring of the CAD data into objects (like robots, sensors, tools or work pieces), sections (single static units e.g. joints of a robot) and components or hulls (describing the geometry e.g. cylinders, spheres or points), is the crucial step in creating the model adumbrated in figure 1- (②). If the CAD data provided are not appropriately structured e.g. with regard to moving parts, the resulting geometrical model is not directly usable for the following functional modelling, nor would be a simulation based on such model. In the worst case a CAD redesign may become necessary to provide the necessary structures in the geometry data.

Having created appropriately structured geometrical models these must be provided with functions (functional modelling) and electrical inputs and outputs (electrical modelling).

Functional Modelling

In the functional modelling stage it is necessary to manually allocate actuator functions such as translation, rotation, gripping etc. and sensor functions to selected parts of the geometrical models, which results in the definition and parameterisation of integrated functional models such as cylinder, turntable, sensor, gripper and so on.

Electrical Modelling

For the final electrical modelling it is necessary to manually add electrical inputs/outputs to the functional models for later connection to inputs/outputs of control programs thus creating complete mechatronic models.

This low-level modelling procedure can only be done by "technically experienced persons with detailed knowledge of mechanical, electrical, pneumatic, and geometric data of single components" (Wischnewski and Freund 2004).

NEW CONCEPTS FOR SIMPLIFIED VC OF MANUFACTURING SYSTEMS

The Model^{CAT} project (Hoyer et al. 2008) at the UASA Hannover demonstrated for chemical processes, that in principle it is possible to automatically generate simulation models of chemical processes for VC based on the data stored in a CAE planning tool. This allows to rapidly conduct a VC after planning, engineering and programming have been completed, at the latest, because then all necessary data for assembly and real commissioning are available in the CAE database.

This current project investigates to what extent the Model^{CAT} concept is transferable to manufacturing systems. A CAE planning tool with object oriented database will be used to hold the planning data and additional data needed for simulation (mechatronic models, control programs). The combination of various data (mechanical data, electrical data, programs if

applicable, and data needed for simulation), together in a CAE planning database tool, is a novel approach. The vision is the combination of a CAE planning tool with a simulation tool like CIROS® which should ideally make it possible to assemble the mechatronic components already existing in the CAE planning tool and then to simulate the system with CIROS®. Thus, the manual high-level modelling procedure in CIROS® could be, at least, partly omitted as the plant simulation model with I/O mapping would be built automatically from the CAE planning database.

In order to validate this concept which would provide simplified generation of virtual manufacturing systems for VC, an overarching tool called the "Prototypical Engineering Modelling and Simulation Environment (PEMS)" was designed (Figure 2). CIROS® as the central simulation tool plays one key role in PEMS, the other key role the CAE planning tool COMOS® with object-oriented database (COMOS 2010). Additional off the shelf tools are used for programming PLCs, robots and HMIs.

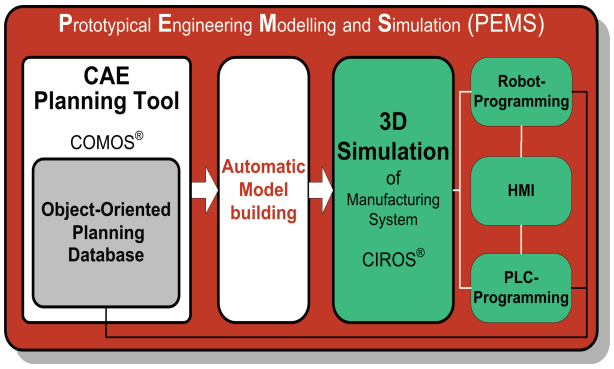


Figure 2: Prototypical Engineering, Modelling and Simulation Environment

Currently, CAE tools such as COMOS® are mainly used for engineering planning and documentation purposes. The integration of data for the simulation of manufacturing systems has not been investigated hitherto.

The first task in developing PEMS is the investigation of how to automatically generate simulation models for CIROS® from the CAE database, with the aid of an additional model building tool, in the case where the component models of the virtual manufacturing system already exist in a component library (*high-level modelling*).

Another task will be the investigation of systematic set-up and extension of component model libraries, because typically projects in manufacturing systems may use new components for which simulation models do not exist in the component model library. Therefore the *low-level modelling* procedure needs systematisation where the formulation of requirements concerning function-oriented structuring of CAD data to geometrical objects by design engineers, is crucial.
In this project the set-up of exemplary simulation models in the component model library of the COMOS® CAE database will be done manually in the first step. However, in future, importing such simulation models into the component database with minimum effort, is an important strategic point: one possible future strategy is the systematic collection of models from component and subsystem vendors (Figure 3), which may allow the provision of such models together with the hardware as predicted in (Schlögl 2007).

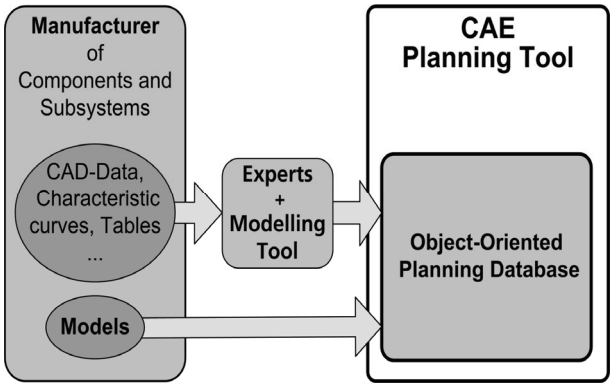


Figure 3: Future component model generation strategy

A promising new approach, supporting this strategy, is the development of AutomationML® (Drath 2010) by companies like Daimler, ABB, KUKA, Rockwell Automation and Siemens along with the Universities of Karlsruhe and Magdeburg and some smaller engineering companies. AutomationML® will provide an intermediate format for automation data exchange including component model data. AutomationML® uses CAEX as top level format for the description of the topology, COLLADA for geometry and kinematics and PLCopen XML for the overall behaviour (including electrical and control). The intention is the reduction of engineering efforts and improvement of quality by interconnecting heterogeneous tools, which may become especially valuable when setting up VC with different tools and exchange of model data using AutomationML® (Drath et al. 2008a).

Test Bed Validation of the Concept

The test bed selected to investigate the "Prototypical Engineering Modelling and Simulation Environment (PEMS)" is a small pilot manufacturing cell, referred to as "Robot-based Flexible Assembly System" at the UASA Hannover. It consists of a Siemens S7 PLC controlled transportation system with autonomous track-bound transport cars, an associated elevator and two robot-based assembling units (SCARA robot and robotic palletiser) as shown in figure 4. The robotic palletiser contains a Berger-Lahr Motion controller/Soft PLC programmed with CoDeSys, the SCARA robot is programmed with a proprietary robot language. The

transportation system includes an identification and data storage system which consists of data read/write stations on the tracks and mobile data carriers in the transport cars.

Figure 4: Robot-based Flexible Assembly System

The validation of the new PEMS approach for a simplified VC procedure will start with the low-level modelling of hardware components while following the defined modelling systematic, and storing component models of the pilot manufacturing cell in the CAE planning database of COMOS®. After re-engineering of the pilot manufacturing cell in COMOS® the next validation step will be the automated (as far as possible) model building of the virtual manufacturing cell for simulation using CIROS®. The validation will end with the execution of a VC, during which the system functions, including mechanical behaviour and control functions, will be tested by simulation. Purposely selected faults will be contrived in order to determine the efficiency of VC.

SUMMARY AND FUTURE PROSPECTS

This project is expected to demonstrate how it could become possible for control system engineers and commissioning engineers to conduct a VC for manufacturing systems based on planning data, without requiring a high level of expertise in model building and simulation. This will necessitate that CAE and simulation tools provide the appropriate functionality, especially high-level simulation model building from the CAE planning database. In addition to the improved tool environment, the object-oriented database containing simulation models of manufacturing system components, will be the subject of detailed investigations and testing. From the authors point of view it is reasonable and essential to relocate the low-level modelling effort as far as possible to the component source, i.e. to the designers, manufacturers and suppliers of components or subsystems. The manufacturers should be encouraged to provide mechatronic simulation models of their components. To support this purpose procedural methods for simplified low-level modelling strategies are investigated in the context of this project. The resulting component or subsystem models should be tool-independent, because it is not reasonable for the manufacturers to create different models for different CAE planning tools and/or simulation tools. For this purpose it will be necessary to define a standardised structure and data format for mechatronic models in future. AutomationML® is an applicable format and has the potential to support such data formats. Customer request for component simulation models, not only CAD data (which are currently often hard to use for simulation) or electronic datasheets, would accelerate this process. Experts from universities, in cooperation with component manufacturers, should make the first move to build up exemplary simulation model libraries of components or subsystems, for use in CAE environments like the proposed PEMS.

The new approach presented here, with the pilot manufacturing cell acting as test bed for the prototype implementation, will show a solution roadmap to reduce the considerable efforts required for the modelling process to such a degree that also moderately resourced small and medium-sized enterprises may become able to employ VC to optimise their engineering processes.

REFERENCES

Auinger, F., Vorderwinkler, M. and Buchtela, G. 1999. "Interface driven domain-independent modeling architecture for 'soft-commissioning' and 'reality in the loop'", *Proceedings of the 1999 Winter Simulation Conference*, Phoenix, AZ, USA, 1, pp. 798-805.

Bani Younis, M. and Frey, G. 2003. "Formalization of existing PLC programs: A Survey", *CESA 2003*, Lille (France).

Comos Industry Solutions. 2010. COMOS, Available: http://www.comos.com [accessed 02. Feb. 2010]

Danielsson, F., Moore, P. and Eriksson, P. 2003. "Validation, off-line programming and optimisation of industrial control logic", *Mechatronics*, 13, 6, pp. 571-585.

Drath, R. (Ed.) 2010. "Datenaustausch in der Anlagenplanung mit AutomationML ", Springer Verlag, Berlin Heidelberg, pp. 326

Drath, R., Lüder, A., Peschke, J. and Hundt, L. 2008a. "AutomationML - the glue for seamless automation engineering", *IEEE International Conference on Emerging Technologies and Factory Automation, ETFA 2008*, Hamburg, pp. 616-623.

Drath, R., Weber, P. and Mauser, N. 2008b. "An evolutionary approach for the industrial introduction of virtual commissioning", *IEEE International Conference on Emerging Technologies and Factory Automation, ETFA 2008*, Hamburg, pp. 5-8.

Freund, E. and Rossmann, J. 1995. "Systems approach to robotics and automation", *Proceedings of the 1995 IEEE International Conference on Robotics and Automation*, Nagoya vol.1, pp. 3-14.

Freund, E., Rossmann, J., Uthoff, J. and Can Der Valk, U. 1994. "Towards realistic simulation of robotic workcells", *Proceedings of the IEEE/RSJ/GI International Conference on Intelligent Robots and Systems '94. 'Advanced Robotic Systems and the Real World', IROS '94.* , Munich, 1, pp. 39-46.

Hoyer, M., Schumann, R., Hoffmann, P. and Premier, G. C. 2008. "Virtuelle Inbetriebnahme mit Model^{CAT} - Vom Prototypen zum industriellen Einsatz / Virtual Start-Up with Model^{CAT} - From Prototypical Realisation to Industrial Implementation", *Automation 2008*, Baden-Baden, VDI-Berichte 2032, pp. 203-206.

Kühn, W. 2006. "Digital Factory - Simulation Enhancing the Product and Production Engineering Process", *Proceedings of the Winter Simulation Conference WSC 06.*, pp. 1899-1906.

Moore, P. R., Pu, J., Ng, H. C., Wong, C. B., Chong, S. K., Chen, X., Adolfsson, J., Olofsgard, P. and Lundgren, J. O. 2003. "Virtual engineering: an integrated approach to agile manufacturing machinery design and control", *Mechatronics*, 13, 10, pp. 1105-1121.

Park, C. M., Bajimaya, S. M., Park, S. C., Wang, G. N., Kwak, J. G., Han, K. H. and Chang, M. 2006. "Development of Virtual Simulator for Visual Validation of PLC Program", *International Conference on Computational Intelligence for Modelling Control and Automation,and International Conference on Intelligent Agents,Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*.

Reinhart, G. and Wünsch, G. 2007. "Economic application of virtual commissioning to mechatronic production systems", *Production Engineering*, 1, 4, pp. 371-379.

RIF - Dortmunder Initiative zur rechnerintegrierten Fertigung (RIF) e. V. 2010. CIROS Engineering, Available: http://www.ciros-engineering.com/en.html [accessed 02. Feb. 2010]

Rossmann, J., Stern, O. and Wischnewski, R. 2007. "Eine Systematik mit einem darauf abgestimmten Softwarewerkzeug zur durchgängigen Virtuellen Inbetriebnahme von Fertigungsanlagen von der Planung über die Simulation zum Betrieb", *atp - Automatisierungstechnische Praxis*, 2007, 07, pp. 52-56.

Schlögl, W. 2007. "Einsatz der Digitalen Fabrik von der Anlagenplanung bis in den Laufenden Betrieb", *GMA-Kongress 2007*, Baden-Baden, VDI-Berichte 1980, pp. 717-725.

Thapa, D., Park, C. M., Dangol, S. and Wang, G.-N. 2006. "III-Phase Verification and Validation of IEC Standard Programmable Logic Controller", *International Conference on Computational Intelligence for Modelling Control and Automation and International Conference on Intelligent Agents,Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*.

VDI-Gesellschaft Fördertechnik Materialfluss Logistik 2008. VDI-Richtlinie 4499 - Blatt 1: Digitale Fabrik - Grundlagen

Westkämper, E., Bierschenk, S. and Kuhlmann, S. 2003. "Digitale Fabrik – nur was für die Großen?", *wt Werkstattstechnik online*, 93, 1/2, pp. 22-26.

Wischnewski, R. 2007. "Virtuelle Inbetriebnahme", *A&D Kompendium 2007/2008*, pp. 64-66.

Wischnewski, R. and Freund, E. 2004. "COSIMIR Transport: Modeling, simulation and emulation of modular carrier based transport systems", *International Conference on Robotics & Automation*, New Orleans, LA, United States, 2004, pp. 5171-5176.

Zäh, M. F. and Wünsch, G. 2005. "Schnelle Inbetriebnahme von Produktionssystemen", *wt Werkstattstechnik online*, 95, 9, pp. 699-704.

Zäh, M. F., Wünsch, G., Hensel, T. and Lindworsky, A. 2006. "Nutzen der virtuellen Inbetriebnahme: Ein experiment - Use of virtual commissioning: An experiment", *ZWF Zeitschrift fuer Wirtschaftlichen Fabrikbetrieb*, 101, 10, pp. 595-599.

AUTHOR BIOGRAPHIES

Peter Hoffmann was born in Hannover, Germany and went to the University of Applied Sciences and Arts Hannover where he studied computer engineering and obtained his Dipl.-Ing. degree in 2000. He is now working as research staff member at the automation engineering lab at the same university, and he is currently a PhD student at the Faculty of Advanced Technology at the University of Glamorgan. His main research interest is in the field of control design for manufacturing systems and more specifically he is interested in verification and validation of PLC programs. His e-mail address is : Peter.Hoffmann@fh-hannover.de

Reimar Schumann received the PhD degree in Automatic Control from the Technical University Darmstadt in 1982. During his industrial career he was responsible for the design and development of a third generation digital process control system. Since 1989 he is teaching process control at the University of Applied Sciences and Arts Hannover. His current research interest is focussed on process and production control and system design. His e-mail address is : Reimar.Schumann@fh-hannover.de

Talal M. A. Maksoud is a Reader in the University of Glamorgan. He has over 30 years experience in research. He has published over 75 papers in Journals and international conferences. He has published several best papers in the field of advanced grinding technology. His main interests are in Grinding technology, heat transfer analyses, computational fluid dynamics, Aerodynamics, and Artificial Intelligence use in Manufacturing. Dr Maksoud has supervised over dozen PhD's as well as MPhil's. He has acted as external and internal examiner for several PhD's. He is a recognised referee for several peer reviewed academic Journals.
His e-mail address is : tmaksoud@glam.ac.uk

Giuliano C. Premier is a Reader in Low Carbon Systems Engineering and senior member of the Sustainable Environment Research Center (SERC). His research activities cover modelling, control, instrumentation, renewable energy systems, biological wastewater treatment. His Ph.D. concerned the modelling and control of anaerobic digestion (AD) processes and he has since been involved with research into anaerobic microbial processes. His research also includes industrial computer aided control system design (through a longstanding collaboration with the University of Applied Sciences, Hannover) and instrumentation for soil and groundwater bio-activity monitoring. He has and continues to contribute to collaborative projects with universities and industrial partners.
His e-mail address is : gcpremier@glam.ac.uk

Multiple Scenarios Computing in the Flood Prediction System FLOREON$^+$

Jan Martinovič, Štěpán Kuchař,
Ivo Vondrák, Vít Vondrák
VŠB - Technical University of Ostrava
FEECS, Department of Computer Science
17. listopadu 15
Ostrava - Poruba
Emails: jan.martinovic@vsb.cz, stepan.kuchar@vsb.cz,
ivo.vondrak@vsb.cz, vit.vondrak@vsb.cz

Boris Šír, Jan Unucka
VŠB - Technical University of Ostrava
Institute of geoinformatics
17. listopadu 15
Ostrava - Poruba
Emails: sir.boris@vsb.cz, jan.unucka@vsb.cz

KEYWORDS

Hydrologic forecasting, rainfall-runoff models, hydrodynamic models, scenarios making, FLOREON$^+$

ABSTRACT

Floods are the most frequent natural disasters affecting the Moravian-Silesian region. Therefore a system that could predict flood extents and help in the operative disaster management was requested. The FLOREON$^+$ system was created to fulfil these requests. This article describes utilization of HPC (high performance computing) in running multiple hydrometeorological simulations concurrently in the FLOREON$^+$ system that should predict upcoming floods and warn against them. These predictions are based on the data inputs from NWFS (numerical weather forecast systems) (e.g. ALADIN) that are then used to run the rainfall-runoff and hydrodynamic models. Preliminary results of these experiments are presented in this article.

INTRODUCTION

There are many types of natural disasters in the world. Many of which depend specifically upon geography. Floods are one of the worst and most recurrent types of natural disasters in our region (Wohl, 2000; Brázdil et al., 2005; Bedient et al., 2007).

Local governments require reliable models for flood simulations and predictions to save on ample funding that must be otherwise invested in post-flood repairs for impacted regions (Brázdil et al., 2005). Therefore, the issue of flood prediction and simulation has been selected as a case of choice for experimental development.

Modelling of the water component in the landscape together with consequent environmental aspects becomes a frequently discussed activity. Water is both element and irreplaceable resource in this context (Unucka et al., 2009).

The geographic information systems and hydrologic models became more frequently used with the development of informatics in the field of hydrology and environmental issues. These products represent the most efficient tools for the hydrologic analyses beyond any doubt. Numerous projects for the improvement of such tools were realized worldwide. Both particular hydrologic and environmental problems were solved together with design of more complex systems, in which communication of GIS and hydrologic models were solved in the system way. ArcHydro is amongst the most sophisticated projects in that way (Maidment, 2002; Bedient et al., 2007; Unucka et al., 2009).

Our project FLOREON$^+$ (FLOods REcognition on the Net) (Vondrák et al., 2008; Martinovič et al., 2008; Unucka et al., 2009) endeavours to achieve such a complex solution to environmental issues. The system is being developed as a tool for disaster management and decision support. It is a prototypal open modular system of environmental risks modelling and simulation, which is based on modern internet technologies and platform independency. Environmental problems are solved via automatic communication of environmental models and GIS, visualised and published via internet interface. The first milestone of the project was the development and operative running of complex decision support and prediction system for hydrologic problems as are the floods. The first aim was to achieve near real-time predictions of hydrographs and flood lakes, which were accomplished. The project is running in operative way nowadays and HPC capabilities are tested within the system. The final product of the project is going to be the system offering online communicational man-machine interface and providing a various types of products for decision support. The project results should help to simplify the process of crisis management and increase its operability and effectiveness.

HPC USE IN HYDROLOGY AND ENVIRONMENTAL MODELING

The speed of the computation is the main advantage of HPC. Solving the tasks occurring in the space and time (such as environmental modelling tasks), the capacity of HPC enables to accomplish the computations in the high resolution, both temporal and spatial (Kumar et al., 2008).

The combination of environmental modelling and HPC is not new. Distributed computers have the poten-

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali
David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

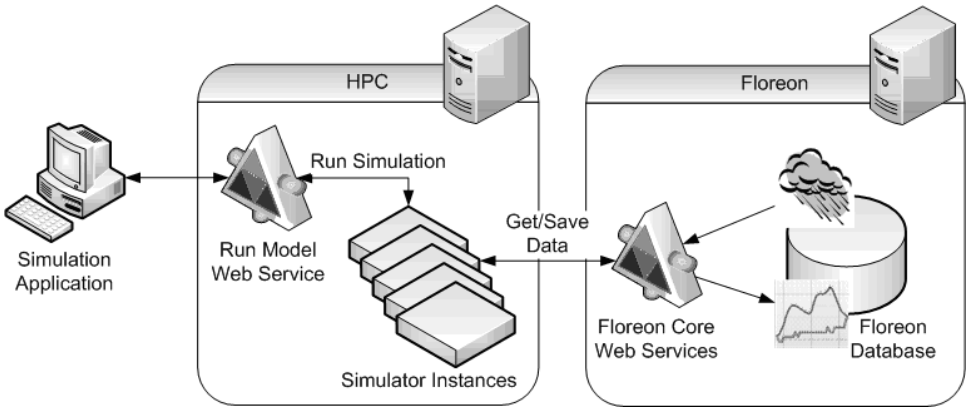


Figure 1: Running hydrological simulations on HPC in the FLOREON$^+$ system

tial to provide an enormous computational resource for solving complex environmental problems, and there is active research in this area to take better advantage of parallel computing resources (Vrugt et al., 2006). Environmental models of many kinds have been computed using HPC cluster. For example, coupled hydrogeological and biogeochemical models were computed by (Gwo et al., 2001) or surface-subsurface flow and reactive transport coupled models were computed by (Gwo and Yeh, 2004). In hydrology field, autocalibration computations were done using HPC (Cheng et al., 2005; Vrugt et al., 2006; Sharma et al., 2006; Tang et al., 2007; Liu et al., 2009), as well as multidimensional hydrodynamic or fluvial-geomorfological modelling and contaminant transport computations (Dortch and Gerald, 2001; Allen et al., 2008). In hydrogeological modelling, contaminant transport and water (including coastal ones as well) quality modelling, cluster computing was applied by (Thompson et al., 1997), (Espino et al., 1997), (Peters et al., 1997), (Wu et al., 2002) or (Hammond et al., 2005), for example. (Li et al., 2006) used HPC advantages to run fully distributed hydrological model, similarly (Lien et al., 2001, 2004) used the capabilities of (to HPC close) grid computing for the distributed solution of flood prediction. Typical area of HPC use is meteorological and climatologic modelling (Sathye et al., 1996; Bougeault, 2008). Another sphere of its use is, for example, a land surface modeling (Tian et al., 2008). Our FLOREON$^+$ system is an example of coupled rainfall-runoff and hydrodynamic models.

RUNNING HYDROLOGICAL SIMULATIONS ON HPC WITHIN THE FLOREON$^+$ SYSTEM

HPC as a parallel environment is able to run many hydrological simulations at the same time. This allows the users to use the environment effectively and it shortens waiting time for simulation results even during the high level of demand (e.g. during critical situations). Parallel computing is also very useful for model calibration, in which many simulations with different calibration parameters can be run simultaneously and their results can be compared gradually.

However, this comes with the implementation cost, because used simulation models are not ready for such simultaneous launching. We had to solve this problem by creating multiple simulation environments integrated with preparation and finalisation code. We named these functional environments *Simulators* and created one instance for each node and computation core that would be used to perform simulations.

Therefore, when a user needs to run a simulation, he uses FLOREON$^+$'s *Simulation Application* to create new simulation and fill it with desired attributes based on the model he wants to use. The Simulation Application then calls the *Run Model Web Service* deployed on the HPC server and sends all given parameters. This web service utilizes the HPC environment to find a suitable Simulator instance in the pool of available instances (see Figure 1). The chosen Simulator prepares the required model and asks *Floreon Core Web Services* for snow thickness, rainfall, temperature and other data, saved in the central *Floreon Database*. These are used as the input data to the model and the Simulator starts the simulation. Results of the simulation are sent to the Floreon Core Web Services to be saved in the Floreon Database for the future use. At the same time, the resulting hydrographs are displayed to the user in the Simulation Application and the Simulator instance is returned to the pool of available instances.

CASE STUDY

As it was mentioned, FLOREON$^+$ is a complex and modular system for hydrologic and environmental modelling. The FLOREON$^+$ system disposes partly of automated hydrometeorological data collecting, partly of automated computational cascade of event rainfall-runoff and hydrodynamic models. The hydrometeorological data are collected from network of gages run professionally by the Czech Hydrometeorological Institute (CHMI) and the Povodi Odry (the river Odra basin board) state enterprise and by methods of remote sensing (radar estimation of precipitation rates, provided by the CHMI). Together with the precipitations predicted by NWFS AL-

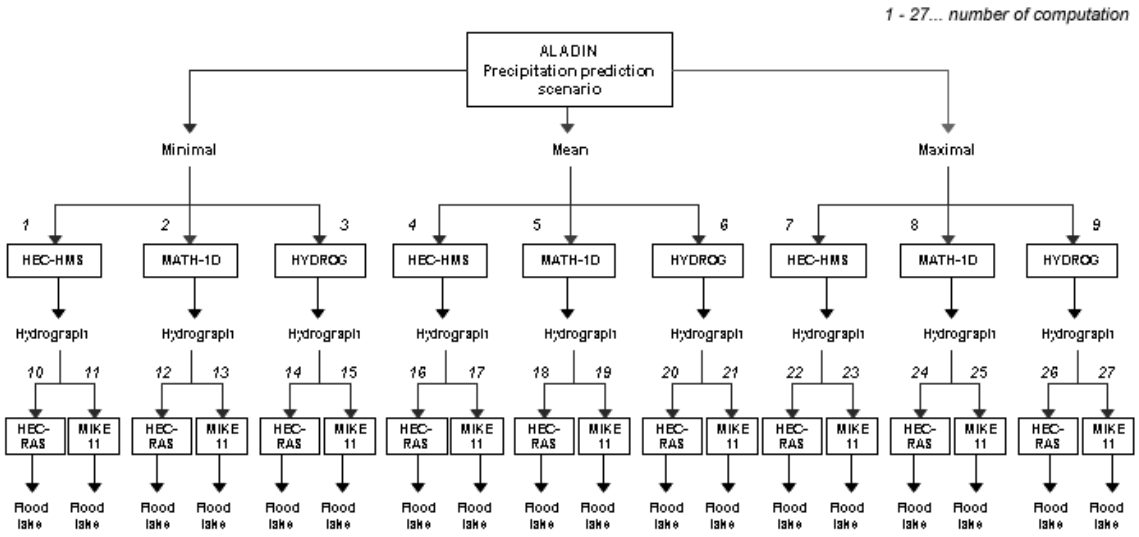


Figure 2: Computation tree for one modelled watershed/river network

ADIN model (Aire Limite, Adaptation Dynamique, Development International), which is meso-beta scale numerical meteorologic model co-developed and computed by CHMI (Zagar, 2000; Řezáčová et al., 2008), they are used as the inputs of the rainfall-runoff models.

Meteorological inputs include precipitation depth in particular, and furthermore the air temperature and the data about snow pack (thickness and water equivalent) during the winter time. The time step of the obtained precipitation depth and temperature data is 1 hour. Hydrological inputs cover data on hourly discharges and water levels from the hydrologic gages. This discussed data are used within the FLOREON$^+$ system as inputs of the runoff response calculation to causal rainfall (possibly runoff caused by snowmelt). The outputs of these models (hydrographs, also in the profiles over and above the professionally monitored gages and gages of professional hydrologic prediction) together with the observed discharges are then used as the boundary conditions inputs to the hydrodynamic models, which solve the water routing in riverbeds, possibly outside riverbeds during floods. One of their outputs is the spatial localization of potential flood lakes. The whole cascade including the input data collecting at the beginning, postprocessing of model outputs, and the web visualisation of the results at the end is constituent of fully automated server solution. The models are run every 6 hours and their prediction is set for 54 hours in advance. Nowadays, the outputs are used as a basis for the decision-making within the disaster management.

Hydrologic prediction is based on the meteorological prediction. In other words, the predicted values of some meteorological factors are used as input data for hydrologic prediction. Meteorological prediction itself is lumbered by uncertainty. This is one of the reasons why the temperature forecast is often published as four degrees wide intervals of temperatures, for example. Therefore, there is no reason why the results of hydrologic prediction should be taken at a face value, both in the aspect of timing or magnitude of predicted discharge. However, in common operational practice, approximately 20% difference of predicted and then observed discharges are taken as a satisfactory prediction. The issue of modelling uncertainties are generally widely discussed in the literature (Beven and Binley, 1992; Anderson and Bates, 2001; Aronica et al., 2002; Ajmi et al., 2007; Beven, 2008; Arnold et al., 2009; Geza et al., 2009; Mishra, 2009; Todini, 2009); and many others.

The important fact is that the predicted meteorological input data coming from the ALADIN NWFS model are not single numbers but intervals (ranges of possible values). Because of these input data uncertainties, it is appropriate to calculate the hydrological prediction via utilization of various precipitation forecast scenarios or ensemble model runs (e.g. combination of various rainfall-runoff models with various methods and parameters). These ensembles are minimal, maximal and mean scenarios of rainfall predictions computations in various types of rainfall-runoff models sense.

Our rainfall-runoff models are computed for 4 watersheds with an approximate area of 1000 km^2 and an average number of 49 subbasins. For each of these watersheds, 3 rainfall-runoff models are computed simultaneously, HEC-HMS (HEC-USACE, 2010), HYDROG (HySoft, 2010), and our own MATH-1D model (Kubíček and Kozubek, 2008), specifically. HEC-HMS (Hydrologic Engineering Center - Hydrologic Modelling System) is a product of HEC-USACE (Hydrologic Engineering Centre - U.S. Army Corps of Engineers). It is a semidistributed rainfall-runoff model working with various advanced methods for hydrologic and hydraulic transformation of rainfall-runoff process. HYDROG

is a semidistributed rainfall-runoff model developed by HySoft company. It is currently operatively run by the CHMI. The model is designed for simulation, operative hydrologic forecast and operative management of watershed system. Besides the HEC-HMS model, it works with the Hortons method for overland flow. MATH-1D model is own product of the FLOREON$^+$ solutions team. It is a rainfall-runoff model accumulating contribution of interflow modeled by convolution integral and simplified differential equation and contribution of surface runoff using non-stationary linear and non-linear isochrones. Computation results of these rainfall-runoff models then serve as the boundary conditions inputs to hydrodynamic models.

For each watershed, the river network of the watershed respectively, two hydrodynamic models are computed, HEC-RAS (HEC-USACE, 2010) and MIKE 11 (DHI, 2010) specifically. As well as HEC-HMS, the model HEC-RAS (Hydrologic Engineering Center - River Analysis System) is a product of HEC-USACE. It is a one-dimensional hydrodynamic model applicable for both steady and unsteady flow computations and computations of subcritical and supercritical flow in the network of natural and constructed channels including technical objects as well. The basic mathematical apparatus are the Bernoulli's and Manning's equations. The computation of flood lakes is fully supported. HEC-RAS and HEC-HMS models are free software products downloadable from the HEC-USACE website. MIKE 11 is a commercial product of DHI (Danish Hydraulic Institute). It is a one-dimensional hydrodynamic model enabling simulation of both steady and unsteady flow computations and computations of subcritical and supercritical flow in both natural and constructed channel network including technical objects as well. In the inundation areas model is able to simulate quasi 2D flow. Its basic mathematical apparatus are the Bernoulli's and energy loss equations together with the Manning's and Chezy's equations. Both models, HEC-RAS and MIKE 11, are industry standards for hydrodynamic simulations. MIKE 11 model is, as well as MIKE SHE model (DHI, 2010) (see bellow), a part of the hydrologic and hydrodynamic models platform MIKE Zero (DHI, 2010).

Because of the ensemble hydrographs (Figure 3) resulting from the rainfall-runoff models computations, the ensemble computation of hydrodynamic models are desirable. It is exacerbated by the fact that the FLOREON+ system is intended for decision making support within operational disaster management.

Since there is quite a big number of computation operations needed in order to compute the whole cascade of models considering the rainfall inputs ensembles (see the Figure 2), HPC capabilities offer a significant increase of computation speed, which is very important in operational practice, especially during the critical events. There is a computation tree for one modelled watershed in the Figure 2. In fact, due to 4 modelled watersheds, the number of computations is 4-times bigger for the complete area covered by FLOREON$^+$ system, if we do not consider different hydrologic computation methods and parameters ensembles.

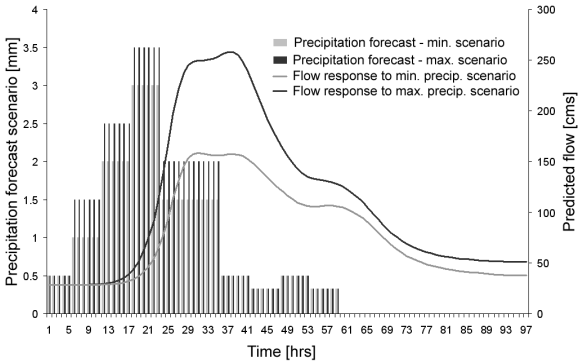


Figure 3: Rainfall-runoff prediction scenarios

Resulting flood lakes are processed by GIS tools and archived with other output data (hydrographs, water levels etc.). The minimal and maximal extents of possible flood lakes (see the Figure 4) and hydrographs scenarios are provided to authorities of crisis management as a support for decision-making and coordination of possible actions. Every other scenarios and ensemble computation outputs are stored and ready to be used in the case of need.

DISCUSSION

The use of the HPC cluster for the computations of FLOREON$^+$ system tasks seems to be necessary for the future. Only in terms of hydrologic modelling, the parallel scenarios computations based on the ensembles of meteorological forecasts will be more demanding on computer performance during a snowmelt episodes. Snowmelt runoff modelling takes into account next to the precipitation air temperatures and snow water equivalent as well. So, at least one more ensemble forecast of meteorological element - temperature - will need to be considered.

Further to the above mentioned, another field of hydrologic modelling that is going to be tested using HPC under the FLOREON$^+$ system is the automatic calibration of models. Since our models are event ones (not continuous for long-term simulations) running on the hourly inputs, the calibration runs automatically started before each model simulation/prediction run are going to be tested. We are aware of the fact that for successful operational deployment of automatically calibrated event models, such a calibration procedures have to be properly tested separately for every single modelled watershed. Parameterisation of the hydrologic model to fit best to actual conditions of the real watershed is difficult, especially in the case of larger and heterogeneous watersheds as well as in routine operation with automatically run event models and automatic continual input data download. Spatial and temporal distribution of rainfall input

Figure 4: Simulated flood extents for variuos precipitation forecast scenarios

is often quite unpredictable and leads to a large heterogeneity in catchment conditions (e.g. saturation of soils), both spatial and temporal. The values of optimized parameters have to fall within the interval of realistic conditions. The automatic optimization can results in a good fit of observed and simulated discharges at the expense of unrealistic parameter values. In operative it could lead to the useless results of the prediction, if the external conditions would suddenly change. Nowadays, we automatically optimise the value of initial abstraction using API (antecedent precipitation index).

Another step of FLOREON$^+$ system development is going to be a fully distributed hydrologic model MIKE SHE deployment. Some hydrologic issues such as runoff response to convective rainstorms or hydrologic balance computations in high resolution have to be solved by this kind of models. Despite of all the quick development in the field of IT, it is still very time-consuming to run distributed models for a larger watershed. HPC thus offers the way of operational running of these models.

In terms of another purpose and use of FLOREON$^+$ system it is good to mention that the system is based on client-server architecture. The three levels of users (divided according to their expertise) will be allowed to run the simulation with different rights of setting the models. So that all computing operations will be performed on HPC servers, by which users are made available only in the form of the award and subsequent results. Using high performance computing brings the possibility to satisfy a big amount of users in real and near-real time (Unucka et al., 2009).

ACKNOWLEDGEMENT

We acknowledge the support of projects SP/2010196 Machine Intelligence and SP/2010192 Possibilities of modelling of natural and environmental risks caused by extreme hydrometeorological situations.

REFERENCES

Ajmi, N.K., Duan, Q.Y., and Sorooshian, S. (2007). An integrated hydrologic Bayesan multimodel combination framework: confronting input, parameter, and model structural uncertainty in hydrologic prediction. In Water Resources Research. 43 (1), W01403.

Allen, J.B., Smith, D.L., Eslinger, O.J., and Valenciano, M.A. (2008). A New Approach to Streambed Modelling and Simulation Using Computational Fluid Dynamics. In 2008 Proceedings of the Department of Defence High Performance Computing Modernization Program: User Group Conference - Solving the Hard Problems. Seattle, WA, 3-8.

Anderson, M.G. and Bates, P.D. (eds.). (2001). Model Validation. Perspective in Hydrological Science. John Wiley & Sons.

Arnold, S., Attingr, S., Frank, K., and Hildebrandt, A. (2009). Uncertainty in parameterisation and model structure affect simulation results in coupled ecohydrological models. In Hydrology and Earth System Sciences. 13 (10), 1787-1807.

Aronica, G., Bates, P.D., and Horrit, M.S. (2002). Assessing the uncertainty in distributed models predictions using observed binary pattern information within GLUE. In Hydrological processes. 16 (10), 2001-2016.

Bedient, P.B. and Huber, W.C. and Vieux, B.C. (2007). Hydrology and floodplain analysis. 4th edition. Prentice Hall, London.

Beven, K.J. (2008). Environmental Modelling: An Uncertain Future? Routledge, London.

Beven, K.J. and Binley A. (1992). The future of distributed models - model calibration and uncertainty prediction. In Hydrological processes. 6 (3), 279-298.

Bougeault, P. (2008). High performance computing and the progress of weather and climate forcasting. In 8th International Conference on High Performance Computing for Computational Science, VECPAR 2008. Toulouse, France.

Brázdil, R. et al. (2005). Historické a současné povodně v České republice. (in Czech). Brno-Praha. MU Brno a ČHMÚ, 369 p.

Cheng, CT, Wu, XY, and Chau, K.W. (2005). Multiple criteria rainfall-runoff model calibration using a parallel genetic algorithm in a cluster of computers. In Hydrological Science Journal. 50(6) December 2005, 1069-1087.

DHI, (2010). DHI - Danish Hydraulic Institute.
URL: http://www.dhigroup.com (13th April 2010)

Dortch, M. and Gerald, T. (2001). Comparison of HPC methods for long-term contaminant modeling. In Estuarine and Costal Modeling: Proceedings of the Seventh International Conference. St. Petersburg, FL, 535-546.

Espino, M., Gonzalez, M., Hermosilla, F., Uriarte, A., Chumbe, S., Garcia, M.A., Borja, A., and Arcilla, A.S. (1997). HPC simulation of pollution events at the San Sebastian Coast (N Spain). In International Conference on Measurements and Modelling in Environmental Pollution, MMEP, Proceedings. Madrid, Spain, 1-12.

Geza, M., Poeter, E.P., and McCray, J.E. (2009). Quantifying predictive uncertainty for mountain-watershed model. In Journal of Hydrology. 376 (1-2), 170-181.

Gwo, J.P., D'Azevedo, E.F., Frenzel, H., Mayes, M., Yeh, GT., Jardine, P.M., Salvage, K.M., and Hoffman, F.M. (2001). HGBC123D: A High-performance Computer Model of Coupled Hydrogeological and Biogeochemical Processes. In Computer & Geosciences. No. 27, 1231-1242.

Gwo, J.P. and Yeh, GT. (2004). High-performance simulation of surface-subsurface coupled flow a reactive transport at watershed scale. In 1st Internationa Conference on Computational Methods (ICCM04). Singapore, Singapore.

Hammond, G.E., Valocchi, A.J., and Lichtner, P.C., (2005). Application of Jacobian-free Newton-Krylov with physics-based preconditioning to biogeochemical transport. In Advances in Water Resources 28, 359-376.

HEC-USACE, (2010). Hydrologic Engineering Center - US Army Corps of Engineers.
URL: http://www.hec.usace.army.mil/ (13th April 2010)

HySoft, (2010). HySoft.
URL: http://www.hysoft.cz/ (13th April 2010)

Kubíček, P. and Kozubek, T., (2008). Mathematic-analytical Solutions of the Flood Wave and its Use in Practice (in Czech) VŠB-TU Ostrava, 150 p.

Kumar, S., Peters-Lidard C., Tian, Y., Riechle, R., Gieger, J., and Alonge, C. (2008). An Integrated Hydrologic Modelling and Data Assimilation Framework. In Computer. Volume 41, Issue 12, 52-59.

Li, TJ., Liu, JH., He, Y., and Wang, GQ. (2006). Application of cluster computing in digital watershed model. In Advances in Water Sceince. Volume 17, Issue 6, November 2006, 841-846.

Lien, HC., Lee, LC., Shiau, YH., Shen, CY., Shih, RJ., Tsai, WF., and Lyu, P. (2001). An integration of Web-based GUI and flow model for a hydroinformatics system. In The Proceeding of the 5th International Confernce and Exhibition on High Performance Computing in the Asia-Pacific Region, HPC Asia 2001. Gold Coast, Queensland, Australia.

Lien, HC., Shiau, YH., Huang, CP., Wu, JH., and Tsai, WF. (2004). The integration and application of computational grid structure in flood forecast system. In Proceedings - Seventh International Conference on High Performance Computing and Grid in Asia Pacific Region, HPCAsia 2004. Tokyo, Japan, 328-331.

Liu, Y., Freer, J., Beven, K., and Matgen, P. (2009). Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error. In Journal of Hydrology. 367 (2009), 93-103.

Maidment, D. R. (ed.) (2002). ArcHydro. GIS for Water Resources. Redlands, ESRI Press.

Martinovič, J., Štolfa, S., Kožusnik, J., Unucka, J., and Vodnrák, I. (2008). FLOREON - the system for an emergent flood prediction. In ECEC-FUBUTEC-EUROMEDIA. Porto, Portugal.

Mishra, S. (2009). Uncertainty and sensitivity analysis techniques for hydrologic modeling. In Journal of Hydroinformatics. 11 (3-4), 282-296.

Peters, J.F., Howington, S.E., Holland, J.P., Tracy, F.T., and Maier, R.S. (1997). High performance computing as a tool for groundwater cleanup. In Journal of Hydraulic Research. Volume 36, Issue 6, 1997, 897-912.

Řezáčová, D. et al. (2008). Fyzika oblaků a srážek (in Czech). Academia, Praha.

Sathye, A., Bassett, G., Droegmeier, K., Xue, M., and Brewster, K. (1996). Experiences using high performance computing for operational storm scale weather prediction. In Concurrency Practice and Experience. Volume 8, Issue 10, December 1996, 731-740.

Sharma, V., Swayne, D.A., Lam, D., and Schertzer, W. (2006). Parallel Shuffled Complex Evolution Algorithm for Calibration of Hydrological Models. In High Performance Computing Symposium. St. John's, New Foundland.

Tang, Y., Reed, P.M., and Kollat, J.B. (2007). Parallelization strategies for rapid and robust evolutionary multiobjective optimization in water resources applications. In Advance in Water Resources. 30 (2007), 335-353.

Thompson, A.F.B., Rosenberg, N.D., Bosl, W.J., Falgout, R.D., Smith, S.G., Shumaker, D.E., and Ashby, S.F. (1997). On the use of high-performance simulation in the management of groundwater resources in large aquifer systems. In Proceedings of the 1997 27th Congress of the International Association of Hydraulic Research, IAHR. San Francisco, CA, USA, 337-342.

Tian, Y., Peters-Lidard, C.D., Kumar, S.V., Geiger, J., Houser, P.R., Eastman, J.L., Dirmeyer, P., Doty, B., and Adams, J. (2008). High-performance land surface modeling with a Linux cluster. In Computer & Geosciences. No 34 (11): 1492-1504 NOV 2008.

Todini, E. (2009). Uncertainties in environmental modelling and consequences for policz making. In NATO Science for Peace and Security. Series C - Environmental Security. Springer, 205-228.

Unucka, J., Martinovič, J., Vondrák, I., and Rapant, P. (2009). Overview of the complex and modular system FLOREON+ for hydrologic and environmental modeling. In River Basin Management V. C.A. Brebbia (Ed.), WIT Press, Southampton, UK.

Vrugt, J.A, Ó Nualláin, B., Robinson, B.A., Bouten, W., Dekker, S.C., and Sloot, P.M.A. (2006). Application of parallel computing to stochastic parameter estimation in environmental models. In Computers and Geosciences. 32 (8) 2006.

Vondrák, I., Martinovič, J., Kožusznik, J., Štolfa, S., Kozubek, T., Kubíček, P., Vondrák, V. , and Unucka, J., (2008). A description of a highly modular system for the emergent ood prediction. In Computer Information Systems and Industrial Management Applications, CISIM '08. 7th, pages 219-224.

Wohl, E.W. (Ed.). (2000). Inland Flood Hazards. Cambridge,Cambridge University Press.

Wu, Y.S., Zhang, K., Ding, C., Pruess, K., Elmroth, E., and Bodvarsson, G.S., (2002). An efficient parallel-computing method for modeling nonisothermal multiphase flow and multicomponent transport in porous and fractured media. In Advances in Water Resources 25, 243-261.

Zagar, M. (2000). Forecasting of the High Resolution Wind and Precipitation with Dynamic Adaptation. Ph.D. thesis. University of Ljubljana & Université Paul Sabatier (Toulouse).

AUTHOR BIOGRAPHIES

JAN MARTINOVIČ studied computer science at the VŠB - Technical University of Ostrava, the Czech Republic and obtained M.S.E. degree in 2004 and Ph.D. degree in 2008. Nowadays, he works as an assistant professor at the same university. He is a researcher dealing with data mining and data processing, with emphasis to web search, data compression and social networks.

JAN UNUCKA studied physical geography and geoecology at the University of Ostrava and obtained M.Sc. degree in 2003 and Ph.D. degree in 2006. For a few years he worked for the Czech Hydrometeorological Institute as a worker of regional hydrological forecast office. Since 2003 he has been working as an assistant professor at the VŠB - Technical University of Ostrava. The main field of his professional interest is environmental modelling and linking GIS with environmental models.

BORIS ŠÍR studied physical geography and geoecology at the University of Ostrava, the Czech Republic and obtained MSc. degree in 2007. In 2008, he obtained M.S.E. degree in environmental engineering at VŠB - Technical University of Ostrava. Nowadays, he is a doctoral student of geoinformatics at VŠB - Technical University.

ŠTĚPÁN KUCHAŘ studied computer science at the VŠB - Technical University of Ostrava and obtained M.S.E. degree in 2007. Nowadays, he is a postgraduate student of computer science and applied mathematics at VŠB - Technical University of Ostrava.

IVO VONDRÁK studied computer science at the VŠB - Technical University of Ostrava and obtained M.S.E. degree in 1983 and Ph.D. degree in 1988. He has been working as assistant professor at the VŠB - Technical University of Ostrava since 1989, since 1991 as associated professor and since 1998 as a professor of engineering informatics. Between 1992 and 2002 he was visiting professor at the Montanuniversitat Leoben, Austria, and in the years 1996 and 1997 he worked as a researcher for Software Research Labs, Texas Instruments, Inc., Dallas, TX, USA. Nowadays, he is the president of the VŠB - Technical University of Ostrava.

VÍT VONDRÁK studied mathematics at the Palacký University in Olomouc, the Czech Republic and obtained M.Sc. degree in 1992 and Ph.D. degree in 2003. Since 1992 he has been working as an assistant professor and since 2007 as an associated professor at the VŠB - Technical University of Ostrava. For a couple of years, he was working for Institute of Thermomechanics of Academy of Sciences of the Czech Republic.

Scientific approaches for the industrial workstations ergonomic design: a review

Terry Bossomaier[(a)], Agostino Bruzzone[(b)], Antonio Cimino[(c)], Francesco Longo[(c)], Giovanni Mirabelli[(c)]

[(a)]Charles Sturt University, Dept of Computer Science, Australia
(b) MISS-DIPTEM, University of Genoa, Italy
[(c)]MSC-LES, Mechanical Department, University of Calabria, Italy

KEYWORDS

Industrial workstations, industrial design, modeling & simulation, computerized models

ABSTRACT

Over the last years ergonomic problems have received growing attention due to their effects on industrial plants efficiency and productivity. Many theories, principles, methods and data relevant to the workstation design have been generated through ergonomics research. However, no general frameworks have been suggested, yet. The time seems to be right for presenting a review paper on the scientific studies whose aim is to achieve the ergonomic design of industrial workstations. To this end, it is the intent of the authors to provide the readers with an accurate overview on the main scientific approaches proposed (during the last two decades) by researchers and scientists working in this specific area. In particular, two main scientific approaches have been identified. The first approach is based on the direct analysis of the real workstations, while the second one uses computerized models to design workstations ergonomically. Each scientific approach will be presented through a detailed description of the research works it involves. The initial search identifies a huge number of articles which were reduced to about 60 studies based on content and quality. Note that the research works description represents the core part of this literature review.

INTRODUCTION

Ergonomics is the application of scientific principles, methods, and data drawn from a variety of disciplines for the development of engineering systems in which people play a significant role (Kroemer 1994). The Institute for Occupational Ergonomics (1999) defines ergonomics as an understanding of the needs, limitations, and abilities of people, and the use of this understanding for the design of products and environments in which people live. Over the last two decades, ergonomics researchers and practitioners have devoted considerable resources to solve the problems associated with the ergonomic design of the working environment. Designers of workplaces have usually three major tasks: one, integrating information about processes, tools, machines, parts, tasks, and human operators; two, satisfying design constraints which often conflict; and three, generating a design acceptable to all parties involved. However, while completing these tasks, designers often have difficulty incorporating ergonomics information about the human operators into their designs. Note that, although today the tasks or processes are being mechanized or automated as the technology has advanced, many tasks are still performed manually in several industrial settings (Chung and Kee 2000). In this context, it seems to be clear that matching the abilities of the operator with the task requirements as well as with working environment physical constraints are important aspects to be faced within the ergonomic workplace design. Over the years many theories, principles, method and data relevant to the workstation design have been generated through ergonomics research. However, no general frameworks have been suggested, yet. It is the intent of the authors to provide the readers with an accurate overview on the main scientific approaches proposed (during the last two decades) by researchers and scientists for achieving the ergonomic design of industrial workstations. In particular, two main scientific approaches have been identified. The first approach is based on the direct analysis of the real workstations, while the second one use computerized models to design workstations ergonomically. Each scientific approach will be presented through a detailed description of the research works it involves. Note that the research works description represents the core part of this literature review.

The paper is structured as follows. Firstly, the international state of the art on the scientific approaches addressing the ergonomic workstation design is proposed. Note that the state of the art description is splitted into two different subsections: the first subsection presents the research works related to the approach based on the direct analysis of the real workstations, while the second one describes the research studies that use computerized models to design workstations ergonomically. Then, some gaps are identified and ongoing research solutions are presented together with some application examples in different manufacturing areas. Finally, a conclusion is drawn.

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali
David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

STATE OF THE ART OVERVIEW

An ergonomic approach for the design of industrial workstations is the attempt to achieve an appropriate balance between the worker's capabilities and worker's requirements as well as provide the worker with physical and mental well-being, job satisfaction and safety (Das and Sengupta 1996). In order to help industrial engineers to achieve effective workplace design, a number of research works have been developed. This section presents an overview of this studies by organizing them into two approaches: the first approach is based on the direct analysis of the real workstations, while the second one use computerized models to design workstations ergonomically. Such studies have been identified by means of Google Scholar, Scopus, and Scirus as research engines. A first run has been made by typing (sometimes alone, some other times combined) the keywords "industrial workstations", "industrial design", "ergonomic effective design", "workstation design", "ergonomic standards", "modelling & simulation", and "computerized models". At a second stage, the abstracts of the peer-reviewed and of the specific-interest-groups publications outcomes have been read, to evaluate the actual pertinence of the abstracts to the research issue. At a later stage, starting from the evaluation of the abstracts only the more relevant papers have been read and classified/considered in the literature review.

Ergonomic design by workstation direct analysis

Here the scientific approach based on the direct analysis of the real workstations is presented by a detailed description of the research studies proposed over the last two decades.

A number of studies in literature try to achieve the ergonomic effective design of manufacturing system workstations by using based observation methods for collecting data, i.e. observation of the worker performing the manufacturing operations is used for collecting information about the work methods. In order to achieve the ergonomic effective design of the manufacturing system workstations, such research works analyze the videotape of the work methods and assume a trial and error methodology (in effect the design methodology is never supported by a well-defined experimental design). The final ergonomic design of the workstations depends on researcher's experience and his/her knowledge about the manufacturing system. Das and Sengupta (1996) provide the conceptual basis for a good workstation design by presenting a systematic ergonomic approach capable of determining the workstation dimensions and layout. The workstation design procedure starts off with the collection of the workstation relevant data through direct observation and videotaping and ends up with constructing a prototype workstation based on the final design. Moreover the authors apply the systematic ergonomic approach to the design of a supermarket checkstand workstation. Kadefors and Forsman (2000) present a method for ergonomic evaluation of complex manual work based on interactive operator assessment of video recordings. The video recordings are displayed on a computer terminal, and the video recorded operators assess the work by clicking on virtual controls on the screen, whenever a situation inducing pain or discomfort appears. The application of the method to a workshop belonging to Volvo Cars shows it is easy to understand and operate by practitioners as well as it provides structured information on the high priority tasks that are relevant and useful for instance in industrial interventions and industrial workstation design. Neumann et al. (2001) identify the trunk position and movement velocity as important parameters to be considered and measured in industrial settings design. To this end, the authors present a video-based posture assessment method capable of measuring trunk angles and angular velocities in industrial workplaces. The video analysis workstation consists of a desktop computer equipped with digital video capture and playback technology, a VCR, and a computer game type joystick. An application example confirms the importance of these factors and demonstrates the utility of a video based method to measure them. Forsman et al. (2002) present a method to amalgamate technical and human aspects in the industrial workstation design. The technical aspects are represented by results from a computer- and video based observation method for time data collection. The human aspects comprised physiological measurements of muscular activity, and of body postures and movements. The integrated procedure allows work activities to be assigned significantly different levels of physical work load. These different levels may be used to predict physical work load in the design and change of production systems.

The possibility to integrate based observation methods and specific ergonomic standards was also investigated by researchers. Even in this case the based observation methods are used as data collection tools, while the ergonomic standards allow the researchers to investigate and analyze the ergonomics of the workplace. Among the ergonomic standards, the following have to be regarded as the most widely used:

- the NIOSH 81 and the NIOSH 91 equations for lifting tasks (NIOSH stands for National Institute for Occupational Safety and Health)
- the OWAS for analysing working postures (OWAS stands for Ovako Working Analysis System)
- the Burandt–Schultetus analysis for lifting tasks involving a large number of muscles
- the Garg analysis for assessing the energy expenditure (EE) for performing an operation

Further information about the cited ergonomic standards can be found in Garg (1976), Schultetus

(1980), Niosh Technical Report 81-122 (1981), Karhu et al. (1981), the Scientific Support Documentation for the Revised 1991 NIOSH Lifting Equation (1991), Waters et al. (1994).

Scott and Lambe (1996) implement the OWAS in a perchery system. The workers have been video recorded performing normal duties within the perchery and the positions of the body have been assessed using the OWAS analysis. Several wrong working postures have been identified and suggestions, in light of the OWAS results, have been proposed for an improved perchery design. Engström and Medbo (1997) develop a procedure for the workstation design that integrates prototype equipment used for workstation data collection and ergonomic analysis for working postures evaluation (OWAS). The prototype equipment consists of a video camera, a video tape recorder, a TV monitor and a personal computer. The OWAS technique carries out a qualitative analysis of the worker's movements during a working process and calculates the stress associated to each body posture. Note that the procedure promotes a design based on empirical data considering ergonomic aspects and work performance analysis. Vedder (1998) presents an easy-to-use video-based posture analysis method for workplaces where task interference has to be minimized and postures have to be observed over a longer period of time. The different worker postures have been video recorded by using a stationary camera and then evaluated by using the OWAS posture analysis system. Such method allows to identify hazardous postures and their causative factors so that appropriate re-design measures can be taken. Herman et al. (1999) propose a practical methodology to analyze the influence of material handling devices on the physical load during the end assembly of cars. First the worker under observation describes the manual actions in detail while performing the task and explains why he does or does not use the tool (important information is recorded in writing), then the NIOSH method (1991) is used to analyze the lifting and lowering aspects of each task, taking into account several distance measurements and the lifting frequency. Finally the authors, according to objective and subjective results of the data analysis, propose several recommendations to the company regarding the use of existing tools for the end assembly of cars. Shuval and Donchin (2005) examine the relationship between ergonomic risk factors and upper extremity musculoskeletal symptoms at a Hi-Tech company in Israel. Ergonomic risk factors were assessed through direct observation of employees' postures at their workstations using the rapid upper limb assessment (RULA) tool. Results of the RULA observations indicate excessive postural loading with no employee in acceptable postures so that the authors point out the need for implementing an intervention program. Lin and Chan (2007) evaluate the effect of ergonomic workstation design on musculoskeletal risk factors (MRFs) and musculoskeletal symptoms (MSSs) reduction among female semiconductor fabrication room (fab) worker. By means of walk-through observations of the working environment, discussing with company's managers and using NIOSH analysis, the authors identify the most prevalent and urgent ergonomic issues to be resolved and modify the layout of the workplace for reducing ergonomic hazards.

In addition to the previous studies, the following use interviews as support tools for the workstation data collection. Grant et al. (1995) describes an investigation conducted to identify and evaluate possible causes of back and lower extremity pain among the workers of a day care facility. The investigation is based on the use of questionnaire, video tape systems and NIOSH lifting equations. Questionnaire results indicated that back/pain discomfort was a common musculoskeletal complaint. Observation and analysis of the work activities indicated that employees spend significant periods of time kneeling, sitting on the floor, squatting, or bending at the waist. The revised NIOSH lifting equation indicated that several employed performing lifting tasks may be at increased risk of low back pain and lower extremity injury. Finally the authors present recommendations for reducing or eliminating these risks by modifying the workplace and changing the organization and methods of work. Grant et al. (1997) analyze lifting tasks at a cabinet company. Workers interviews have been used to assess the magnitude of the musculoskeletal problems. Videotape systems have been used for observing material handling activities and finally the revised NIOSH lifting equation has been used for analyzing representative lift tasks. The research study identifies several lifting hazards and specific recommendations for reducing physical workload have been suggested. Chung and Kee (2000) propose a procedure based on the use of a questionnaire survey as well as the 1991 revised NIOSH lifting equations for the evaluation of lifting tasks frequently performed during fire brick manufacturing processes. A questionnaire survey shows that weight of the load significantly influence the incidence of back injuries. The NISOH lifting equation identifies risk factors that may cause musculoskeletal disorders among the operators. The research results suggest that several tasks should be redesigned ergonomically simply by making horizontal locations closer to a worker or by reducing the asymmetric angles. White and Kirby (2003) propose an ergonomic evaluation of health-care workers in a rehabilitation center. The authors present a procedure based on the integration of questionnaire, video tape systems and OWAS analysis. Workers completed a brief questionnaire that elicited information on the subject's age, gender and occupation. The videotape system was used to ensure relevant qualitative data, as well as providing data that could be coded and scored. The OWAS analysis was used for identifying the

wrong working postures. The research study reveals that health-care workers use a variety of methods, many of which include bent and twisted back postures that may carry a risk of injury. Note that the authors do not provide any information concerning the improvement of the operators work methods.

The literature overview analysis reveals that the majority of the studies developed for facing the ergonomic effective design problem are based on the single use of specific ergonomic standards. Examples of such research works are proposed on the sequel as they run through the literature. Carrasco et al. (1995) use the OWAS analysis for evaluating three different designs of checkout workstation, which require the operator to stand when they scan the products, pack them into the plastic bags and transfer the packed bags to the customer. The evaluation points out significant musculoskeletal load and exertion associated with the different checkouts and several suggestions have been presented for an improved workstation design in terms of postural load reduction and productivity increase as well. Van Wendel de Joode et al. (1996) use the OWAS analysis in order to quantify workers physical load within two ship maintenance companies. Postural load was measured and awkward postures were identified affecting workers back, neck/shoulder and arms. On the light of such results, the authors reduced workers physical load by proposing several technical adaptations and applications as well as by enlarging task rotation. Temple and Adams (2000) use the NIOSH analysis in order to establish ergonomic acceptable limits for an industrial lifting station. Through the analysis of several factors the authors define a cumulative lifting index and use such index for detecting ergonomic problems during lifting tasks. They successively modify the lifting station for reducing ergonomic risks and preventing lower back related injuries. González et al. (2003) apply the RULA method for the ergonomic evaluation of industrial workplace. The authors propose a methodology that consists of three steps: the first includes the selection of the profile of the firm to study, while the second and the third will, respectively, consist in the choice of the workplace and the gathering and treatment of the representative data of the levels of ergonomics and quality. Having identified the ergonomic problems (by using the RULA method) within a metalworking firm, a series of improvements were then implemented, analyzing whether significant alterations in quality levels took place in parallel as a result of these ergonomic improvements. Massaccessi et al. (2003) investigate work-related disorders in truck drivers using the RULA method. Such method allowed to perform a rapid and correct evaluation of the loading to which neck and trunk are exposed while driving. RULA evidences that the posture adopted in street washing trucks during cleaning operations was associated with a major risk for back pain, especially with non-adjustable seats. On the light of the analysis results, the authors recommend ergonomic interventions aiming at modifying the truck's workstation with a view to prevent musculo-skeletal disorders. Choobineh et al. (2004) use the RULA technique for carrying out ergonomic intervention in carpet mending operation. The authors identify several ergonomic problems affecting workers knees, back and shoulders and propose a new workstation configuration improving working postures noticeably. Mäkelä and Hentilä (2005) estimate the physical workload and strain of dairy farming in loose housing barns. The authors use the OWAS analysis for evaluating workers postures during the feeding and removing manure and spreading of bedding activities. On the basis of the OWAS results, the authors provide some recommendations for building new loose-housing barns providing enough space for automated feeding and cleaning systems.

In order to achieve achieving multiple and simultaneous ergonomic improvements, several researchers propose the integration of two or more ergonomic standards. Jones et al. (2005) use the RULA method and the NIOSH lifting equation in order to examine three common pub occupations (bartending, waitressing and cooking) with the aim of determining the biomechanical loads of job tasks, assessing the potential risk of musculoskeletal injury in these job tasks, and recommending injury prevention measures. Jones and Kumar (2007) compare the results of 5 ergonomic risk assessment methods (RULA, REBA, ACGIH TLV, Strain Index and OCRA) in a repetitive high-risk sawmill occupation, examine the effect of multiple definitions of the posture and exertion variable on the risk assessment methods, describe the variability in risk assessment scores between workers, examine the ability of risk assessment component scores to differentiate between facilities with significantly different levels of exposure, and examine the association between risk output and recorded incidence rates. Russell et al. (2007) compare the results of different ergonomic standards (NIOSH, ACGIH TLV, Snook, 3DSSPP and WA L&I) for evaluating ergonomic risks in lifting operations. Each ergonomic standard is applied to a uniform task (lifting and lowering two different types of cases) with the aim of choosing the best work methods by appropriately interpreting the results of the ergonomic analysis.

Finally, another important issue to take into consideration is the relation between the concepts of work measurement and ergonomics. In effect the work measurement and the ergonomics affect each other: ergonomic interventions affect the time required for performing the operations as well as any change to the work method affects the ergonomics of the workplace. Such relation was investigated by some researchers. Resnick and Zanotti (1997) underline that ergonomic principles can potentially be used to improve productivity as well. The authors propose an application example for remarking that a workstation

can be designed to maximize performance and reduce costs by considering both ergonomics and productivity together. Laring et al. (2002) develop an ergonomic complement to a modern MTM system called SAM that gives the production engineer a first insight into the future ergonomic quality of a planned production. In particular, the authors propose a tool that gives the possibility to estimate simultaneously the consumption of time in the envisaged production and the biomechanical load inherent in the planned tasks. The method was tested at the Torslanda final assembly plant of Volvo Car Corporation and at the ITT Flygt plant. The results show that the method identifies the events causing high biomechanical load on the operator so that they can be redesigned.

Ergonomic design by computerized models

In the past, workplace ergonomic considerations have often been reactive, time-consuming, incomplete, sporadic, and difficult. Usually the analysis of the real workstations is quite expensive (in terms of money and time) because it requires to ‘‘disturb’’ processes and activities of the manufacturing system. There are now emerging technologies supporting simulation-based engineering to address this in a proactive manner. These allow the workplaces and the tasks to be simulated even before the facilities are physically in place. Over the years, researchers propose several computer aided methodologies to face the ergonomic effective design problem within industrial workstations. A number of such existing research works considers virtual environment (VE) as a potential tool to support the ergonomic design of workstations. Wilson (1997) proposes an overview on attributes and capabilities of virtual environments (devoted to support ergonomic design) and describes a framework for their specification, development and evaluation. According to the author, virtual environment has potential as a tool to support many types of ergonomics contribution, including assessments of office and workplace layouts giving egocentric viewpoints for testing consequences for reaching and accessing, reconfiguring and testing alternative interface designs, training for industrial and commercial tasks, and teaching in special needs or general sectors. Jayaram et al. (2006) propose two distinct approaches to link virtual environments (VE) and quantitative ergonomic analysis tools in real time for occupational ergonomic studies. The first approach aims at creating methods to integrate the VE with commercially available ergonomic analysis tools for a synergistic use of functionalities and capabilities. The second approach aims at creating a built-in ergonomic analysis module in the VE. The authors present the two integration strategies and test them using case studies conducted with real industrial company. Chang and Wang (2007) propose a method of conducting workplace ergonomic evaluations and re-design in a digital environment for the prevention of work-related musculoskeletal disorders. First, the real workplace and human task can be converted into the digital environment through building digital mock-ups and using a motion capture technique. Second, the ergonomics evaluation models can be applied to evaluate the assembly task in the digital environment. The method has been applied to evaluate automobile assembly tasks and some ergonomic improvements have been implemented during assembly tasks in the automotive sector. Note that the proposed method allows to verify the improvements in the digital space and then implement them in the real space.

During the 1990s many researchers developed virtual models to work with CAD systems in order to achieve the workstation ergonomic effective design. SAFEWORK (Fortin et al., 1990) is one example of this technique. More commonly, human form models and analysis tools have been designed for access from within a CAD system. These systems take advantage of the designers familiarity with the terminology, techniques, and command structures of commercially available CAD programs. Examples include MINTAC (Kuusito and Mattila,1990), ErgoSHAPE (Launis and Lehtela, 1992), HUMAN (Sengupta and Das 1997), RAMSIS (Seidl 1997), and commercial systems such as ANYBODY (Porter et al. 1995) and Mannequin. Ulin et al. (1990), Grobelny (1990), Kayis and Iskander (1994) and Jung and Kee (1996) describe other examples.

Moreover the literature analysis reveals also that the majority of the research studies propose three-dimensional CAD programs with built-in ergonomics assessment capabilities. Such ergonomic CAD systems have been described in the literature and among the others the following has to be regarded as the most known: APOLIN (Grobelny et al. 1992), CAAA (Hoekstra 1993), COMBIMAN and Crew Chief (McDaniel 1990), Deneb/ERGO (Nayar 1995), ERGOMAN (Mollard et al. 1992), JACK (Badler et al. 1995), and TADAPS (Westerink et al. 1990). In addition to the previous cited research works, Feyen et al. (2000) propose a PC-based software program that allows a designer to quantify a worker's biomechanical risk for injury based on a proposed workplace design. The program couples an established software tool for biomechanical analysis, the Three-Dimensional Static Strength Prediction Program (3DSSPP), with a widely used computer-aided design software package, AutoCAD. The software program allows the authors to study ergonomic issues during the design phase taking into consideration different design alternatives. The use of this 3DSSPP/AutoCAD interface in the proactive analysis of an automotive assembly task is described and the results compared with an independent assessment using observations of workers performing the same task.

The workstation ergonomic design has been also achieved by using commercial simulation software available for ergonomic studies. Hanson (2000)

presents a survey of the following three tools: ANNIE-Ergoman, JACK, and RAMSIS, used for human simulation and ergonomic evaluation of car interiors. The tools are compared and the comparison shows that all three tools have excellent potential in evaluating car interiors ergonomically in the early design phase. Gill et al. (1998) provide an analysis of the Jack (a simulation software used for human simulation and ergonomic evaluation of car interiors) to highlight the usefulness for applications in the manufacturing industry. Eynard et al. (2000), describe a methodology using Jack to generate and apply body typologies from anthropometric data of Italian population and compare the results with a global manikin. The study identified the importance of using accurate anthropometric data for ergonomic analysis. Sundin et al. (2000), present two case studies to highlight benefits of the use of Jack analysis, one in the design phase of a new Volvo bus and the other in the design phase of the Cupola, a European Space Agency (ESA) module for manned space flights for the International Space Station. Marcos et al. (2006) aim at reducing the stress and strain of the medical staff during laparoscopic operations, and, simultaneously, at increasing the safety and efficiency of an integrated operation room (OR) by an ergonomic redesign. This was attempted by a computer simulation approach based on the integration of the CAD software (CATIA) and the simulation software (RAMSIS). The proposed approach, after defining ergonomically ideal postures, allows to evaluate the optimal solutions for key elements of an ergonomic design of the OR (position and height of the image displays, height of the OR table and the Mayo stand) with special regard to the different individual body size of each member of the team. Cimino and Mirabelli (2009) use the simulation software eMWorkplace to develop a simulation model capable of recreating, with satisfactory accuracy, the evolution over the time of the real workstations. The actual workstation configuration is then analyzed and several workstation modifications are incorporated in the simulation model. The effects (ergonomic risk level) of those changes are analysed for designing an improved workstation configuration in terms of interaction between operators and their industrial working environment. Cimino et al. (2009) address the industrial workstations design issue by proposing an approach based on the integration of Modeling & Simulation tools, several ergonomic standards and the most known work measurement tools. The Modeling and simulation tools allow to implement a three-dimensional environment capable of recreating in a virtual environment the real workstations. The ergonomic standards consent to evaluate the ergonomic risks level within the system being considered. The work measurement tools permit to calculate the time required for performing all the workstations operations. The effective design of the workstations is achieved by using the simulation model for comparing workstations' alternative configurations designed according to the authors' experience. The comparison is based on ergonomic and time indexes related to the ergonomic standards and the work measurement tools. Such comparison allows to choose the workstations final configurations. Finally, Bruzzone et al. (2004) develop a methodology for modeling the human behavior in industrial facilities.

Gap identification and ongoing researches

The high number of studies addressing the workstation design problem reveals that huge research efforts have been carried out in this specific area over the last two decades. However, it seems to be clear that even if research activities have brought significant and high quality results, some further improvements could be achieved. In effect, the literature analysis points out that the majority of the research works assume a trial and error methodology in order to design the workstations ergonomically: the final ergonomic design of the workstations usually depends on researcher's experience and his/her knowledge about the manufacturing system and it is never supported by a well defined experimental design. On the basis of such considerations, further researches can be carried out for giving a significant contribution to the state of the art related to this specific area. A new methodology based on the integration of commercial simulation software, specific standards for the ergonomic analysis and a well planned Design of Experiments (DOE) could be developed. The commercial simulation software would allow to recreate in a virtual environment the real workstations in order to not "disturb" the industrial processes. The ergonomic standards could be applied through the simulation model and would allow to evaluate the ergonomic risks level of the workstations under consideration. Finally, a well defined experimental design (DOE) would allow to generate different workstation configurations; the evaluation of each workstation configuration in terms of ergonomic risks level, carried out by using the ergonomic standards, would allow to choose the final design of the considered workstation.

Workstations Ergonomic Effective Design: some application examples

In the following the authors propose some application examples in which they used the approach described in the previous section to achieve the effective ergonomic design of real industrial workstations.

Figure 1 depicts the 3D virtual model of an assembly line for heaters production (Longo and Mirabelli, 2009). The effective design of assembly line is achieved by simultaneously considering assembly workstations time balancing and ergonomic issues and risks in each workstation. In this case the simulation model is used for a twofold objective: (i) to carry out time analysis in order to increase the assembly line

output (heaters/time unit); (ii) reduce ergonomic risks such as musculoskeletal disorders and wrong working postures.

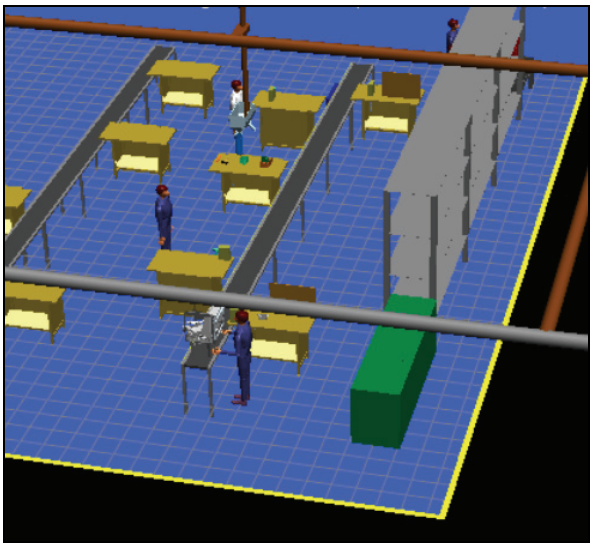


Fig. 1 – 3D Virtual Simulation model of an assembly line for heaters production

Figure 2 depicts a 3D virtual simulation model of a workstation devoted to assemble high pressure hydraulic hoses (Cimino et al. 2009). After a preliminary phase conducted within the real industrial plant (with the aim of collecting data and information to be used for simulation model implementation) the authors used the simulation model combined with Design of Experiment for investigating the behaviour of multiple ergonomic performance measures under the effect of multiple design parameters.

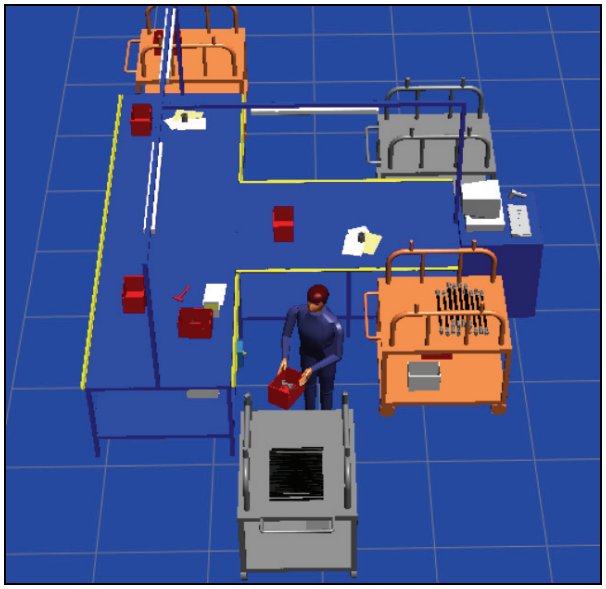


Fig. 3 – 3D virtual simulation model of an assembly workstation devoted to assemble high pressure hydraulic hoses

In this case some of the simulation results include the following: 19% reduction of the total amount of energy required for performing assembly operations, about 14% reduction of the total assembly time, new T-shape layout for the assembly workstation and reduction of not-allowed lifting activities.

Finally figure 3 depicts two workstations devoted to produce leathers goods (De Sensi et al., 2007). In this case simulation is used for the validating the new workstations modular design proposed by the company top management (according to the initial simulation results obtained by the actual workstations layout). The animation during the simulation ha been used for detecting ergonomic problems that otherwise would be difficult to detect (i.e. wrong working postures and wrong disposition of tools).

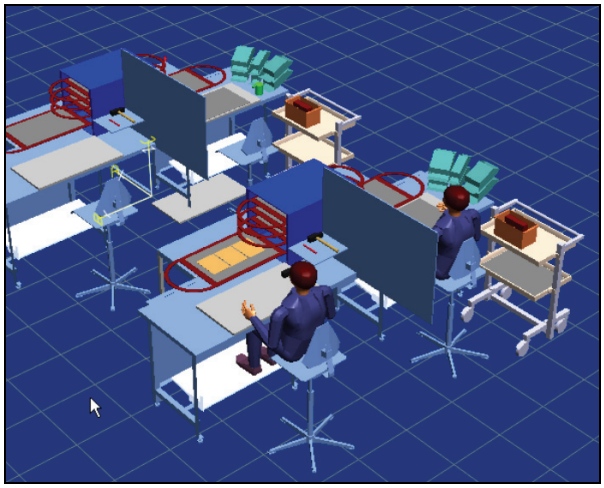


Fig. 3 – Modular design for two workstations used for producing leather goods.

CONCLUSION

The main objective of the paper is to present a literature review concerning the ergonomic effective design of industrial workstations. The initial search identifies a huge number of articles which were reduced to about 60 studies based on content and quality. Such studies have been identified by means of Google Scholar, Scopus, and Scirus as research engines. A first run has been made by typing (sometimes alone, some other times combined) the keywords “industrial workstations”, “industrial design”, “ergonomic effective design”, “workstation design”, “ergonomic standards”, “modeling & simulation”, and “computerized models”. At a second stage, the abstracts of the peer-reviewed and of the specific-interest-groups publications outcomes have been read, to evaluate the actual pertinence of the abstracts to the research issue. At a later stage, starting from the evaluation of the abstracts only the more relevant papers have been read and classified/considered in the literature review. The research works were clustered according to the scientific approach they propose. In this regards, the authors identify two different scientific approaches based on different principles, methods and tools. The

first approach is based on the direct analysis of the real workstations, while the second one use computerized models to design workstations ergonomically. Finally, the literature review is completed with the identification of some gap and a brief description of ongoin research activities that give a significant contribution to the actual state of the art. To this end three different application examples in three different manufacturing areas (heaters production, high pressure hydraulic hoses production and leather goods production) are briefly presented and discussed.

REFERENCES

Badler N.I., Becket W.M., Webber B.L., 1995. Simulation and analysis of complex human tasks for manufacturing. Proc. SPIE - Int. Soc. Opt. Eng. 2596, 225-233.

Bocca E. and Longo F., 2008. Simulation Tools, Ergonomics Principles and Work Measurement Techniques for Workstations Design. In: Proceedings of Summer Computer Simulation Conference. Edinburgh, 16-19 June, SAN DIEGO: pp. 481-486, ISBN/ISSN: 1-56555-323-3

Bruzzone A.G., Viazzo S., Briano C., Massei M., 2004. Modelling Human Behaviour in Industrial Facilities & Business Processes'. Proceedings. of ASTC2004, Arlington, VA, April.

Carrasco C., Coleman N., Healy S., Lusted M., 1995. Packing products for customers: an ergonomics evaluation of three supermarkets checkouts.

Chang S.-W. and Wang M.-J. J, 2007. Digital Human Modeling and Workplace Evaluation: Using an Automobile Assembly Task as an Example. Human Factors and Ergonomics in Manufacturing, 17, 445-455.

Choobineh A., Tosian R., Alhamdi Z., Davarzanie M., 2004. Ergonomic intervention in carpet mending operation. Applied Ergonomics, 35, 493-496.

Chung M. K. and Kee D., 2000. Evaluation of lifting tasks frequently performed during FIre brick manufacturing processes using NIOSH lifting equations. International Journal of Industrial Ergonomics, 25, 423-433.

Cimino A., Mirabelli G., 2009. Modeling, simulation and ergonomic standards as support tools for a workstation design in manufacturing system. International Journal of Simulation and Process Modeling, 5(2), pp 138-148

Cimino A., Longo F., Mirabelli G., Papoff E., 2009. Industrial Workstations Design: A Real Case Study. Proceedings of the 8th International Workshop on Modeling & Applied Simulation (MAS09), Tenerife, Spain

Cimino A., Curcio D., Longo F., Mirabelli G., 2009. Improving workers conditions within Industrial Workstations. In: Proceedings of the Summer Computer Simulation Conference. Istanbul, Turkey, 13-16 July, 2009 Society for Modeling & Simulation International.

Cimino A., Curcio D., Longo F., Papoff E., 2008. Workstation productivity enhancement within hydraulic hoses manufacturing process. In: Proceedings of the International Workshop on Modelling & Applied Simulation. Campora S. Giovanni, 17-19 September, Vol. I, p. 268-274, ISBN/LISSN: 978-88-903724-1-4

Cimino A., Longo F., Mirabelli G., Papoff E. 2008. Most and MTM for Work Method Optimization: a Real Case Study based on Modeling & Simulation. In: Proceedings of the International Workwhop on Harbour Maritime Multimodal Logistics Modelling & Simulation. Campora S. Giovanni (CS), Italy, 17-19 September, Vol. I, p. 35-41, ISBN/ISSN: 978-88-903724-2-1.

Das B. and Sengupta A. K., 1996. Industrial workstation design: a systematic ergonomics approach. Applied Ergonomics, 27, 157-163.

De Sensi G., Longo F., Mirabelli G., 2007. Modeling & Simulation for workplaces analysis in the leather industry. In: Proceedings of the International Mediterranean Modelling Multiconference. Bergeggi, Italy, October 4-6, pp. 225-230, ISBN/ISSN: 88-900732-6-8

Engström T. and Medbo P., 1997. Data collection and analysis of manual work using video recording and personal computer techniques. International Journal of Industrial Ergonomics, 19, 291-298.

Eynard E., Fubini E., Masali M., Cerrone M., Tarzia A., 2000. Generation of virtual man models representative of different body proportions and application to ergonomic design of vehicles, in: Proceedings of the XIVth Triennial Congress of the International Ergonomics Association and 44th Annual Meeting of the Human Factors and Ergonomics Association, Ergonomics for the New Millennium', 29 July to 4 August 2000, San Diego, CA, USA.

Feyen R., Liu Y., Chaffin D., Jimmerson G., Joseph B., 2000. Computer-aided ergonomics: a case study of incorporating ergonomics analyses into workplace design. Applied Ergonomics, 31, 291-300.

Forsman M., Hansson G-Å, Medbo L., Asterland P., Engström T., 2002. A method for evaluation of manual work using synchronised video recordings and physiological measurements. Applied Ergonomics, 33, 533-540.

Fortin C., Gilbert R., Beuter A., Laurent F., Schiettekatte J., Carrier R., Dechamplain B., 1990. SAFEWORK: a microcomputer-aided workstation design and analysis. New advances and future developments. In: Karkowski, W., Genaidy, A.M., Asfour, S.S. (Eds.), Computer-Aided Ergonomics. Taylor and Francis, London, pp. 157-180.

Garg A., 1976. A Metabolic Rate Prediction for Manual Materials Handling Jobs, Dissertation, University of Michigan, Ann Arbor, Michigan, USA.

Gill S.A. and Ruddle R.A., 1998. Using virtual humans to solve real ergonomic design problems, in: Proceedings of the 1998 International Conference on Simulation, IEEE Conference Publication 457, pp. 223–229.

González B. A., Adenso-Dìaz B., González Torre P., 2003. Ergonomic performance and quality relationship: an empirical evidence case. International Journal of Industrial Ergonomics, 31, 33-40.

Grant K. A., Habes D. J., Tepper A. L., 1995. Work activities and musculoskeletal complaints among preschool workers. Applied Ergonomics, 26, 405-410.

Grant K. A., Habes D. J., Bertsche P.K., 1997. Lifting hazards at a cabinet manufacturing company: evaluation and recommended controls. Applied occupational and environmental Hygiene, 12, 253-258.

Grobelny J., 1990. Anthropometric data for a driver's workplace design in the AutoCAD system. In: Karkowski, W., Genaidy, A.M., Asfour, S.S. (Eds.), Computer-Aided Ergonomics. Taylor & Francis, London, pp. 80-89.

Grobelny J., Cysewski P., Karkowski W., Zurada J., 1992. APOLIN: a 3-dimensional ergonomic design and analysis system. In: Mattila, M., Karkowski, W. (Eds.), Computer Applications in Ergonomics, Occupational Safety and Health. Elsevier, Amsterdam, pp. 129-135.

Hanson L., 2000. Computerized tools for human simulation and ergonomic evaluation of car interiors, in: Proceedings of the XIVth Triennial Congress of the International Ergonomics Association and 44th Annual Meeting of the Human Factors and Ergonomics Association, Ergonomics for the New Millennium', 29 July to 4 August 2000, San Diego, CA, USA.

Hoekstra P.N., 1993. Some uses of active viewing in computer aided anthropometric assessment. Des. Diversity Proc. Human Factors Ergon. Soc. 1, 494-498.

Jayaram U., Jayaram S., Shaikh I., Kim Y., Palmer C., 2006. Introducing quantitative analysis methods into virtual environments for real-time and continuous ergonomic evaluations. Computers in Industry, 57, 283-296.

Jones T., Strickfaden M., Kumar S., 2005. Physical demands analysis of occupational tasks in neighborhood pubs. Applied Ergonomics, 36, 535-545.

Jones T. and Kumar S., 2007. Comparison of ergonomic risk assessments in a repetitive high-risk sawmill occupation: Saw-filer. International Journal of Industrial Ergonomics, 37, 744-753.

Jung E.S., Kee D., 1996. Man-machine interface model with improved visibility and reach functions. Comp. Ind. Eng. 30 (3), 475-486.

Kadefors R. and Forsman M., 2000. Ergonomic evaluation of complex work: a participative approach employing video computer interaction, exemplified in a study of order picking. International Journal of Industrial Ergonomics, 25, 435-445.

Karhu O., Harkonen, R., Sorvali, P. and Vepsalainen, P. (1981) 'Observing working postures in industry: Examples of OWAS application', Applied Ergonomics, Vol. 12, No. 1, pp.13–17.

Kayis B. and Iskander P.A., 1994. Three-dimensional human model for the IBM/CATIA system. Applied Ergon. 25 (6), 395-397.

Kuusito A. and Mattila M., 1990. Anthropometric and biomechanical man models in computer-aided ergonomic design structure and experiences of some programs. In: Karkowski, W., Genaidy, A.M., Asfour, S.S. (Eds.), Computer-Aided Ergonomics. Taylor& Francis, London, pp. 104-114.

Laring J., Forsman M., Kadefors R., Örtengren R., 2002. MTM-based ergonomic workload analysis. International Journal of Industrial Ergonomics, 30, 135-148.

Launis M. and Lehtela J., 1992. ergoSHAPE: a design oriented ergonomic tool for AutoCAD. In: Mattila, M., Karkowski, W. (Eds.), Computer Applications in Ergonomics, Occupational Safety and Health. Elsevier, Amsterdam, pp. 121-128.

Lin R.-T. and Chan C.-C., 2007. Effectiveness of workstation design on reducing musculoskeletal risk factors and symptoms among semiconductor fabrication room workers. International Journal of Industrial Ergonomics, 37, 35-42.

Longo F. and Mirabelli G., 2009. Effective Design of an Assembly Line using Modeling & Simulation. Journal of Simulation, Vol. 3; p. 50-60, ISSN: 1747-7778, doi: 10.1057/JOS.2008.18

Longo F., Mirabelli G., Papoff E.,, 2006. Simulation Models for Work Methods Optimization in a Real Manufacturing Plant. In: Proceedings of the Modern Information Technology in The Innovation Processes of The Industrial Enterprises. September 11th – 12th 2006, Budapest, Hungary, p. 463-468, ISBN/ISSN: 963-86586-5-7

Marcos P., Seitz T., Bubb H., Wichert A., Feussner H., 2006. Computer simulation for ergonomic improvements in laparoscopic surgery. Applied Ergonomics, 37, 251-258.

McDaniel J.W., 1990. Models for ergonomic analysis and design: COMBIMAN and CREW CHIEF. In: Karkowski, W., Genaidy, A.M., Asfour, S.S. (Eds.), Computer-Aided Ergonomics. Taylor & Francis, London, pp. 138-156.

Mollard R., Ledunois S., Ignazi G., Coblentz A., 1992. Researches and developments on postures and movements using CAD techniques and ERGODATA. In: Mattila, M., Karkowski, W. (Eds.), Computer Applications in Ergonomics, Occupational Safety and Health. Elsevier, Amsterdam, pp. 337-343.

Nayar N., 1995. Deneb/ERGO: a simulation based human factors tool. Winter Simulation Conference Proceedings 1995, 427-431.

Neumann W.P., Wells R.P., Norman R.W., Kerr M.S., Frank J., Shannon H.S., OUBPS Working Group, 2001. Trunk posture:reliability, accuracy, and risk estimates for low back pain from a video based assessment method. International Journal of Industrial Ergonomics, 28, 355-365.

Niosh Technical Report 81–122, 1981. National Institute for Occupational Safety and Health (Hrsg.), Work Practices guide for Manual Lifting, NTIS, Center for Disease Control, US Department of Health and Human Services, Cincinnati, OH, USA.

Porter J.M., Freer M., Case K., Bonney M.C., 1995. Computer aided ergonomics and workspace design. In: Wilson, J.R., Corlett, E.N. (Eds.), Evaluation of Human Work: A Practical Ergonomics Methodology, 2nd Edition. Taylor and Francis, London, pp. 574-620.

Resnick M.L. and Zanotti A., 1997. Using ergonomics to target productivity improvements. Computers and Industrial Engineering, 33, 185-188.

Russell S. T., Winnemuller L., Camp J. E., Johnson P. W., 2007. Comparing the results of five lifting analysis tools. Applied Ergonomics, 38, 91-97.

Schultetus W., 1980. Daten, Hinweise und Beispiele zur Ergonomischen Arbeitsgestaltung, Montagegestaltung, Verlag TÜV Rheinland Gmbh, Köln.

Scientific Support Documentation for the Revised 1991 NIOSH Lifting Equation (1991) Technical Contract Reports, 8 May, NTIS No. PB-91-226-274.

Scott G.B. and Lambe N.R., 1996. Working practices in a perchery system, using the OVAKO Working posture Analysing System (OWAS). Applied Ergonomics, 27, 281-284.

Seidl A., 1997. RAMSIS: a new CAD-tool for ergonomic analysis of vehicles developed for the German automotive industry. Automotive Concurrent/Simultaneous Engineering SAE Special Publications, Vol. 1233, pp. 51-57.

Sengupta A.K., Das B., 1997. Human: an AutoCAD based three dimensional anthropometric human model for workstation design. Int. J. Ind. Ergon. 19 (5), 345-352.

Shuval K. And Donchin M., 2005. Prevalence of upper extremity musculoskeletal symptoms and ergonomic risk factors at a Hi-Tech company in Israel. International Journal of Industrial Ergonomics, 35, 569-581.

Sundin A., Christmansson M., Ortengren R., 2000. Methodological differences using a computermanikin in two case studies:Bus and space module design, in: Proceedings of the XIVth Triennial Congress of the International Ergonomics Association and 44th Annual Meeting of the Human Factors and ErgonomicsAssociation,Ergonomics for theNewMillennium', 29 July to 4 August 2000, San Diego, CA, USA, pp. 496–498.

Temple R. and Adams T., 2000. Ergonomic Analysis of a Multi-Task Industrial Lifting Station Using the NIOSH Method. Journal of Industrial Technology, 16, 1-6.

Ulin S.S., Armstrong T.J., Radwin R.G., 1990. Use of computer aided drafting for analysis and control of posture in manual work. Appl. Ergon. 21 (2), 143-151.

van Wendel de Joode B., Burdorf A., Verspuy C., 1997. Physical load in ship maintenance: Hazard evaluation by means of a workplace survey. Applied Ergonomics, 28, 213-219.

Vedder J., 1998. Identifying postural hazards with a video-based occurrence sampling method. International Journal of Industrial Ergonomics, 22, 373-381.

Waters T. R., Vern P. A., Garg A., 1994. Application Manuals for the Revised NIOSH Lifting Equation. National Institute for Occupational Safety and Health, Cincinnati, OH, USA.

Westerink J.A,, Tragter H., Van Der Star A., Rookmaaker D.P., 1990. TADAPS: a three-dimensional CAD man model. In: Karkowski, W., Genaidy, A.M., Asfour, S.S. (Eds.), Computer-Aided Ergonomics. Taylor & Francis, London, pp. 90-103.

White H. A. and Kirby L. R. , 2003. Folding and unfolding manual wheelchairs:an ergonomic evaluation of health-care workers. Applied Ergonomics, 34, 571-579.

Wilson J., 1997. Virtual environments and ergonomics: needs and opportunities. Ergonomics, 40, 1057-1077.

AUTHOR BIOGRAPHIES

Terry Bossomaier is professor in Information Technology and Director of CriCs at Charles Sturt University, School of Computing & Mathematics. His research interests includ, among others, Neural networks, Complex systems, Evolutionary computing and Parallel computing. He is authors and co-authors of numerous papers on international journals and conferences. His web page is available at http://www.csu.edu.au/faculty/business/comp-math/staff/tbossomaier.htm

Agostino Bruzzone since 1991, he has taught "Theories and Techniques of Automatic Control" and in 1992 he has become a member of the industrial simulation work group at the ITIM University of Genoa; currently he is Full Professor in DIPTEM. He has utilized extensively simulation techniques in harbour terminals, maritime trading and sailboat racing sectors. He has been actively involved in the scientific community from several years and served as Director of the McLeod Institute of Simulation Science (MISS), Associate Vice-President and Member of the Board of the SCS (Society for Modelling & Simulation international), President of the Liophant Simulation, VicePresident of MIMOS (Movimento Italiano di Simulazione) and Italian Point of Contact for the ISAG (International Simulation Advisory Grouyp) and Sim-Serv. He has written more than 150 scientific papers in addition to technical and professional reports in partnerships with major companies (i.e. IBM, Fiat Group, Contship, Solvay) and agencies (i.e. Italian Navy, NASA, National Center for Simulation, US Army). He teaches "Project Management" and "Industrial Logistics" at the University for students in the Mechanical Engineering (4th year), Management Engineering (4th year) and Logistics & Production Engineering (3rd Year) Degree Courses. His Web-page can be found at http://www.liophant.org/liophant/memb/agostino.html.

Antonio Cimino took his degree in Management Engineering, summa cum Laude, in September 2007 from the University of Calabria. He is currently PhD student at the Mechanical Department of University of Calabria. He has published more than 20 papers on international journals and conferences. His research activities concern the integration of ergonomic standards, work measurement techniques, artificial intelligence techniques and Modeling & Simulation tools for the effective workplace design. His e-mail address is: acimino@unical.it and his Web-page can be found at http://www.ingegneria.unical.it/impiantiindustriali/index_file/Cimino.htm

Francesco Longo received his Ph.D. in Mechanical Engineering from University of Calabria in January 2006. He is currently Assistant Professor at the Mechanical Department of University of Calabria and Director of the Modelling & Simulation Center – Laboratory of Enterprise Solutions (MSC-LES). He has published more than 80 papers on international journals
and conferences. His research interests include Modeling & Simulation tools for training procedures in complex environment, supply chain management and security. He is Associate Editor of the "Simulation: Transaction of the society for Modeling & Simulation International". For the same journal he is Guest Editor of the special issue on Advances of Modeling & Simulation in Supply Chain and Industry. He is Guest Editor of the "International Journal of Simulation and Process Modelling", special issue on Industry and Supply Chain: Technical, Economic and Environmental Sustainability. He is Editor in Chief of the SCS M&S Newsletter and he works as reviewer for different international journals. His e-mail address is: f.longo@unical.it and his Web-page can be found at http://www.ingegneria.unical.it/impiantiindustriali/index_file/Longo.htm

Giovanni Mirabelli is currently Assistant Professor at the Mechanical Department of University of Calabria. He has published more than 60 papers on international journals and conferences. His research interests include ergonomics, methods and time measurement in manufacturing systems, production systems maintenance and reliability, quality. His e-mail address is: g.mirabelli@unical.it and his Web-page can be found at
http://www.ingegneria.unical.it/impiantiindustriali/index_file/Mirabelli.htm

MODELLING AND SIMULATION OF DRY ANAEROBIC FERMENTATION

Zdenka Prokopová and Roman Prokop
Faculty of Applied Informatics
Tomas Bata University in Zlín,
Nad Stráněmi 4511, 760 05 Zlín, Czech Republic
E-mail: prokopova@fai.utb.cz

KEYWORDS
Modelling, Simulation, Anaerobic Fermentation, Waste treatment, Biogas production.

ABSTRACT
The paper is focused on mathematical modelling and computer simulation of the anaerobic fermentation mechanism for neutral or slightly acid and acid fermentation. Degradation of organic substances to final products, methane and carbon dioxide, involves their coordinated metabolic cooperation. A product of one microorganism group turns into substrate for the subsequent ones. Generally, the studied anaerobic fermentation processes progress in four stadiums, therefore a mathematical model of four-level decomposition is used. All mentioned processes were modeled through differential equations and computed and simulated in the MATLAB+SIMULINK environment.

INTRODUCTION

The necessity for alternative "green" energy from renewable resources has enhanced the role of environmental management of ecosystems. Today, anaerobic fermentation is widely accepted as a sound technology for many waste treatment applications, and novel reactor designs are being applied on a commercial scale. In spite of this acceptance, advances are still being made, and our developments are concentrating on the uses of small amount of biodegradable mass - especially for dry discontinuous fermentation processes.
Anaerobic fermentation is a biological process of organic mass decay which proceeds without oxygen (air). This process runs naturally in country e.g. in marshes, at the lake bottom but it is used also in different types of wastes (communal waste dump, cow and poultry manure, liquids from the agro-industries etc.). Mixed culture of microorganism in several steps decay organic mass during this process. A product of one microorganism group turns into substrate for the other group (Ahring, 2003).
The fermentation process can be divided into four main phases:

- Hydrolysis: by the activity of extracellular enzymes, macromolecular materials are outside the cell split into simpler organic substances, first of all fatty acids, alcohols, carbon dioxide (CO2) and molecular hydrogen (H2).
- Acidogenesis: products of hydrolysis are inside the cell rotted into simpler substances (acids, alcohols, carbon dioxide and molecular hydrogen). By the fermentation of these substances is generating mixture of products whose composition depends on initial substrate and reaction conditions. Under the low concentration of hydrogen is generating acetic acid. Under the higher concentration of hydrogen is generating lactic acid and alcohol. Another important factor is pH value of reaction mixture. When the pH is neutral or slightly acid it dominates the butyric fermentation and when the pH is more acid (3-4) it dominates lactic fermentation.
- Acetogenesis: in this step substances produced by acidogenesis are spread out into molecular hydrogen, carbon dioxide and acetic acid.
- Methanegenesis: it is the last stadium of the anaerobic decay when from the acetic acid, hydrogen and carbon dioxide rises methane - CH4. This step is performing by methanogene microorganisms which are strictly anaerobic organism and oxygen is poison for them.

The main product of anaerobic fermentation of organic mass is biogas. Biogas is colorless gas consisting primarily of methane (approx. 60%) and carbon dioxide (approx. 40%). It is able to contain small quantities of N2, H2S, NH3, H2O, ethane and lower hydrocarbons. As a secondary product there is a stabilized anaerobic material (fermentation remainder, digestat, ferment) which is mostly exploited as a fertilizer material (Straka, 2003).
A fermentation processes usually run in large heated and mixed (stirred) tanks – fermentation reactors. It is a continuous or semicontinuous process. The tank size is given by quantity and quality of material, quantity of active biomass in the reactor and the desired time delay. These parameters significantly influence the production of biogas and quality of output materials.
In light of reactionary temperatures we can divide anaerobic processes, according to optimal temperature for microorganism to psychrophilic (5-30°C), mesophilic (30-40°C), thermophilic (45-60°C) and extremely thermophilic (up 60°C). Most common applications are processes mesophilic at temperature approximately 38°C (Froment & Bischotf, 1990).

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

Wet fermentation technology

Most widely used technology of biogas production is so-called "wet fermentation", which processes substrates with resulting dry matter content <12%. Wet anaerobic fermentation proceeds in reserved vessels (fermenters/reactors). These vessels are heated on designed operational temperature and mixed.

Dry fermentation technology

The technology of dry fermentation processes substrates about 30 - 35% of dry matter. It can be used for biomass production which is not possible to work up by wet fermentation (e.g. slurry, sawdust, straw, grass, foliage, leaf litter, wood waste). Generally, there are mesophilic conditions of anaerobic process; range of used reactionary temperatures is 32-38°C. The optimal pH is usually between 6.5 – 7.5. In principle, it is possible to divide out technologies on discontinuous (batch) and continuous one.
The discontinuous technology consists of several reaction chambers (metallic containers or bricked chambers) and a buffer stock. An anaerobic process is treated by the dosage of procedural liquids. For inoculation/vaccinations needs is exploited partly periodic injection of so-called percolate (material with content of suitable anaerobic cultures) and additions of fermentative remainder from previous cycle to the fresh substrate. From the investment and operational point of view discontinuous technologies are essentially less exacting than continuous ones.

THEORETICAL BACKGROUND

The simplest dynamic quantitative model of a complex fermentation process represents a set of four simple differential equations according to the following scheme:

$$c_1 \xrightarrow{k_1} c_2 \xrightarrow{k_2} c_3 \xrightarrow{k_3} c_4 \xrightarrow{k_4} c_5 \qquad (1)$$

where k_1, k_2, k_3, k_4 are speed constants for each subsequent reaction.
Quantitative models of steady states are based on the mass and thermal balance of particular processes. It is necessary to emphasize that quantitative models rise from laws of mass and energy conservation equations. Hydrolysis of the cellulose is described in a lot of publications e.g. (Swift, 1998), (Smith, 1998) we shall deal only with acidogenesis, acetogenesis and methanogenesis. The total anaerobic decomposition of the glucose as a product of the cellulose at usage of 100% substrate is possible to describe by the following formula:

$$C_6H_{12}O_6 \rightarrow 3CH_4 + 3CO_2 \qquad (2)$$

According to the presented chemical equation from 1 kmol of 100% glucose can be obtained 6 kmol (134.4 m^3) of biogas (50% methane and 50% carbon dioxide) by anaerobic fermentation.
The quantitative description of the fermentation process depends on the idea of chemical mechanism of all four reactions - hydrolysis, acidogenesis, acetogenesis and methanogenesis. With the assumption that the source material is substrate, whose main biodegradable component forms cellulose, we can compile two basic mathematical models according the acidogenesis mechanism (Kodriková, 2004), (Kolomazník & Kodrikova, 2008).

Neutral or slightly acid fermentation

The butyric fermentation dominates When the pH value of reaction mixture is neutral or slightly acid. The complex mass balance of the chemical mechanism is then described by the following equations:

Hydrolysis: with speed constant = k_0

$$(C_6H_{12}O_5)_n + nH_2O =_n C_6H_{12}O_6 \qquad (3)$$

Acidogenesis: with speed constant = k_1

$$C_6H_{12}O_6 = C_4H_8O_4 + 2CO_2 + 2H_2 \qquad (4)$$

Acetogenesis: with speed constant = k_2

$$C_4H_8O_4 + 2H_2O = 2C_2H_4O_2 + 2H_2 \qquad (5)$$

Methanogenesis: with speed constant = k_4

$$2C_2H_4O_2 = 2CH_4 + 2CO_2 \qquad (6)$$

and with speed constant = k_3

$$CO_2 + 4H_2 = CH_4 + 2H_2O \qquad (7)$$

The whole chemical status described by Equations (3) – (7) is illustrated by the following kinetic graph.

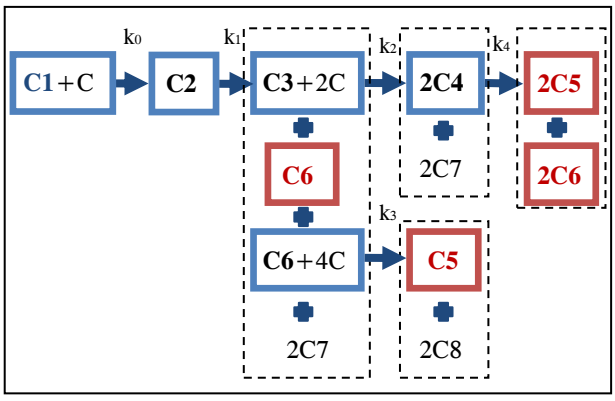


Figure 1: Kinetic graph of neutral or slightly acid fermentation

In Figure 1, abbreviations where C1 is cellulose, C2 is glucose, C3 is butyric acid, C4 is acetic acid, C5 is

methane, C6 is carbon dioxide, C7 is hydrogen and C8 is water are used.

Acid fermentation

The lactic fermentation dominates for the lower (3-4) pH value of the reaction mixture. Then the complex mass balance of the chemical mechanism is described by the following equations:

Hydrolysis: with speed constant = k'_0

$$(C_6H_{12}O_5)_n + nH_2O =_n C_6H_{12}O_6 \quad (8)$$

Acidogenesis: with speed constant = k'_1

$$C_6H_{12}O_6 = 2C_3H_6O_3 \quad (9)$$

Acetogenesis: with speed constant = k'_2

$$2C_3H_6O_3+2H_2O = 2C_2H_4O_2+4H_2+2CO_2 \quad (10)$$

Methanogenesis: with speed constant = k'_4

$$2C_2H_4O_2 = 2CH_4+2CO_2 \quad (11)$$

and with speed constant = k'_3

$$CO_2 + 4H_2 = CH_4+2H_2O \quad (12)$$

The whole chemical status described by Equations (8) – (12) is illustrated by the following kinetic graph.

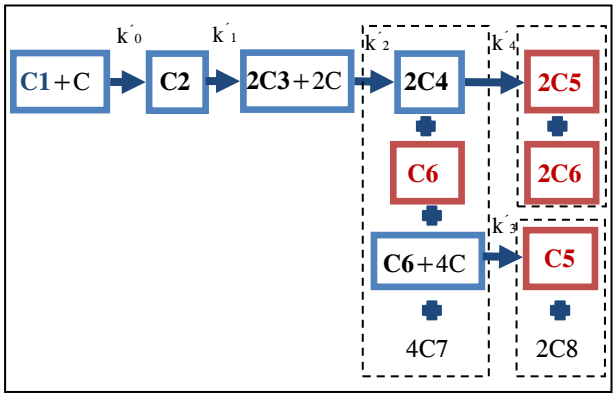


Figure 2: Kinetic graph of acid fermentation

In Figure 2, abbreviations where C1 is cellulose, C2 is glucose, C3 is is lactic acid, C4 is acetic acid, C5 is methane, C6 is carbon dioxide, C7 is hydrogen and C8 is water are used.

MATHEMATICAL MODEL

Now, the dynamic mathematical model can be summarized according to the mentioned chemical reactions, mechanisms and flows.

Neutral or slightly acid fermentation

The first set of differential equations describes the chemical processes shown in Figure 1 representing neutral and/or slightly acid fermentation:

$$\frac{dc_1}{dt} = -k_0c_1c_8 \quad (13)$$

$$\frac{dc_2}{dt} = k_0c_1c_8 - k_1c_2 \quad (14)$$

$$\frac{dc_3}{dt} = k_1c_2 - k_2c_3c_8^2 \quad (15)$$

$$\frac{dc_4}{dt} = k_2c_3c_8^2 - k_4c_4 \quad (16)$$

$$\frac{dc_5}{dt} = k_4c_4 + k_3c_6c_7^4 \quad (17)$$

$$\frac{dc_6}{dt} = k_1c_2 + k_4c_4 - k_3c_6c_7^4 \quad (18)$$

$$\frac{dc_7}{dt} = k_1c_2 + k_2c_3c_8^2 - k_3c_6c_7^4 \quad (19)$$

$$\frac{dc_8}{dt} = -k_2c_3c_8^2 + k_3c_6c_7^4 \quad (20)$$

where c_1 represents cellulose, c_2 glucose, c_3 butyric acid, c_4 acetic acid, c_5 methane, c_6 carbon dioxide, c_7 hydrogen and c_8 water respectively.

Acid fermentation

The second dynamic mathematical model for the acid fermentation described previously and summarized in Figure 2 is presented by the set of the following differential equations:

$$\frac{dc_1}{dt} = -k'_0c_1c_8 \quad (21)$$

$$\frac{dc_2}{dt} = k'_0c_1c_8 - k'_1c_2 \quad (22)$$

$$\frac{dc_3}{dt} = k'_1c_2 - k'_2c_3^2 \quad (23)$$

$$\frac{dc_4}{dt} = k'_2c_3^2c_8^2 - k'_2c_6^2 - k'_2c_7^4 - k'_4c_4^2 \quad (24)$$

$$\frac{dc_5}{dt} = k'_4c_4^2 + k'_3c_6c_7^4 \quad (25)$$

$$\frac{dc_6}{dt} = k'_2c_3^2c_8^2 + k'_4c_4^2 - k'_3c_6c_7^4 \quad (26)$$

$$\frac{dc_7}{dt} = k'_2c_3^2c_8^2 - k'_3c_6c_7^4 \quad (27)$$

$$\frac{dc_8}{dt} = -k'_0c_1c_8 - k'_2c_3^2c_8^2 + k'_3c_6c_7^4 \quad (27)$$

where c_1, c_2, c_4, c_5, c_6, c_7 and c_8 represents the same sense as in Equations (13)-(20), c_3 represents lactic acid.

SIMULATIONS

Simulation experiments with dry fermentation of biomass has been provided in order to get the information about possibilities of its utilization for liquidation of biodegradable sorted communal waste dump. The results of simulations and obtained values are utilized for the design and development of technological reactors and plants.

As the fermenter we plan to use the ordinary commercial composter with some construction modifications. For this purpose we need to find out time behaviors of particular processes - time dependence of initial substances, intermediate and final products concentrations, their speed constants and so on.

Neutral or slightly acid fermentation

Experiment 1:

Speed constants were chosen as k_0=0.9, k_1=0.8, k_2=0.7, k_3=0.6, k_4=0.6.

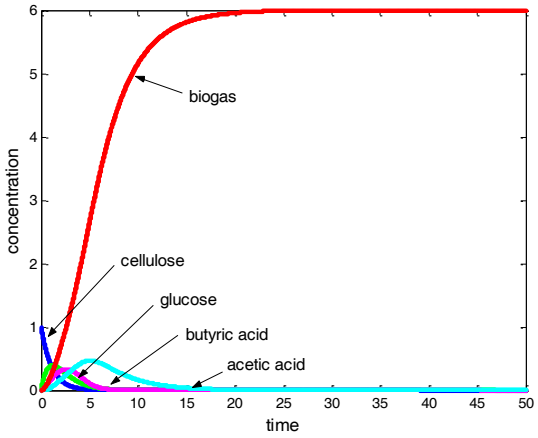


Figure 3: Time dependence of the initial substance, intermediate products and biogas concentration-Exp.1.

Experiment 2:

Speed constants were chosen as k_0=0.9, k_1=0.8, k_2=0.7, k_3=0.2, k_4=0.2.

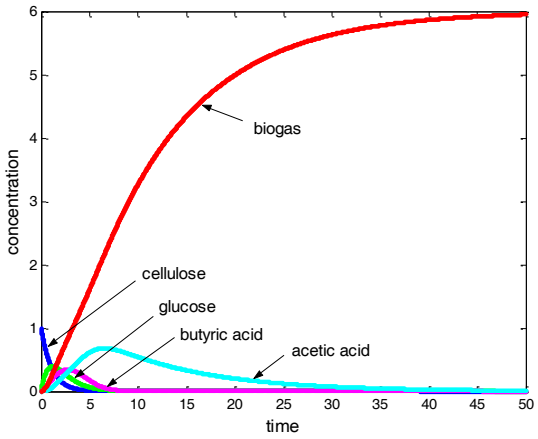


Figure 4: Time dependence of the initial substance, intermediate products and biogas concentration-Exp.2.

Experiment 3:

Speed constants were chosen as k_0=0.9, k_1=0.8, k_2=0.7, k_3=0.6, k_4=0.2.

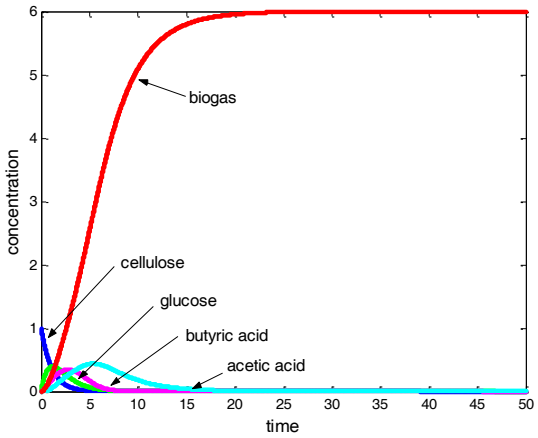


Figure 5: Time dependence of the initial substance, intermediate products and biogas concentration-Exp.3.

Experiment 4:

Speed constants were chosen as k_0=0.9, k_1=0.8, k_2=0.7, k_3=0.2, k_4=0.6.

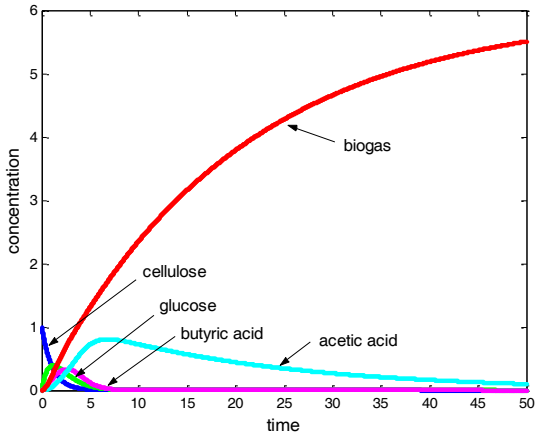


Figure 6: Time dependence of the initial substance, intermediate products and biogas concentration-Exp.4.

Experiment 5:

Speed constants were chosen as k_0=0.6, k_1=0.5, k_2=0.4, k_3=0.3, k_4=0.2.

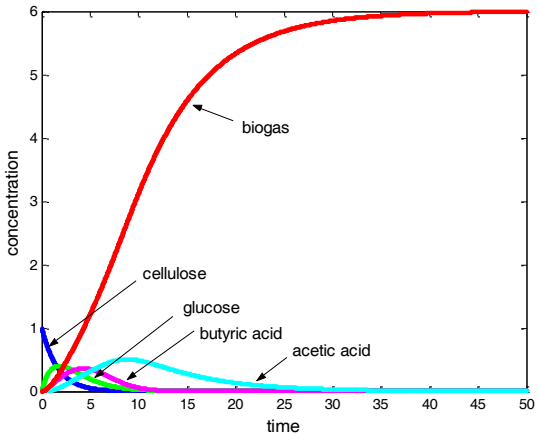


Figure 7: Time dependence of the initial substance, intermediate products and biogas concentration-Exp.6.

Experiment 6:

Speed constants were chosen as k_0=0.2, k_1=0.3, k_2=0.4, k_3=0.5, k_4=0.6.

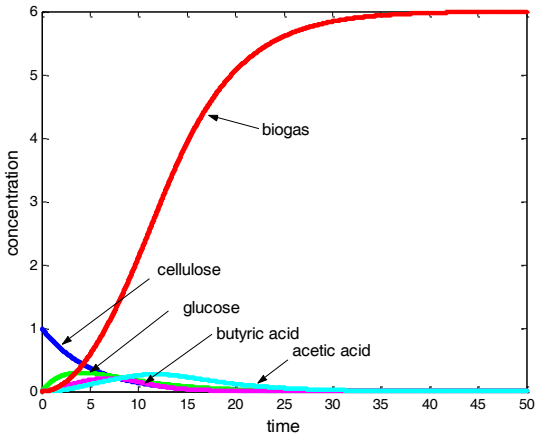


Figure 8: Time dependence of the initial substance, intermediate products and biogas concentration-Exp.6.

Acid fermentation

Experiment 7:

Speed constants were chosen as k'_0=0.9, k'_1=0.8, k'_2=0.7, k'_3=0.6, k'_4=0.6.

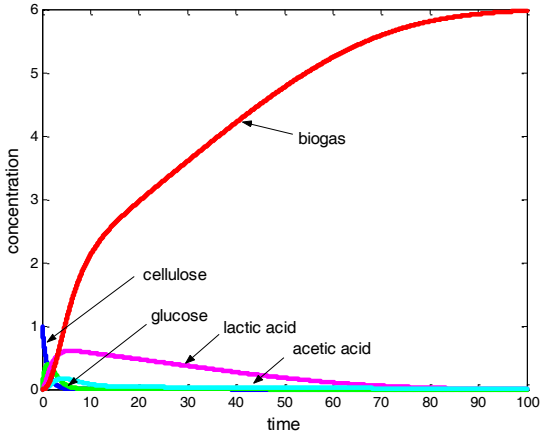


Figure 9: Time dependence of the initial substance, intermediate products and biogas concentration-Exp.7.

Experiment 8:

Speed constants were chosen as k'_0=0.9, k'_1=0.8, k'_2=0.7, k'_3=0.2, k'_4=0.2.

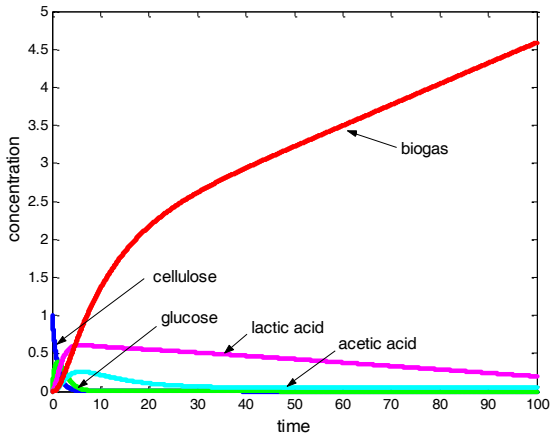


Figure 10: Time dependence of the initial substance, intermediate products and biogas concentration-Exp.8.

Experiment 9:

Speed constants were chosen as k'_0=0.9, k'_1=0.8, k'_2=0.7, k'_3=0.8, k'_4=0.8.

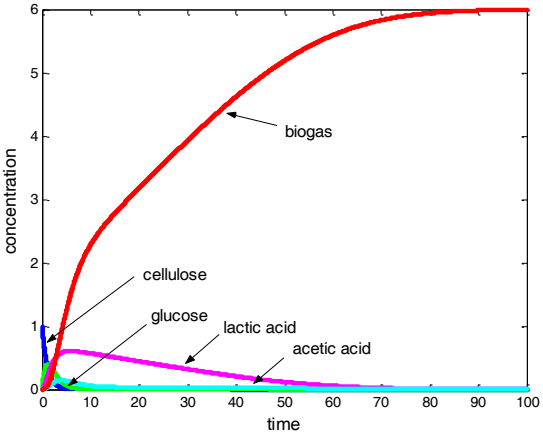


Figure 11: Time dependence of the initial substance, intermediate products and biogas concentration-Exp.9.

Experiment 10:

Speed constants were chosen as k'_0=0.9, k'_1=0.8, k'_2=0.7, k'_3=0.2, k'_4=0.8.

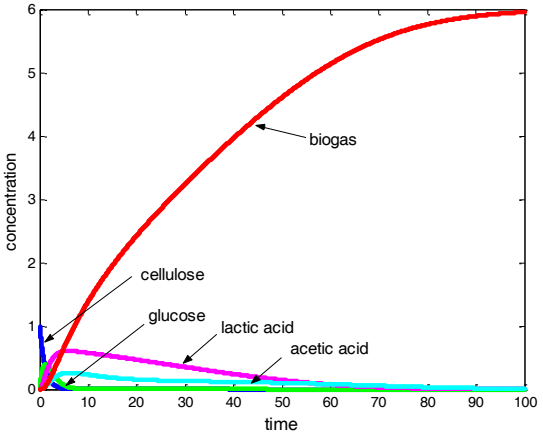


Figure 12: Time dependence of the initial substance, intermediate products and biogas concentration-Exp.10.

Experiment 11:

Speed constants were chosen as k'_0=0.5, k'_1=0.6, k'_2=0.7, k'_3=0.8, k'_4=0.9.

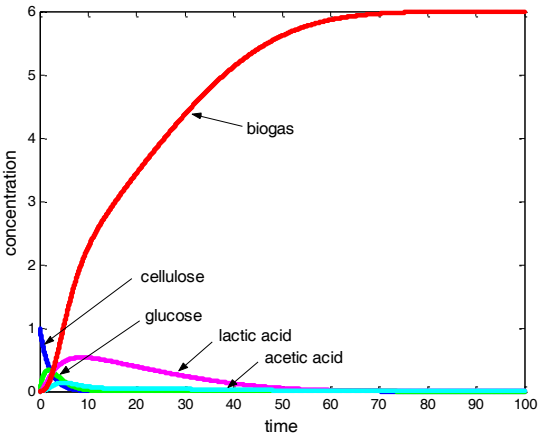


Figure 13: Time dependence of the initial substance, intermediate products and biogas concentration-Exp.11.

Experiment 12:

Speed constants were chosen as k'_0=0.2, k'_1=0.3, k'_2=0.4, k'_3=0.5, k'_4=0.6.

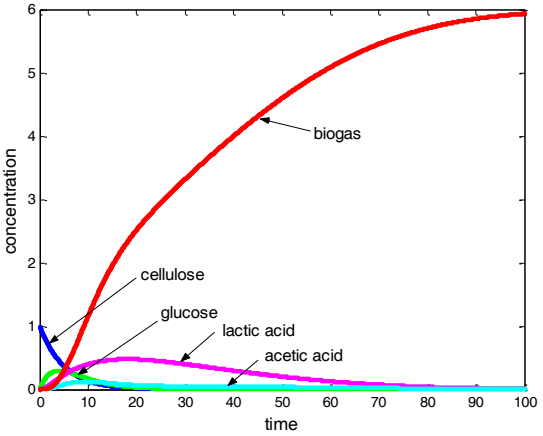


Figure 14: Time dependence of the initial substance, intermediate products and biogas concentration-Exp.10.

CONCLUSION

Biogas production combines the short-term economic needs of human communities with nature of conservation and the end of ecological degradation. This process producing alternative energy also generate other valuable materials as fertilizers, soil conditioners, animal and fish feed and so on. Further benefits of this technology can be seen also in e.g. odor problem reduction, microorganism pathogens control, water resources protection etc.

Deeper understanding and insight into special anaerobic technology is a key point for practical production of the biogas. This paper brings a look to the behavior of anaerobic decay of cellulose and its time progression under the various conditions. Mathematical models have been based only on the mass balance of particular processes. We have started from presumption that anaerobic reaction is exothermal and fermenter works without external heating. Temperature influence of reactions (interactions among microorganism) is partly hidden in speed constants. The obtained results and performed experiments demonstrate that the neutral or slightly acid fermentation is more fast (quick) than acid fermentation. The speed of fermentation depends on the activity of catalyzer which is used as an accelerator of chemical reactions. The presented figures and graphs disclose that the complex behavior of reactions is strongly influenced by the speed constants k_3 (resp. k'_3) and k_4 (resp. k'_4).

All analytical dynamic mathematical models were calculated and simulated in the MATLAB-SIMULINK environment.

Acknowledgments

This work was supported by the Ministry of Education of the Czech Republic in the range of the project No. MSM 7088352102.

REFERENCES

Ahring, B.K. (2003). Perspectives for Anaerobic Digestion. In: Advantages in Biochemical Engineering/Biotechnology, Vol 81, Springer Verlag, Berlin.

Amon, T. et al. (2007). Methane production trough anaerobic digestion of various energy crops grown in sustainable crop rotations. Bioresource Technology, vol. 98, No 17, pp. 3204-3212.

Boone, D. R., Chynoweth, D. P., Mah, R. A., Smith, P. H., Wilkie, A. C. (1993). Ecology and microbiology of biogasification. Biomass and Bioenergy, No. 3-4, vol. 5, pp. 191-202, 1993.

Froment, F. & Bischotf, K. B. (1990). *Chemical Reactor Analysis and Design,* 2nd ed., Wiley, New York.

Klein, Donald W.; Prescott, Lansing M.; Harley, John (2005). Microbiology. New York: McGraw-Hill. ISBN 0-07-255678-1.

Kodriková, K. (2004). Evaluation of amaranth components and research of the polysaccharide component fermentation to fractional fission products. Thesis. UTB Zlín.

Kolomazník, K. & Kodrikova, K. (2008). Biomass dry fermentation and sorted biodegradable waste (in Czech), *Project partial report MPO ČR FI-IMT/183*.

Nickolas J, Themelis S, Verma S. (2004). Anaerobic digestion of organic waste in MSW. Waste Management World. pp. 41–47.

Schausberger, P., Bosch, P., Friedl, A. (2008). Modeling and simulation of coupled ethanol and biogas production. In: Proc. of 11th Conference on Process Integration, Modelling and Optimisation for Energy Saving and Pollution Reduction, Prague, pp. 163-170.

Smith, J. M. (1998). *Chemical Engineering Kinetics.* McGRAW-HILL, New York. p. 131.

Straka, F. et al. (2003). Biogas (in Czech). In: Použití bioplynu v podmínkách ČR. Říčany. ISBN 80-7328-029-9

Swift, G. (1998). Requirements for biodegradable water-soluble polymers, *Polymer degradation and stability*, vol.59, 19-24.

Vogel, T., Ahlhaus, M., Barz, M. (2009). Optimisation of biogas production from grass by dry-wet fermentation. In: Proc. of 8th International Scientific Conference on Engineering for Rural Development, MAY 28-29, 2009 Jelgava, LATVIA, pp. 21-26.

AUTHOR BIOGRAPHIES

ZDENKA PROKOPOVÁ was born in Rimavská Sobota, Slovak Republic. She graduated from Slovak Technical University in 1988, with a master's degree in automatic control theory. Doctor's degree she has received in technical cybernetics in 1993 from the same university. Since 1995 she is working in Tomas Bata University in Zlin, Faculty of Applied Informatics. She is now working there as an associating professor. Research activities: mathematical modeling, simulation, control of technological systems, programming and application of database systems. Her e-mail address is : prokopova@fai.utb.cz.

ROMAN PROKOP was born in Hodonin, Czech Republic in 1952. He graduated in Cybernetics from the Czech Technical University in Prague in 1976. He received post graduate diploma in 1983 from the Slovak Technical University. Since 1995 he has been at Tomas Bata University in Zlín, where he presently holds the position of full professor of the Department of Automation and Control Engineering and a vice-rector of the university. His research activities include algebraic methods in control theory, robust and adaptive control, autotuning and optimization techniques. His e-mail address is : prokop@fai.utb.cz.

OPTIMIZATION AND CONTROL OF A DYNAMICAL PROCESS BY GENETIC ALGORITHM

Tran Trong Dao
Faculty of Information Technology & Applied Mathematics, Ton Duc Thang University
98 Ngo Tat To St.,Ward 19, Binh Thanh Dist, Ho Chi Minh City,Vietnam.
E-mail: trantrongdao@tut.edu.vn

KEYWORDS

Simulation; Optimization; Genetic algorithm;
Continuous stirred tank reactor (CSTR).

ABSTRACT

In this paper, the modeling of a dynamic chemical engineering process is presented in a highly understandable way using a unique combination of the simplified fundamental theory and direct hands-on computer simulation. The main aim is to use them for analysis, optimise and adaptive control of behavior of dynamical system, especially of a given chemical reactor. A non-linear mathematical model is required to describe the dynamic behaviour of a continuous stirred tank reactor (CSTR). Evolutionary algorithm from the field of artificial intelligent - Genetic algorithm (GA) is used in this investigation.

The optimizations and control of a chemical reactor processes have been performed in several ways, each one for a different set of reactor parameters or different cost function. The optimized and adaptive control chemical reactor processes were used in simulations with optimization by genetic algorithm and the results are presented in graphs. Finally, experimental results are reported, followed by conclusion.

INTRODUCTION

Continuous stirred tank reactors (CSTRs) belong to a class of nonlinear systems where both steady-state and dynamic behaviour are nonlinear. Their models are derived and described in e.g. (Ogunnaike and Ray, 1994), (Schmidt, 2005) and (Corriou, 2004). verification can be found in (Stericker and Sinha, 1993).

Chemical process control requires intelligent monitoring due to the dynamic nature of the chemical reactions and the non-linear functional relationship between he input and output variables involved. CSTR is one of the major processing unit in many chemical, pharmaceutical and petroleum industries as well as in environmental and waste management engineering. In spite of continuing advances in optimal solution techniques for optimization and control problems, many of such problems remain too complex to be solved by the known techniques (Emuoyibofarhe O.Justice, Reju A Sunday, 2008).

In chemical engineering, evolutionary optimization has been applied by the author and others to system identification (Pham and Coulter, 1995); a model of a process is built and its numerical parameters are found by error minimization against experimental data. Evolutionary optimization has been widely applied to the evolution of neural networks models for use in control applications (e.g. Li & Haubler, 1996).

The optimization of dynamic process has received growing attention in recent years because it is essential for the process industry to strive for more efficient and agile manufacturing in face of saturated market and global competition (T. Backx, O. Bosgra 2000).

Evolutionary algorithms such as evolution strategies and genetic algorithms have become the method of choice for optimization problems that are too complex to be solved using deterministic techniques such as linear programming or gradient (Jacobian) methods. The large number of applications (Beasley (1997)) and the continuously growing interest in this field are due to several advantages of EAs compared to gradient based methods for complex problems (Ivo F. Sbalzarini, Sibylle Muller and Petros Koumoutsakos 2000).

Designing optimal reactor parameters including control constitutes is one of the most complex tasks in process engineering. The situation is particularly complicated by the fact that the precise mechanism of chemical reaction kinetics is very often unknown. For this reason it is necessary to carry out extensive measurements of input and output concentration dependencies of components on time, temperature, etc.

In this work, the methods of artificial intelligence by evolutionary algorithm GA is presented for optimizing chemical engineering processes, particularly those in which the genetic algorithm is used for static optimization and adaptive control of a chemical CSTR reactor.

MATHEMATICAL PROBLEMS

Consider a CSTR with the first order consecutive exothermic reaction according to the scheme $A \xrightarrow{k_1} B \xrightarrow{k_2} C$ and with a perfectly mixed cooling jacket. Using the usual simplifications, the model of the CSTR is described by four nonlinear differential equations (see Dostál, P., Gazdoš, F., Bobál, V., Vojtěšek, J., 2007).

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali
David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

$$\frac{dc_A}{dt} = -\left(\frac{Q_r}{V_r} + k_1\right)c_A + \frac{Q_r}{V_r}c_{Ai} \quad (1)$$

$$\frac{dc_B}{dt} = -\left(\frac{Q_r}{V_r} + k_2\right)c_B + k_1 c_A + \frac{Q_r}{V_r}c_{Bi} \quad (2)$$

$$\frac{dT_r}{dt} = \frac{h_r}{(\rho c_p)_r} + \frac{Q_r}{V_r}(T_{ri} - T_r) + \frac{A_h U}{V_r(\rho c_p)_r}(T_c - T_r) \quad (3)$$

$$\frac{dT_c}{dt} = \frac{Q_c}{V_c}(T_{ci} - T_c) + \frac{A_h U}{V_c(\rho c_p)_c}(T_r - T_c) \quad (4)$$

with initial conditions $c_A(0) = c_A^s$, $c_B(0) = c_B^s$, $T_r(0) = T_r^s$ and $T_c(0) = T_c^s$. Here, t is the time, c are concentrations, T are temperatures, V are volumes, ρ are densities, c_p are specific heat capacities, Q are volumetric flow rates, A_h is the heat exchange surface area and U is the heat transfer coefficient. The subscripts are denoted $(\cdot)_r$ for the reactant mixture, $(\cdot)_c$ for the coolant, $(\cdot)_i$ for feed (inlet) values and the superscript $(\cdot)^s$ for steady-state values. The reaction rates and the reaction heat are expressed as

$$k_j = k_{0j}\exp\left(\frac{-E_j}{RT_r}\right), j = 1,2 \quad (5)$$

$$h_r = h_1 k_1 c_A + h_2 k_2 c_B \quad (6)$$

where k_0 are pre-exponential factors, E are activation energies and h are reaction entalpies. The values of all parameters, feed values and steady-state values are given in Tab 1. (Dostál, P., Gazdoš, F., Bobál, V., Vojtěšek, J., 2007).

Tab 1. Parameters, inlet values and initial conditions.

$V_r = 1.2$ m^3	$Q_r = 0.08$ m^{3}min^{-1}
$V_c = 0.64$ m^3	$Q_c^s = 0.03$ m^{3}min^{-1}
$\rho_r = 985$ kg m^{-3}	$c_{pr} = 4.05$ kJ kg^{-1}K^{-1}
$\rho_c = 998$ kg m^{-3}	$c_{pc} = 4.18$ kJ kg^{-1}K^{-1}
$A = 5.5$ m^2	$U = 43.5$ kJ m^{-2}min^{-1}K^{-1}
$k_{10} = 5.616 \cdot 10^{16}$ min^{-1}	$E_1/R = 13477$ K
$k_{20} = 1.128 \cdot 10^{18}$ min^{-1}	$E_2/R = 15290$ K
$h_1 = 4.8 \cdot 10^4$ kJ kmol^{-1}	$h_2 = 2.2 \cdot 10^4$ kJ kmol^{-1}
$c_{Ai} = 2.85$ kmol m^{-3}	$c_{Bi} = 0$ kmol m^{-3}
$T_i = 323$ K	$T_{ci} = 293$ K
$c_A^s = 0.1649$ kmol m^{-3}	$c_b^s = 0.9435$ kmol m^{-3}
$T_r^s = 350.19$ K	$T_c^s = 330.55$ K

STATIC OPTIMIZATION REACTOR

In this model of CSTR the parameters were optimized: the parameters of volumetric flow rates of the reactant mixture and the coolant Q_r, Q_c, the parameter of concentration for feed values c_{Ai} and temperature reactant mixture and colant T_{ri}, T_{ci}.(see Tab. 2)

Tab 2. Parameters of reactor, "highlight color" were optimized

$V_r \rightarrow \text{Meter}^3$	$V_c \rightarrow \text{Meter}^3$
$A_r \rightarrow \text{Meter}^2$	$Q_r \rightarrow \frac{\text{Meter}^3}{\text{min}}$
$c_{pc} \rightarrow \frac{\text{KiloJoule}}{\text{KilogramKelvin}}$	$U \rightarrow \frac{\text{KiloJoule}}{\text{Meter}^2\ \text{min Kelvin}}$
$h_1 \rightarrow \frac{\text{KiloJoule}}{\text{KiloMole}}$	$h_2 \rightarrow \frac{\text{KiloJoule}}{\text{KiloMole}}$
$c_{Ai} \rightarrow \frac{\text{KiloMole}}{\text{Meter}^3}$	$c_{Bi} \rightarrow \frac{\text{KiloMole}}{\text{Meter}^3}$
$\rho_r \rightarrow \frac{\text{Kilogram}}{\text{Meter}^3}$	$\rho_c \rightarrow \frac{\text{Kilogram}}{\text{Meter}^3}$
$Q_c \rightarrow \frac{\text{Meter}^3}{\text{min}}$	$c_{pr} \rightarrow \frac{\text{KiloJoule}}{\text{KilogramKelvin}}$
$\frac{E_1}{R} \rightarrow \text{Kelvin}$	$\frac{E_2}{R} \rightarrow \text{Kelvin}$
$k_{10} \rightarrow \frac{1}{\text{min}}$	$k_{20} \rightarrow \frac{1}{\text{min}}$
$T_{ri} \rightarrow \text{Kelvin}$	$T_{ci} \rightarrow \text{Kelvin}$
$T_r^s \rightarrow \text{Kelvin}$	$T_c^s \rightarrow \text{Kelvin}$

The Cost Function (CF)

In this research, the objective used to minimize the area arising as a difference of the process between the observed and real selected time interval, which was the duration of a CSTRs cycle. With the inlet concentration $c_{Bi} = 0$ kmol m^{-3}, the cost function, that was minimized is given in (7). In the cost function, we multiplied by (-1) in order to transfer from maximization into minimization.

$$f_{\cos t} = (-1) * \sum_{t=0}^{t} |c_B[t]| \quad (7)$$

Genetic Algorithm

Genetic Algorithms (GA) imitate the evolutionary processes with emphasis on genotype based operators (genotype/phenotype dualism). The GA works on a population of artificial chromosomes, referred to as individuals. Each individual is represented by a string of L bits. Each segment of this string corresponds to a variable of the optimizing problem in a binary encoded form.

The population is evolved in the optimization process mainly by crossover operations. This operation recombines the bit strings of individuals in the population with a certain probability *Pc*. Mutation is secondarily in most applications of a GA. It is responsible to ensure that some bits are changed, thus allowing the GA to explore the complete search space

even if necessary alleles are temporarily lost due to convergence.

The following pseudocode describes the general principle of a genetic algorithm (see JCell Documentation of A. Zell, http://www.ra.cs.uni-tuebingen.de/software/JCell/tutorial/tutorial.html):

```
t = 0;
initialize(P(t=0));
evaluate(P(t=0));
while is NotTerminated() do
P_p (t) = P(t).selectParent();
P_c(t) = reproduction(P_p);
mutace(P_c(t));
evaluate(P_c(t));
P(t+1) = buildNextGenerationForm(P_c(t), P(t));
t=t+1;
end
```

Figure 1. Pseudocode of GA

Parameter Setting

The control parameter settings have been found empirically and are given in Tab. 3. The main criterion for this setting was to keep the same setting of parameters as much as possible and of course the same number of cost function evaluations as well as population size (parameter PopSize). Number of optimized reactor parameters and their range inside represents in Tab. 4.

Tab. 3 GA parameter setting

	A
PopSize	20
MutationCostant	0.2
Generations	200
Individual Length	6
CF Evaluations	4000

Tab. 4 Optimized reactor parameters and their range inside which has been optimization done

Parameter	Range
Q_c [m^3min^{-1}]	0.015 – 0.1
Q_r [m^3min^{-1}]	0.05 – 0.012
c_{Ai} [kmol m^{-3}]	2 – 3.5
T_{ri} [K]	303 – 333
T_{ci} [K]	288-303

EXPERIMENTAL RESULTS OF STATIC OPTIMIZATION REACTOR

Evolutionary algorithms (in this work using genetic algorithm) are partly of stochastic nature, a large set of simulations has to be done in order to get data for statistical data processing. The algorithms GA have been applied 100 times in order to find the optimum of process parameters and. All important data has been visualized directly or/and processed for graphs demonstrating performance of this algorithms. Estimated parameters and their diversity (minimum, maximum and average) are depicted (see Tab. 5). From Fig. 2 it is visible that results from GA algorithm are showed detail "optimal points". For the demonstration are graphically the solutions shown in Fig. 3- 6.

Tab. 5 Estimated parameters for GA

Parameter	Min	Avg	Max
Q_c [m^3min^{-1}]	0.0152769	0.0706135	0.0994583
Q_r [m^3min^{-1}]	0.105348	0.115416	0.119818
c_{Ai} [kmol m^{-3}]	2.00811	2.69678	3.49589
T_{ri} [K]	314.177	320.645	324.491
T_{ci} [K]	290.091	299.845	302.908
Q_c [m^3min^{-1}]	0.0152769	0.0706135	0.0994583

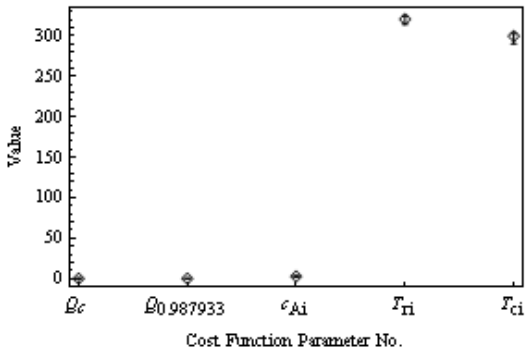


Figure 2. Parameter variation

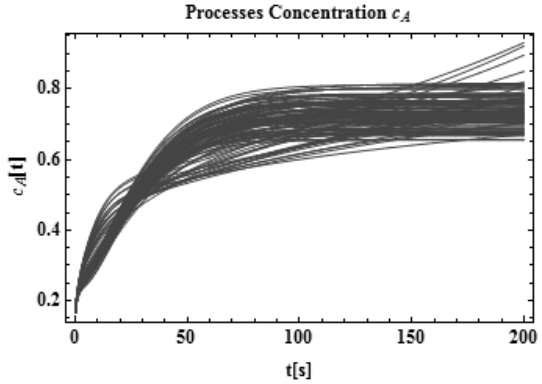


Figure 3. 100 simulations for c_A

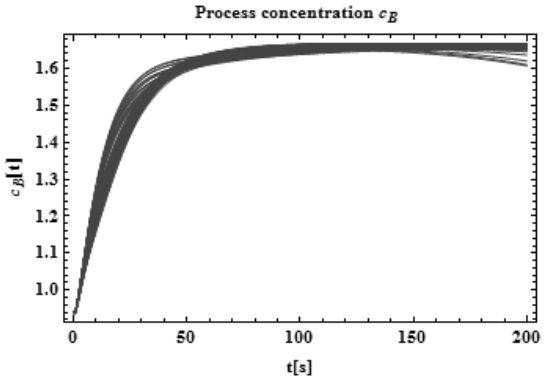


Figure 4. 100 simulations for c_B

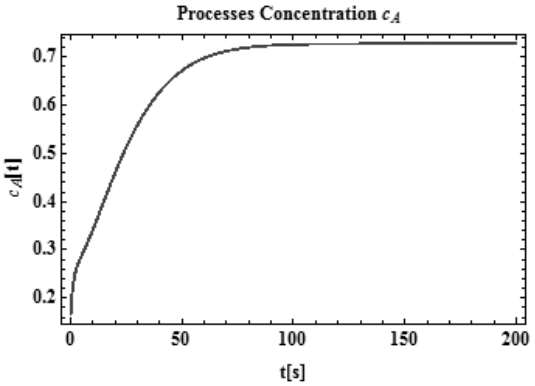


Figure 7. Best solution for c_A

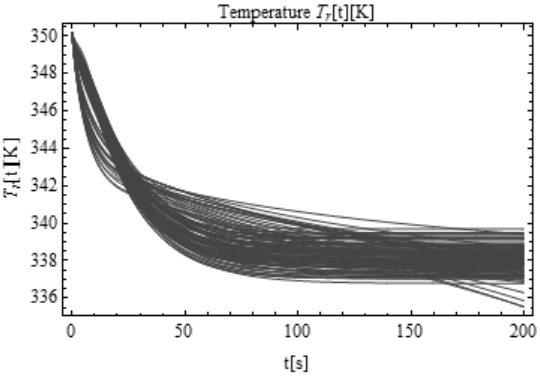


Figure 5. 100 simulations for T_r

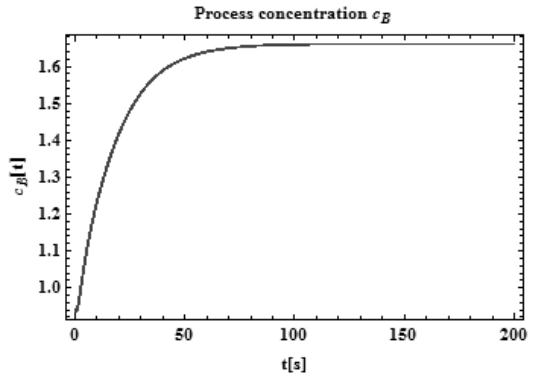


Figure 8. Best solution *for* c_B

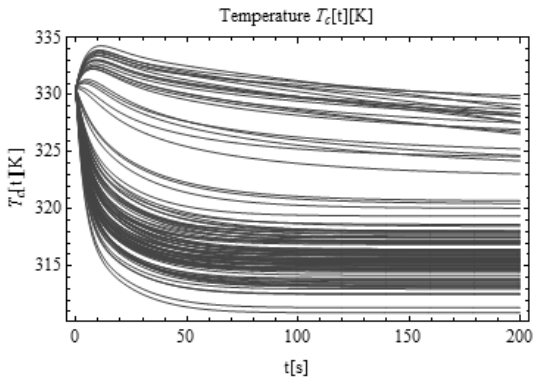


Figure 6. 100 simulations for T_c

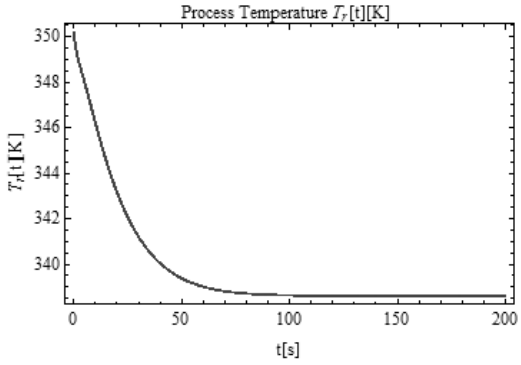


Figure 9. Best solution for T_r

From investigation on optimization of process parameter of CSTR we can see, that evolutionary algorithm GA has presented diversity of parameters (100 lines in diagram are processes of 100 simulation repeated). On optimization of GA, it is evident that the courses of algorithm are densities in a thin. Alongside it, sometime few values drift out of the actual solution. But by the repetition of simulation was recorded the best result. On Fig. 7 – Fig. 10, the processes of parameters by GA algorithm obtained best solutions for the optimization.

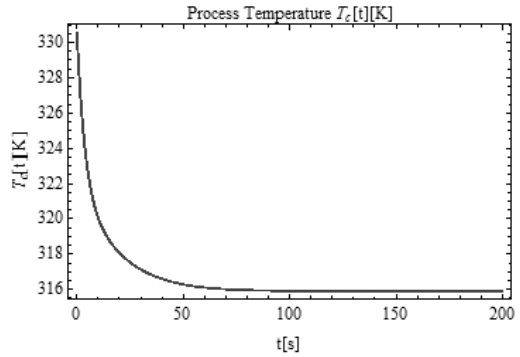


Figure 10. Best solution for T_c

OPTIMIZATION REACTOR WITH ADAPTIVE CONTROL

Adaptive control by mean evolutionary algorithms is very robust method, particular in system with many disturbing effects and failures. It's also a powerful tool in the search for optimal solutions to very complex problem in the field of control process. The basic idea is to find a set of action that lead to the principle optimization with required value. The block of adaptive control is shown in Fig. 11. (Tran, T.D., 2009).

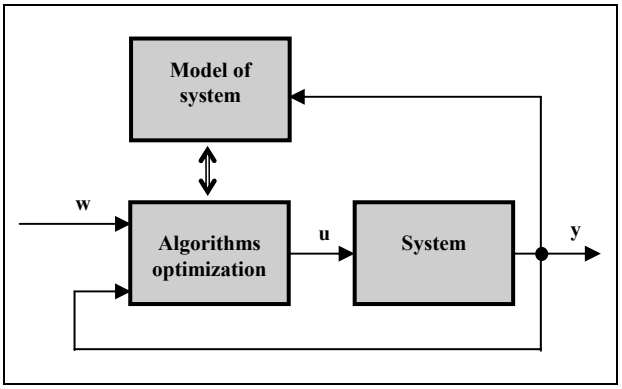


Figure 11. Principle of adaptive control by evolutionary algorithm

In block „Algorithms optimization“ are evolutionary algorithm (here is genetic algorithm), the adaptive control is selected by minimising the commonly cost function *J*:

$$J(N_1,N_2,N_u)=\sum_{j=N_1}^{N_2}[y(k+j)-w(k+j)]^2 + \sum_{j=1}^{N_u}\lambda(j)[\Delta u(k+j-1)]^2 \quad (8)$$

Here *y* is the output of system, *u* is actuating signal, *w* is the controlled value, Δu is the control value change, *k* is the control step, N_1 is the lower and N_2 is the upper output prediction horizon, *Nu* is control horizon and λ is a weight sequence control of action.

Block "system" is a control process and block "model system" is used to adaptive its behaviour, it often represented in the appropriate algorithms of artificial intelligent, commonly artificial neural network.

Adaptive control proceed when change of required value running optimization algorithms in conjunction with model of system and minimising cost function (8) is found optimal action, which is for chosen system.

This optimization was minimized the area arising as a difference between the required and real temperature profile of the reaction mixture in a selected time interval, which was the duration of a CSTR cycle. The cost function was minimized is given in (9) for T_c temperature and (10) for T_r.

$$f_{\cos t}=\sum_{t=0}^{t}\left|w_1-T_c[t]\right| \quad (9)$$

$$f_{\cos t}=\sum_{t=0}^{t}\left|w_2-T_r[t]\right| \quad (10)$$

Where w_1, w_2…are required values (control point)

For static optimization of CSTR reactor with adaptive control, we have added required value for simulation of temperature T_r and T_c belong following below Tab. 6. The range inside of temperature T_r and T_c for adaptive control is <273.;380> [K] (see Tab. 7).

Tab. 6 Range inside for adaptive control of CSTR

Parameter	Range
T_r [K]	273 – 380
T_c [K]	273 – 380

Tab.7 Parameters setting for adaptive control

Time simulation[s]	**Required value *for* T_r [K]**	**Required value *for* T_c [K]**
0 - 80	360	340
80 - 150	340	320
150 - 200	320	300
200 - 300	370	370

Simulation

Simulations were conducted so, that the first minimising cost function using the adaptive horizon is found within the optimal action and intervention that was held for the control horizon. After that, it was calculated new intervention and has been applied during control horizon etc. till to do filling of reactor. For the CSTR reactor, required values were selected: N_1 = 1, N_2 =300 a N_u = till to change.

EXPERIMENTAL RESULTS OF OPTIMIZATION REACTOR WITH ADAPTIVE CONTROL

On investigation of adaptive control chemical reactor CSTR, at the same principle setting parameter of GA and the algorithm have been applied 100 times. The evolutionary processes of temperature that control by GA algorithm show in follow graphs from Fig. 12-13.

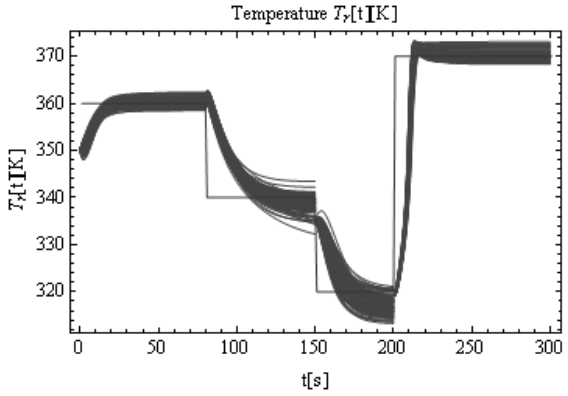


Figure 12. Evolutionary process for T_r

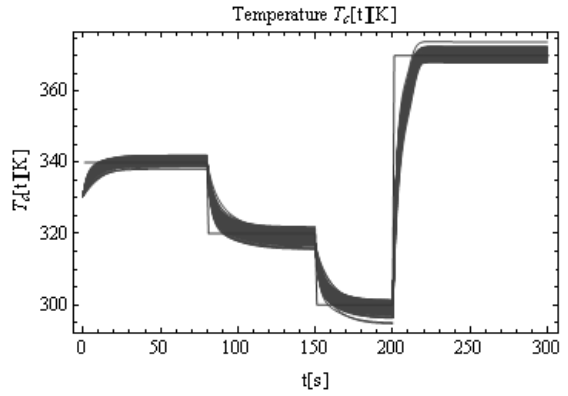


Figure 13. Evolutionary process for T_c

CONCLUSION

This work was performed static optimization and adaptive control on non-linear system using genetic algorithms. Based on these results it can be stated as follows:

- Genetic algorithm is used successfully to static optimise and adaptive control of a chemical reactor process.
- Calculation was 100 times repeated and the best, worst and average result (individual) was recorded from the last population in each simulation. All one hundred triplets (best, worst, average) were used to create Tab. 5.
- From the graphs, we have chosen the best solution of process parameters. Alongside it, sometime few values drift out of the actual solution. From Fig. 3 - Fig. 6 we have defined the cost function by concentration c_B and it obtained optimal value for the optimization of process parameters CSTR. On the pictures of Fig. 12 and Fig. 13 shown the evolutionary processes of temperatures T_r and T_c of reactor by definition of the cost function in (9) & (10).
- The optimizations and control chemical reactor have been performed in several ways, each one for a different set of reactor parameters or different cost function. From the obtained results, it is possible to say that all simulations give satisfactory results and thus genetic algorithm is capable of solving this class of difficult problems and the quality of results does not depend only on the problem being solved but they are extremely sensitive on the proper definition of the cost function, selection of parameters setting of algorithm.

From this work GA have shown good potential and ability to solve complex problems of optimization, not only at the filed of chemical engineering process but also in diverse industrial fields.

ACKNOWLEDGEMENT

The author would like to thanks Associate Professor. Ivan Zelinka for his offer advise on evolutionary computation.
This work was supported at the Ton Duc Thang University, Ho Chi Minh, Vietnam.

REFERENCES

Babu, B.V. and R. Angira (2001a). Optimization of Non-linear functions using Evolutionary Computation. Proceedings of 12. ISME Conference, India, January 10−12, 153-157 (2001).

Back, T., Fogel, D.B., Michalewicz, Z., 1997. Handbook of Evolutionary Computation. Institute of Physics, London.

Bhaskar et al., 2000. Applications of multiobjective optimization in chemical engineering. Reviews in Chemical Engineering. v16 i1.

Beyer, H.-G., 2001. Theory of Evolution Strategies. Springer, New York. Cerny, V., 1985. Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm. Journal of Optimization Theory and Applications 45 (1), 41–51.

Beasley, D. 1997 Possible applications of evolutionary computation. In B¨ack, T., Fogel, D. B. &Michalewicz, Z., editors, Handbook of Evolutionary Computation. 97/1, A1.2, IOP Publishing Ltd. and Oxford University Press.

Bruce Nauman, E. Chemical Reactor Design, Optimization, and Scale up, IBSN 0-07-137753-0.

Corriou, J.P. (2004). Process control. Theory and applications. Springer-Verlag, London.

Dostál, P., V. Bobál and M. Blaha (2001). One approach to adaptive control of nonlinear processes. In: Proc. IFAC Workshop on Adaptation and Learning in Control and Signal Processing ALCOSP 2001, Cernobbio-Como, Italy, 407-412.

Dostál, P., Gazdoš, F., Bobál, V., Vojtěšek, J.: Adaptive control of a continuous stirred tank reactor By two feedback controllers. In: 9th IFAC Workshop Adaption and Learning in Kontrol and Sinal Processing ALCOSP'2007, Saint Petersburg, Russia, August 29-31, 2007, P5-1 – P5-6.

Emuoyibofarhe O.Justice., Reju A Sunday, An optimum solution for a process control problem (continuous stirred tank reactor) using a hybrid neural network. Journal of Theoretical and Applied Information

Technology. Volume 4, Number 10, Pages: 906-915. 31 October 2008.
F. Glover, M. Laguna and R. Martí. Advances in Evolutionary Computation: Theory and Applications, A. Ghosh and S. Tsutsui (eds.), Springer-Verlag, New York, pp. 519-537 (2003).
Glover F., Laguna M. a Martí R., Scatter Search in Ghosh A. a Tsutsui S. (Eds.), Advances in Evolutionary Computation: Theory and Applications, Springer-Verlag, New York, pp. 519-537, 2003.
Gross B.; Roosen P. Total process optimization in chemical engineering with evolutionary algorithms. Computers & Chemical Engineering. Volume 22, Supplement 1, 15 March 1998, Pages S229-S236. European Symposium on Computer Aided Process Engineering-8.
Hanne, T. Global multiobjective optimization using evolutionary algorithms. Journal of Heuristics, 2000.
Ivo F. Sbalzarini, Sibylle Muller and Petros Koumoutsakos 2000. Multiobjective optimization using evolutionary algorithms. Center for Turbulence Research, Proceedings of the Summer Program 2000.
Ingham, J., Dunn, I.J., Heinzle, E., J.E.P 2000. Chemical Engineering Dynamic, ISBN 3-527-29776-6.
Kirkpatrick, S. Gelatt, CD. Vecchi, MP. Optimization by simulated annealing -Page 1. 13 May 1983, Volume 220, Number 4598.
Levenspiel O. Chemical reaction engineering. New York: John Wiley and Sons, 1962. 578 p.
Li, Y. and Haubler, A. Artificial evolution of neural networks and its application to feedback control. Artificial Intelligence in Engineering 10, 143-152 (1996).
Liu et al., 2007 Li Liu, Wenxin Liu and David A. Cartes, Particle swarm optimization-based parameter identification applied to permanent magnet synchronous motors, Engineering Applications of Artificial Intelligence (2007) 10.1016/j.engappai.2007.10.002.
Ogunnaike, B.A. and W.H. Ray (1994). Process dynamics, modelling, and control. Oxford University Press, New York.
Pham, Q.T. Dynamic optimization of chemical engineering processes by evolutionary method (2005).
Pham, Q.T. and Coulter, S. Modelling the chilling of pig carcasses using an evolutionary method. Proc. Int. Congress of Refrig. Vol. 3a pp.676-683 (1995).
Stericker, D.L. and N.K. Sinha (1993). Identification of continuous-time systems from samples of input-output data using the δ-operator. Control-Theory and Advanced Technology, 9, 113-125.
Schmidt, L.D. (2005). The engineering of chemical reactions. Oxford University Press, New York.
Smith, R. Chemical Process, Design and Integration 2005, ISBN 0-471-48681-7.
Srinisavan B., Palanki S., Bonvin D. (2002). Dynamic optimization of batch processes II. Role of Measurement in handling uncertainly. Computers and Chemical Engineering, Vol. 27, p. 27-44.
Srinivasan et al., 2003 B. Srinivasan, S. Palanki and D. Bonvin, Dynamic optimization of batch processes: I. Characterization of the nominal solution, Computers & Chemical Engineering 27 (2003), pp. 1–26.
Stefanis et al., (1997). Environmental impact considerations in the optimal design and scheduling of batch processes. Computers and Chemical Engineering. v21 i10. 1073-1094.
T. Backx, O. Bosgra, W. Marquardt, Integration of model predictive control and optimization of processes, in: Proceedings ADCHEM 2000, vol. 1, 2000, pp. 249–260.
Thomas Bäck, Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms, Oxford University Press, Oxford, 1996Thomas Hanne, Global Multiobjective Optimization Using Evolutionary Algorithms , Journal of Heuristics, 6: 347–360 (2000).
Tran, T.D., The evolutionary computation techniques in chemical engineering. Thesis, 2009, Faculty of Applied Informatics, Tomas Bata University in Zlín.
Wolpert, D.H., Macready, W.G. (1997), "No Free Lunch Theorems for Optimization," IEEE Transactions on Evolutionary Computation. Pages 68-72, Volume 1.

AUTHORS BIOGRAPHIES

TRAN TRONG DAO was born in Vietnam, and went to the VSB-Technical university of Ostrava in 2001, where he studied Automatic Control and Engineering Informatics and obtained his degree in 2006. He obtained Ph.D. degree in Technical Cybernetics in 2009 at Tomas Bata University in Zlin. Now he is a lecturer at the Faculty of Information Technology & Applied Mathematics, Ton Duc Thang University, Ho Chi Minh city, Vietnam.

ENHANCING FUZZY INFERENCE SYSTEM BASED CRITERION-REFERENCED ASSESSMENT WITH AN APPLICATION

Kai Meng Tay
Electronic Engineering Department
Faculty of Engineering,
University Malaysia Sarawak
kmtay@feng.unimas.my

Chee Peng Lim
School of Electrical & Electronic Engineering
University of Science Malaysia, Malaysia
cplim@eng.usm.com

Tze Ling Jee
Electronic Engineering Department
Faculty of Engineering,
University Malaysia Sarawak
jessie_jtl@yahoo.co.uk

KEYWORDS
Fuzzy inference system, education assessment, criterion-referenced assessment, monotonicity property.

ABSTRACT

An important and difficult issue in designing a Fuzzy Inference System (FIS) is the specification of fuzzy sets, and fuzzy rules. The aim of this paper is to demonstrate how an additional qualitative information, i.e., monotonicity property, can be exploited and extended to be part of an FIS designing procedure (i.e., fuzzy sets and fuzzy rules design). In this paper, the FIS is employed as an alternative to the use of addition in aggregating the scores from test items/tasks in a Criterion-Referenced Assessment (CRA) model. In order to preserve the monotonicity property, the *sufficient conditions* of the FIS is proposed. Our proposed FIS based CRA procedure can be viewed as an enhancement for the FIS based CRA procedure, where monotonicity property is preserved. We demonstrate the applicability of the proposed approach with a case study related to a laboratory project assessment task at a university, and the results indicate the usefulness of the proposed approach in the CRA domain.

INTRODUCTION

Educational assessment is a process of forming judgment about quality and extent of students' achievement or performance, and, therefore, by inference a judgment about the learning that has taken place. Judgment usually is based on information obtained by requiring students to attempt some specified tasks, and to submit their work for an appraisal of its quality. Scoring refers to the process of representing students' achievement by numbers or symbols.

With respect to criterion-referenced assessment (CRA), ideally, students' grade should be determined by comparing their achievements with a set of clearly stated criteria for learning outcomes and standards for some particular levels of performance. The aim of CRA is to report students' achievement with reference to a set of objective reference points. It can be a simple pass-fail grading schema, or a single grade or percentage (Sadler, 2005). From the literature, the use of CRA in essay writing (While, 2002) clinical performance (Nicholson *et al*., 2009) have been reported.

Scoring usually refers to test items/tasks rather than to the overall achievement (Sadler, 2005, White, 2002, Nicholson *et al*., 2009, Joughin, 2008). To ease the assessment process, in common practice, a score is given to each item or task, with the use of rubric. Scores are then aggregated to produce a final score. Scores from different test items/tasks are usually added together and then projected (Sadler, 2005). A score can be weighted before being added to reflect the relative importance of each task (Sadler, 2005).

The use of fuzzy set related techniques in education assessment models is not new. Biswas (1995) presented a fuzzy set related method to evaluate students' answer scripts. This work was further enhanced by Chen and Lee (1999). Ma and Zhou (2000) presented a fuzzy set related method to assess student-centred learning. Saliu (2005) suggested the use of the FIS in CRA, as a Constrained Qualitative Assessment (CQA) method, with a case study.

In this paper, an FIS-based CRA model is explained. The model can be viewed as an alternative to the use of addition in aggregating the scores from all test items/tasks, and to produce a final score. The idea of replacing simple or weighted addition with a more complicated algorithm is not new (Sadler, 2005). It is pointed out that aggregation of scores can be done by some dedicated algorithm or mathematical equation. The FIS is used owing to several reasons. First, the criteria in rubric can be qualitative rather than quantitative (Sadler, 2005). As an example, a score of 4 in a rubric does not mean two times better than that of a score of 2. The FIS acts as a solution to qualitative assessment, and keeps qualitative assessment accountable. Second, the relative importance of each task can be different. The importance of each task

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

depends on the learning outcomes. Third, an FIS can be used as an alternative approach to model or to customize the relationship between the score of each task and the aggregated score.

With respect to the FIS, it can be viewed as a method to construct a multi-input, non-linear model in an easy manner (Jang, *et al.*, 1997). In this work, our investigation focuses on the monotonicity property of an FIS-based CRA model. The importance of the monotonicity property in FIS-based CRA has been pointed out by Saliu (2005). It was suggested that the failure of an FIS-based CRA model to fulfil monotonicity property is an anomaly, and effort should be put to overcome this problem. However, there are relatively few articles addressing the problem of designing monotonic FIS (Kouikoglou and Phillis, 2009). Noted that the importance of the monotonicity property in other assessment and selection problems has been highlighted in Kouikoglou and Phillis (2009), Broekhoven and Baets (2008, 2009).

In this paper, the monotonicity property of an FIS and the *sufficient conditions* for the FIS to be of monotonicity, as pointed in Kouikoglou and Phillis (2009) as well as Tay and Lim (2008a, 2008b), are reviewed. An FIS-based CRA model is then presented. The monotonicity property for the FIS-based CRA model is defined. The *sufficient conditions* for the FIS to be of monotonicity is applied to CRA. The applicability of our proposed approach is demonstrated with a case study related to laboratory project at Universiti Malaysia Sarawak, Malaysia.

THE PROPOSED FIS-BASED CRA MODEL

Figure 1 depicts a flow chart of our proposed FIS-based CRA model. Learning usually starts with definition of learning objectives and learning outcomes. From the learning objectives and learning outcomes, test items/tasks and their assessment criteria are designed. Consider a laboratory project with three test items/tasks, i.e., *electronic circuitry design*, *electronic circuitry development*, and *presentation*. Table 1 shows the scoring rubric for *electronic circuitry design*. Each partition of the rubric can be represented by a fuzzy set. Figure 2 depicts the membership functions of *electronic circuitry design*. Each membership function is assigned a *linguistic term*. For example, a score of 3 to 5 is assigned to *Satisfactory*, which refers to *"The circuit is simple (3~4 necessary ICs). Some unnecessary components are included. Able to apply moderately the learned knowledge. Simulate only parts of circuit and briefly explain the circuit operation*". The same is applied to the rubrics of *electronic circuitry development*, and *presentation*.

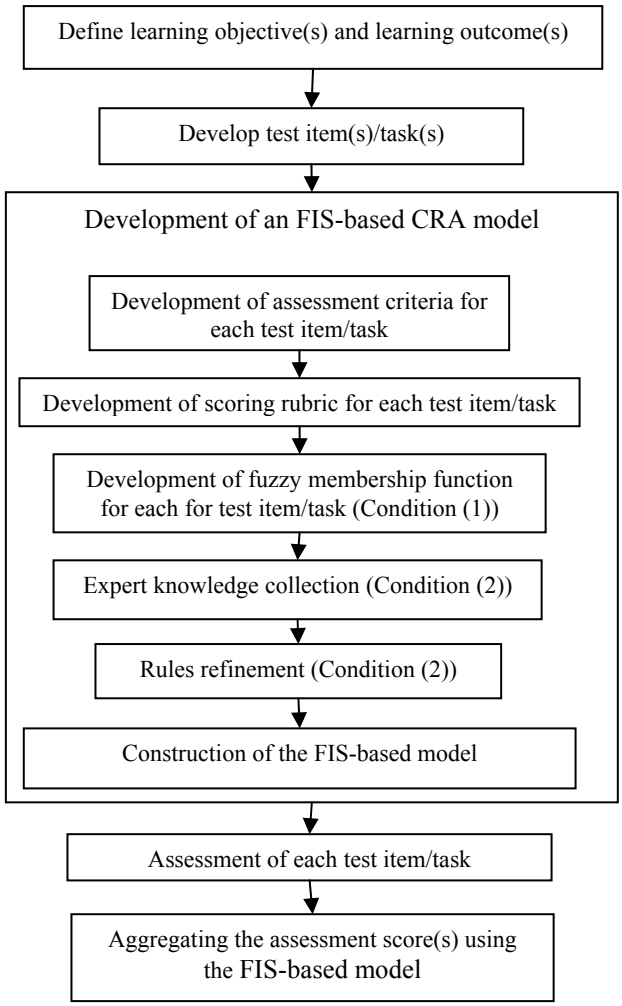


Figure 1**.** The proposed CRA procedure using the fuzzy inference system

In this paper, membership functions of *electronic circuitry design*, *electronic circuitry development*, and *presentation* are labeled as μ_{Dg}^{a}, μ_{Dv}^{b}, and $\mu_{\Pr}^{c}$, respectively. For example, for the test item/task *electronic circuitry design*, a score from 3 to 5 is represented by μ_{Dg}^{2}. The *final score* varies from 1 to 100, and is represented by seven fuzzy membership functions, i.e., "*Excellent*", "*Very good*", "*Good*", "*Fair*", "*Weak*", "*Very weak*" and "*Unsatisfactory*", respectively. The corresponding b scores are assumed to be the point where membership value of B is 1.

Table 1. Scoring Rubric of *Electronic Circuitry Design*

Score	Linguistic Terms	Criteria
10	Excellent	The circuit is complex (≥ 10 necessary ICs). Able to apply knowledge in circuit design. Able to simulate and clearly explain the operation of designed circuit.
9~8	Very good	The circuit is moderate (7~9 necessary ICs). Able to apply most of the learned knowledge. Able to simulate and clearly explain the operation of the circuit.
7~6	Good	The circuit is moderate (5~6 necessary ICs/Components). Some unnecessary components are included. Able to apply most of the learned knowledge. Able to simulate the circuit and briefly explain circuit operation.
5~3	Satisfactory	The circuit is simple (3~4 necessary ICs). Some unnecessary components are included. Able to apply moderately the learned knowledge. Simulate only parts of circuit and briefly explain the circuit operation.
2~1	Unsatisfactory	The circuit is simple (1~2 necessary ICs). Some components are not included and unnecessary components are added. Only able to apply some of the learned knowledge. Unable to simulate and explain the operation of designed circuit.

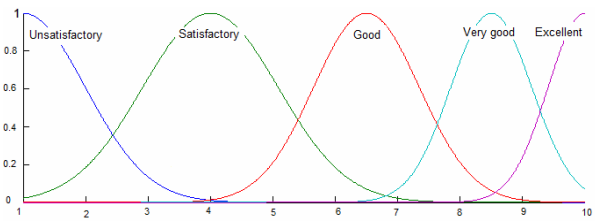


Figure 2. Membership function for one of the test item, i.e. *electronic circuitry design*

A fuzzy rule base is a collection of knowledge in the *If-Then* format from experts. It describes the relationship between *electronic circuitry design*, *electronic circuitry development*, and *presentation* and the *final score*. As an example, Figure 3 shows two rules collected from lecturers who are responsible for the assessment.

Rule 1
If *electronic circuitry design* is **Good** and *electronic circuitry development* is **Good** and *presentation* is **Unsatisfactory** then *Final Score* is **Weak**

Rule 2
If *electronic circuitry design* is **Very good** and *electronic circuitry development* is **Very good** and *presentation* is **Good** then *Final Score* is **Good**

Figure 3 An example of two fuzzy production rules

In this paper, a simplified Mamdani FIS is used to evaluate the *final score*, as shown in Equation (1), which is a zero-order Sugeno FIS model.

$$Final\ score = \frac{\sum_{a=1}^{M_{Dg}} \sum_{b=1}^{M_{Dv}} \sum_{c=1}^{M_{\Pr}} \mu_{Dg}^{a} \times \mu_{Dv}^{b} \times \mu_{\Pr}^{c} \times b^{a,b,c}}{\sum_{a=1}^{M_{Dg}} \sum_{b=1}^{M_{Dv}} \sum_{c=1}^{M_{\Pr}} \mu_{Dg}^{a} \times \mu_{Dv}^{b} \times \mu_{\Pr}^{c}} \quad (1)$$

A REVIEW ON THE *SUFFICIENT CONDTIONS*

If for all x^a and x^b such that $x^a < x^b$, then for a function f to be monotonically increasing or decreasing, the condition $f(x^a) \le f(x^b)$ or $f(x^a) \ge f(x^b)$ must be fulfilled, respectively. From the literature, there are a lot of investigations on the monotonicity property of FIS models. One attempt is to differentiate the output of an FIS with respect to its input(s). Won *et al.* (2002) derived the *sufficient conditions* for the first-order Sugeno fuzzy model with this approach. The *sufficient conditions* for a zero-order Sugeno FIS model to be monotonicity has also been reported (Kouikoglou and Phillis, 2009, Tay and Lim, 2008a, 2008b).

For an FIS to be monotone, the *sufficient conditions* state that two conditions are needed. *Condition (1)* is related to how a membership function should be tuned in order to ensure that the FIS satisfies the monotonicity property. Assume both μ^p and μ^q are differentiate-able. For $\mu^p \prec \mu^q$, Equation (2) has to be fulfilled

$$\frac{\mu^p{}'(x)}{\mu^p(x)} \le \frac{\mu^q{}'(x)}{\mu^q(x)} \quad (2)$$

Assume that the Gaussian membership function, $G(x) = e^{-[x-c]^2/2\sigma^2}$, is used in the FIS-based CRA model. The derivative of $G(x)$ is $G'(x) = -\left((x-c)\right)/\sigma^2 \left)G(x)$. Using Equation (2), the ratio of the Gaussian membership function returns a

linear function, i.e., $E(x) = G'(x)/G(x) = -\left(1/\sigma^2\right)x + \left(c/\sigma^2\right)$. *Condition (2)* highlights the importance of having a monotonic rule base in the FIS model.

AN FIS-BASED CRA MODEL WITH THE *SUFFICIENT CONDITIONS*

The monotonicity property is important to the FIS-based CRA model to allow valid and meaningful comparisons among students' performance to be made. It describes the relationship between a single test item/ task with the aggregated final score. Generally, it is possible to explain the importance of the monotonicity property in CRA with the theoretical properties of a *length function* e.g. monotonicity and sub-additivity (Inder, 2005). For example, if a student obtains a higher score in *electronic circuitry design*, he/she should have a higher *final score*. For two students who are awarded the same scores in *electronic circuitry development* and *presentation*, the student with a higher score in *electronic circuitry design* should not have a lower score than that of the other.

To preserve this property, the *sufficient conditions* is applied to the FIS-based CRA model. *Condition (1)* is used to generate the membership function for *electronic circuitry design*, *electronic circuitry development*, and *presentation*, as illustrated in Figure 1. Figure 2 depicts the membership functions of *electronic circuitry design* that obey the condition. The membership functions of *electronic circuitry design* (as illustarted in Figure 2) can be projected, and $E(x) = G'(x)/G(x)$ allows the membership functions of *electronic circuitry design* to be visualized, as in Figure 4. For example, the membership function of *Excellent* is projected, and its linear line is greater than that of *Very Good* over the universe of discourse. Since $E_{Excellent}(x) > E_{VeryGood}(x) > E_{Good}(x) > E_{Satisfactory}(x) > E_{Unsatisfactory}(x)$, *Condition (1)* is fulfilled.

Condition (2) is used to check the validity of the corrected rule base. If the rule base collected does not fulfill *Condition (2)* , a feedback is sent to lecturer incharge so that rule set that fulfill *Condition (2)* is provided.

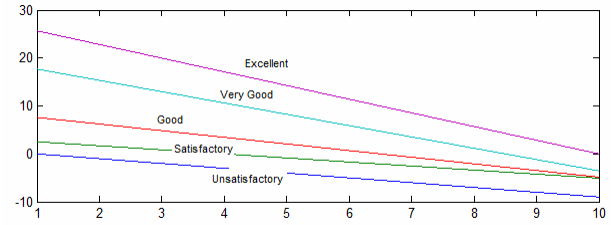


Figure 4 Visualization of the membership functions of *electronic circuit design*

A CASE STUDY

A case study was conducted to evaluate the proposed FIS-based CRA model. Asseemsent of a laboratory project by second year undergraduate students at Universiti Malaysia Sarawak was performed. The students were required to perform three test items/tasks: (1) to design a digital electronic system based on the knowledge learned at their digital system subject, as well as their creativity and technical skills; (2) to develop the system either using a *printed circuit board* or a *breadboard*; (3) to present and demonstrate their work(s).

Table 2 summarizes the assessment results with the FIS-based CRA model. Column "*No.*" shows the label of each student's project. Columns "*Dg*", "*Dv*", and "*Pr*" list the score of each test item/task, respectively. Column "*Final score*" shows the results from the FIS-based CRA model. Column "*Expert's knowledge*" shows the linguistic term associated with each project.

Figure 5 depicts one of the completed projects. The project was given a score of 7 for *electronic circuit design* because it consisted of about five components, and the student was able to explain the operations of the designed system. The student was given a score of 6 for *electronic circuit development* as the system worked well, and all electronic components were installed on the *breadboard* correctly. However, the electronic system was messy. The student was awarded a score of 7 for *project presentation*. The *final score* obtained by the student was 50.0102 (from the FIS-based CRA model).

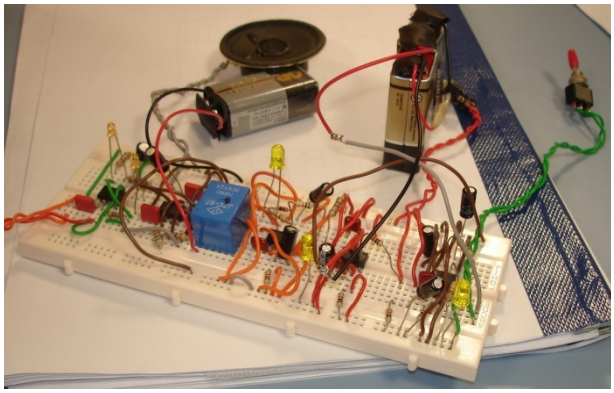


Figure 5 A digital system built by student #8

Table 2 Assessment with FIS based CRA

No.	Score of each task			Assessment with FIS based CRA	
	Dg	*Dv*	*Pr*	*Final score* (%)	Expert's knowledge
					Linguistic term
1	4	4	6	39.9691	Weak
2	5	4	6	40.2218	Weak
3	5	4	7	40.4292	Weak
4	7	4	6	40.4292	Weak
5	5	5	7	42.1384	Weak
6	6	8	5	48.8705	Fair
7	5	7	7	49.2797	Fair
8	7	6	7	50.0102	Fair
9	7	7	6	50.8660	Fair
10	8	6	6	51.1921	Fair
11	7	7	8	63.4142	Good
12	7	9	8	75.3164	Very good
13	8	8	10	78.1755	Very good
14	10	8	8	93.2697	Excellent
15	10	9	8	94.1681	Excellent

From the experiment, the FIS-based CRA model is able to produce an aggregated final score in accordance with expert's knowledge. This can be observed as the *final score* is in agreement with *Expert's knowledge*.

The importance of the monotonicity property can be explained by comparing the performance of students labeled "1" and "2", with scores of 4 4 6 and 5 4 6 ("*Dg*", "*Dv*" and "*Pr*" respectively). Both the students obtained the same score for *Dv*" and "*Pr*. However, student "1" was awarded a lower score (*Dg*=4) than that of student "2" (*Dg*=5) in the design task. The monotonicity property suggests that the final score of student "2" should not be lower than that of student "1", in order to allow a valid comparison of their performance.

From the observation, the FIS-based CRA model is able to fulfill the monotonicity property. There are no illogical predictions found in this case study. Figure 6 depicts a surface plot of the *total score* versus *electronic circuit design* and *project presentation* when *electronic circuit development=5*. An monotonic curve is obtained. In summary, as long as *Condition (1) and Condition (2)* are fulfilled, the monotonicity property can be ensured.

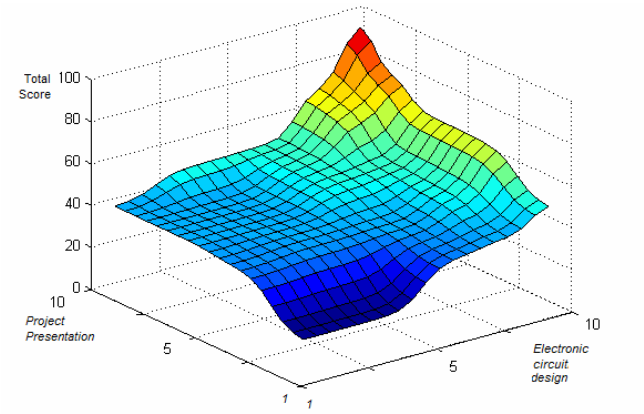


Figure 6 A surface plot of the *total score* versus *electronic circuit design* and *project presentation* when *electronic circuit development=5*

SUMMARY

In this paper, we have proposed an approach to construct an FIS-based CRA model. It is argued that the FIS-based CRA model should possess some theoretical properties of a length function. This is important to ensure the validity of the model and to allow valid and meaningful comparisons among students' performance to be made. The *sufficient conditions* is incoperated totp an FIS to ensure that the monotonicity property is fulfiled.

A case study has been conducted to evaluate the proposed approach. The experiment was conducted with data and information obtained from a laboratory project assessment problem at a university in Malaysia. The proposed FIS-based CRA model has been demonstrate to produce results that are in line with expert's knowledge.

For further work, we plan to examine how the FIS-based CRA model is able to fulfil another properties of a length function, i.e., sub-additivity. Further studies are also needed in order to vindicate the usefulness of the proposed FIS-based CRA model in the education assessment domain.

REFERENCES

Biswas. R. (1995) An application of fuzzy sets in students' evaluation. *Fuzzy Sets and Systems*, 74, 187-194.

Broekhoven, E.V. and Baets, B.D. (2008), Monotone Mamdani–Assilian models under mean of maxima defuzzification, *Fuzzy Sets and Systems*, 159, 2819 – 2844

Broekhoven, E.V. and Baets, B.D. (2009), Only Smooth Rule Bases Can Generate Monotone Mamdani–Assilian Models Under Center-of-Gravity Defuzzification, *IEEE Trans. on Fuzzy Systems*, 17, 1157-1174.

Chen, S.M. and Lee, C.H. (1999) New methods for students' evaluation using fuzzy sets. *Fuzzy Sets and Systems*, 104, 209-218.

Inder K. R. (2005), *An introduction to Measure and integration*, Alpha Science International.

Jang, J.S.R., Sun, C.T. and Mizutani, E. (1997) *Neural-Fuzzy and soft Computing*. Prentice-Hall.

Joughin, G., (2008), *Assessment, Learning and Judgement in Higher Education*. London: Springer Netherlands.

Kouikoglou, V.S. and Phillis, Y.A. (2009) On the monotonicity of hierarchical sum-product fuzzy systems, *Fuzzy sets and systems*, 160(24), 3530-3538

Ma, J. and Zhou. D. (2000) Fuzzy Set Approach to the Assessment of Student-Centered Learning. *IEEE Transactions on Education*, 43(2), 237-241.

Nicholson, P. Gillis, S., and Dunning, A., (2009) The use of scoring rubrics to determine clinical performance in the operating suite. *Nurse Education Today*, 29(1), 73-82.

Sadler, D.R (2005) Interpretations of criteria-based assessment and grading in higher education. *Assessment and Evaluation in Higher Education*, 30 (2), 175–194.

Saliu, S. (2005) Constrained Subjective Assessment of Student Learning, *Journal of Science Education and Technology,* 14(3), 271-284.

Tay, K.M., and Lim, C.P. (2008a) On the Use of Fuzzy Inference Techniques in Assessment Models: Part I: Theoretical Properties, *Fuzzy Optimization and Decision Making*, 7(3), 269-281.

Tay, K.M. and Lim, C.P. (2008b) On the Use of Fuzzy Inference Techniques in Assessment Models: Part II: Industrial Applications, *Fuzzy Optimization and Decision Making*, 7(3), 283.-302.

White, D.C. (2002) Criterion Based Assessment using the Support of a Computer**.** *Proceedings of the 35th Hawaii International Conference on System Sciences.*

Won, J.M., Park, S.Y. and Lee, J.S. (2002) Parameter conditions for monotonic Takagi-Sugeno-Kang fuzzy system, *Fuzzy Sets and Systems*, 132, 135-146.

COMPARISON OF COMPUTATIONAL EFFICIENCY OF MOEA\D AND NSGA-II FOR PASSIVE VEHICLE SUSPENSION OPTIMIZATION

Tey Jing Yuen and Rahizar Ramli
Department of Mechanical Engineering
University of Malaya
50603 Kuala Lumpur, Malaysia

KEYWORDS

Passive suspension optimization, multi-objective evolutionary algorithms, genetic algorithms, quarter vehicle model, ride comfort.

ABSTRACT

This paper evaluates new optimization algorithms for optimizing automotive suspension systems employing stochastic methods. This method is introduced as an alternative over the conventional approach, namely trial and error, or design of experiment (DOE), to efficiently optimize the suspension system. Optimizations algorithms employed are the multi-objective evolutionary algorithms based on decomposition (MOEA\D), and non-sorting genetic algorithm II (NSGA-II). A two-degree-of-freedom (2- DOF) linear quarter vehicle model (QVM) traversing a random road profile is utilized to describe the ride dynamics. The road irregularity is assumed as a Gaussian random process and represented as a simple exponential power spectral density (PSD). The evaluated performance indices are the discomfort parameter (ACC), suspension working space (SWS) and dynamic tyre load (DTL). The optimised design variables are the suspension stiffness, K_s and damping coefficient, C_s. In this paper, both algorithms are analyzed with different sets of experiments to compare their computational efficiency. The results indicated that MOEA\D is computationally efficient in searching for Pareto solutions compared to NSGA-II, and showed reasonable improvement in ride comfort.

INTRODUCTION

In vehicle dynamics, the suspension system isolates the vehicle body from the roughness of the road surfaces. Optimizing the suspension system in terms of ride is crucial to maintain the passenger comfort when traversing these road profiles. Traditionally, experienced engineers employ trial and error approach or DOE to tune the suspension. The major drawback of these methods is that, it is time consuming and does not provide reliable global optimal solution. These drawback can be overcome by employing the stochastic method as proposed in this paper. In this paper, the main aim is to evaluate the computational efficiency and to provide the best Pareto optimal solution to optimize a passive suspension system using stochastic methods.

Many researchers (Reimpell et al. 2001; Crolla and Whitehead 2003; Mastinu et al. 2006; Schiehlen 2007; Guglielmino et al. 2008; Jazar 2008; Genta and Morello 2009a; Genta and Morello 2009b) have attempted to derive suspension model with basic engineering rules. In this paper, stochastic optimization is employed in a QVM to obtain the optimized design variable i.e. the spring stiffness (K_s) and damping coefficient (C_s). To achieve the best ride comfort, the QVM with passive suspension is optimized against the suspension working space (SWS), the discomfort parameter (ACC) and the dynamic tire load (DTL). However, there is a design constraint of the SWS since the suspension travel is limited. Therefore, stochastic optimization is required to find the best compromise solutions for K_s and C_s within the design constraint.

Alternative approach in solving similar problem can be done using gradient-based optimization (Lina and Zhang 1989; Tamboli and Joshi 1999; Sun et al. 2000; Els and Uys 2003; Thoresson et al. 2009a; Thoresson et al. 2009b). The drawback of this technique is that it may trap in local minimal point (depending on the starting point). Additionally, it requires complex auxiliary equations to execute and to derive the system dynamic as the degree of freedom increases. It is proven that optimization using genetic algorithms performs better than gradient-based method as reported by (Baumal et al. 1998).

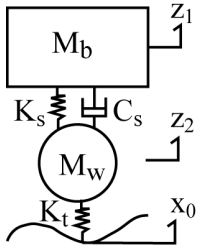
In stochastic method, Alkhatib et. al. (Alkhatib et al. 2004) employs simple genetic algorithm (GA) method for a 2 DOF QVM, aims at minimizing absolute acceleration (RMS) sensitivity to changes in relative displacement (RMS). However, the use of simple GA is not robust thus, does not provide a reliable solution near to the global minimum as compared to the presently improved algorithms (Konaka et al. 2006). In this paper, two recently develop optimization methods to produce reliable and robust of the solution for a linear two DOF QVM are examined. They are the non-sorting genetic algorithm-2 (NSGA-II) and multi-objective evolutionary algorithms based on decomposition (MOEA\D). NSGA-II is the current state-of-art optimization technique known for its fast convergence and robust (Deb et al. 2002). MOEA\D is the new algorithm developed by Zhang et. al. (Zhang et al. 2009). It employs decomposition method to decompose a multi-objective optimization problem into a number of scalar optimization sub-problems and optimizes them

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

simultaneously. This method produced less computational complexity at each generation as compared to NSGA-II and it is capable of generating evenly distributed solutions, as reported by Zhang et. al. (Zhang et al. 2009).

QUARTER VEHICLE MODEL (QVM)

The simplest form of analytical derivation for the estimation of dynamic response of vehicle traversing a randomly road profiled is QVM. Sharp (Sharp 1987), reported that QVM has proven to produce reasonably accurate prediction of the dynamic behavior in terms of ACC, SWS and DTL. The random road profile can be represented as a simple exponential power spectral density (PSD). This one slope PSD is accurate in estimating the amplitudes of the road irregularity, especially at low frequency excitation. There are other complex power spectral densities but is generally complex and impractical to solve using analytical approach as reported by (Gobbi and Mastinu 2001; Karamihas 2005). A generic QVM propose by (Crolla and Whitehead 2003) is used as illustrated in Figure 1.



Where,
$M_b = body\ mass$
$M_w = wheel\ mass$
$K_s = spring\ stiffness$
$C_s = damping\ coefficient$
$K_t = tire\ stiffnes\ coefficient$
$x_0 = road\ profile\ input$

Figures 1: Quarter Vehicle Model Representation

The equation of motion for the QVM can be formulated using Newtonian methods as shown,

$$M_w \ddot{z_1} = K_t(x_0 - z_1) - K_s(z_1 - z_2) - C_s(\dot{z_1} - \dot{z_2}) \quad (1)$$

$$M_b \ddot{z_2} = K_s(z_1 - z_2) + C_s(\dot{z_1} - \dot{z_2}) \quad (2)$$

The frequency response functions of z_1/x_0 and z_2/x_0 can be found by solving the equation (1) and (2) using state-space approach. Since the road profile input is in PSD, the output of PSD can be calculated using the following formula:

$$PSD_{output} = |H(\omega)|^2 . PSD_{input} \quad (3)$$

where, *H(ω)* is the frequency response function and ω is the frequency. The objective function of optimization is defined as the following:

Discomfort parameter (ACC) evaluates ride quality by combining the frequency components according to the ISO recommended weighting scheme shown in Figure 2. These weighting functions are applied to the multi frequency acceleration spectra prior to integration, to give a root mean squared (RMS) value of frequency-weighted acceleration or discomfort parameter.

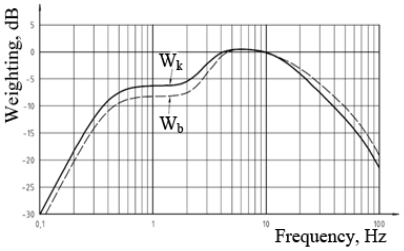
The discomfort parameter (ACC) is calculated as follows:

$$H(\omega)_{M_b} = \left|-\omega^2 . \frac{Z_2}{X_0}\right| \quad (4)$$

$$PSD_{M_b} = \left|-\omega^2 . \frac{Z_2}{X_0}\right|^2 . Sf \quad (5)$$

where, *Sf = road input power spectral density*. Then PSD_{M_b} is transformed into the discomfort parameter measured on human ($PSD_{weighted}$) using ISO 2631 in Figures 2.

$$ACC = RMS\ PSD_{weighted} = \sqrt{\int PSD_{weighted}} \quad (6)$$



W_k – weighting curve for vertical acceleration

W_b – acceptable approximation of the general curve W_k

Figures 2: ISO2631 Weighting Function of Acceleration Spectra (ISO 2001)

Suspension working space (SWS) is a parameter defined as the RMS value of wheel to body displacement (Z_2 - Z_1). It measures the variation of the displacement about its static position.

Suspension working space (SWS) is calculated as follows:

$$H(\omega)_{SWS} = \left|\frac{Z_2}{X_0} - \frac{Z_1}{X_0}\right| \quad (7)$$

$$PSD_{SWS} = \left|\frac{Z_2}{X_0} - \frac{Z_1}{X_0}\right|^2 . Sf \quad (9)$$

$$SWS = RMS\ PSD_{SWS} = \sqrt{\int PSD_{SWS}} \quad (10)$$

Dynamic tyre load (DTL) is defined as the RMS value of tire load variation about the static value. This parameter can be considered a measure of the vehicle's road holding ability, since a variation in the tire load results in a varying contact length and a net reduction in side or braking force.

Dynamic tyre loading (DTL) is calculated as follows:

$$H(\omega)_{DTL} = K_t \left|\frac{Z_1}{X_0} - 1\right| \quad (11)$$

$$PSD_{DTL} = \left(K_t \left|\frac{Z_1}{X_0} - 1\right|\right)^2 . Sf \quad (12)$$

$$DTL = RMS\ PSD_{DTL} = \sqrt{\int PSD_{DTL}} \quad (13)$$

OPTIMIZATION ALGORITHMS

Non-sorting genetic algorithms II (NSGA-II)

NSGA-II developed by Deb et. al. (Deb 2002) was an improved version of NSGA (Srinivas N 1994). The development of the improved version of NSGA-II employs different strategy to solve the previous three main issues in NSGA approach.

```
Initialize Population
  * Generate random population
  * Evaluate objective values (SWS, ACC, and
    DTL)
  * Assign rank (level) based on Pareto dominance
  * Generate child population
        - Tournament selection
        - Recombination and mutation

For i = 1 to number of generations
  * Parent and child population are assign rank based on
    Pareto
  * Generate sets of non-dominated fronts
  * Determine the crowding distance between points on
    each front
  * Select points based on crowding distance calculation
    and fill into the parent population until full.
  * Create next generation
  * Tournament Selection
  * Recombination and Mutation
  * Evaluate Objective Values (SWS, ACC, and DTL)
  *Increment generation index
End.
```

Figures 3: NSGA-II Pseudo Code

The main issues in NSGA approach are:

- The non-dominated sorting approach is time consuming and involves high computational complexity.
- It lacks of elitism, which is important in preventing the loss of good solutions once they are found.
- It requires the user to specify the sharing parameters, which is difficult for the user to know the ideal value for the parameters.

Therefore, in NSGA-II, two new approaches have been proposed to solve these issues. First, it employs a fast non-dominated sorting aims at reducing the complexity of sorting as compare to old non-dominating sorting. Secondly, it introduces a crowd-comparison approach to replace the sharing parameter needed in NSGA. The pseudo code of the NSGA-II is shown in Figure 3.

Multi-objective evolutionary algorithms based on decomposition (MOEA\D)

In recent years, the progresses on evolutionary algorithms (EAs) that deal with multi-objective problems have increased significantly. MOEA\D is one of the multi-objective evolutionary algorithms (MOEAs) aims at finding a set of representative Pareto optimal solutions in a single run. MOEA\D is also one of the Pareto dominance-based MOEA like NSGA-II (Deb 2002; Zhang et al. 2009) . The algorithm is developed by Zhang (Zhang et al. 2009) that include the use of Tchebycheff approach as the decomposition method, with dynamical resource allocation, to improve the efficiency of the MOEA\D. The following are the details of the algorithms:

```
Define [termination condition, N (number of sub-
problems), a uniform spread weight vectors, T (number of
the weight vectors in the neighborhood of each weight
vector)]

Initialization
  * Generate initial population by uniformly spreading and
    randomly sampling from search space
  * Calculate the reference point for the Tchebycheff
    approach.
  * Evaluate Objective Values (SWS, ACC, and DTL)
  * Selection using tournament selection method based on
    utility π^i
  * Selection of mating and updating range
  * Reproduction
  * Repair
  * Update of solutions

While (not equal to termination condition)
  * Evaluate Objective Values (SWS, ACC, and DTL)
  * Selection using tournament selection method base on
    utility π^i
  * Selection of mating and update range
  * Reproduction
  * Repair - if the searching element is out of boundary
 * Update the solutions

If (generation is a multiplication of a pre-set value of x),
   * Update utility function;
End

End
```

Figures 4: Pseudo Code of MOEA\D

The Tchebycheff decomposition approach (Zhang et al. 2009) can be described as the scalar optimization problems of the form:

$$\text{Minimize } g^{te}(x|\lambda, z^*) = max\ 1 \le i \le m\ \{\lambda_i | f_i(x) - z^*|\}$$
$$\text{subject to } x \in \Omega \qquad (14)$$

where $z^* = (z_1^*, \dots, z_m^*)^T$ is the reference point, i.e. $z_i^* = min\{f_i(x)|x \in \Omega\}$ for each $i = 1, \dots, m$. Under certain mild conditions, in each Pareto optimal point x^*, there exists a weight vector, λ such that x^* is the optimal solution of (14) where each one is a Pareto optimal solution of the objective function ($minimize\ F(x) = \left(f_1(x), \dots, f_m(x)\right)^T$). Eventually, this allows the user to obtain different Pareto optimal solutions by solving a set

of single objective optimization problem defined by the Tchebycheff approach with different weight vectors.

MOEA\D with Dynamical Resource Allocation:

MOEA\D (Zhang et al. 2009) minimizes the entire *N* objective simultaneously in a single run. Neighbourhood relations among these single objective sub-problems are defined from the distances among their weight vectors. In the previous version of MOEA\D proposed by (Zhang and Li 2007), all the sub-problems are treated equally and received about the same amount of computational effort. However, in the real situation, each sub-problems may encounter different level of difficulty in obtaining the solution. Therefore, the new version of MOEA\D with a dynamical resource allocation (MOEA\D-DRA) is introduced (Zhang et al. 2009) by computing a utility parameter π^i for each of the sub-problems *i*, allowing computational efforts to be distributed based on their utilities.

RESULTS AND DISCUSSIONS

The computational efficiency of both algorithms i.e. NSGA-II and MOEA\D is evaluated based on the number of iteration and population size. Clearly, this is crucial as the complexity of the vehicle degree-of-freedom increases. The experiment is conducted with two different termination conditions i.e. at 10 iterations and at 20 iterations. For each termination conditions, two different number of population of 500 and 1000 are chosen, where the average solution time is evaluated. Additionally, the diversity of the Pareto solutions is also evaluated. In MOEA\D, the parameter is set as the default value (Zhang et. al. 2009). In NSGA-II, the parameter is also set as the default value from MATLAB GA toolbox. All experiments are executed on a standard desktop PC powered by Intel Processor E7300 with 4GB RAM.

For each of the experiment, the solution time is averaged for five repeats. The optimization aims at minimizing all three objectives i.e. SWS, ACC, and DTL for the QVM with the following variables shown in Table 1. Here, the optimization routine will search for the best sets of solutions for the spring stiffness, K_s and the damping coefficient, C_s. Therefore, a range of K_s and C_s must be specified (Table 1). The reference value of K_s and C_s are used as the design target since these values represent the experimentally optimized solution for the suspension setting.

The results shown in Figures 5 indicate that MOEA\D algorithm is significantly faster than NSGA-II for same number of iteration and population size. This is because NSGA-II has higher computational complexity for each generation which can be expressed as mN^2, where *m* is the number of the objectives and *N* is its population size (Deb et al. 2002). As the number of population size increases, the time needed by NSGA-II increases exponentially affected by the term N^2. However, in MOEA\D the computational complexity is only mNT, where *T* is the number of the weight vectors in the neighborhood, usually has smaller value than *N*. Therefore, MOEA\D can solve much faster than NSGA-II at each generation (Table 4).

Table 1: Quarter vehicle model parameter (Crolla and Whitehead 2003)

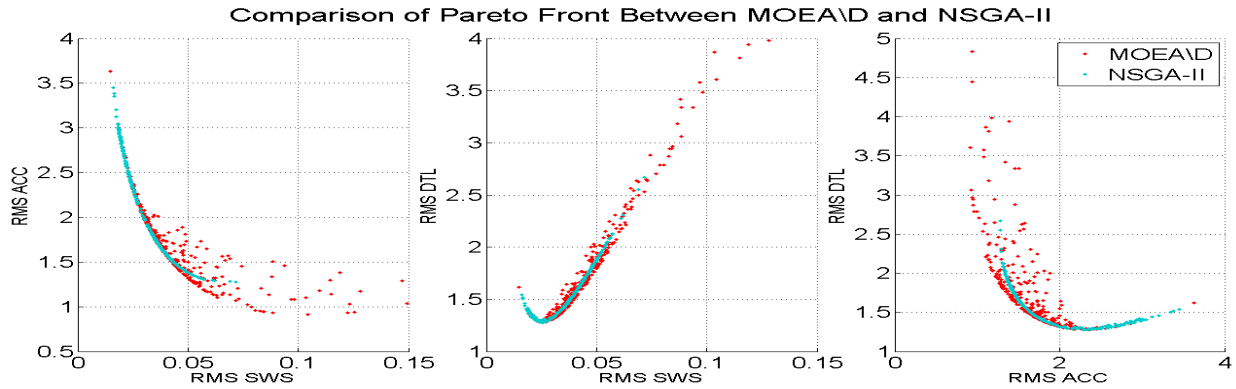
Body mass, M_b	317,5 kg
Wheel mass, M_w	45,4 kg
Spring stiffness, K_s	1 - 30 kN/m (**Ref: 22 kN/m**)
Damping coefficient, C_s	1 - 5 kNs/m (**Ref: 1,5 kNs/m**)
Tire stiffness coefficient, K_t	192 kN/m
Vehicle traveling speed, V	20 m/s
PSD road input	$\frac{5 \times 10^{-6}.V^{1.5}}{f^{2.5}}$
Working frequency range, f	0 - 20 Hz

Table 2: Computational Efficiency of NSGA-II

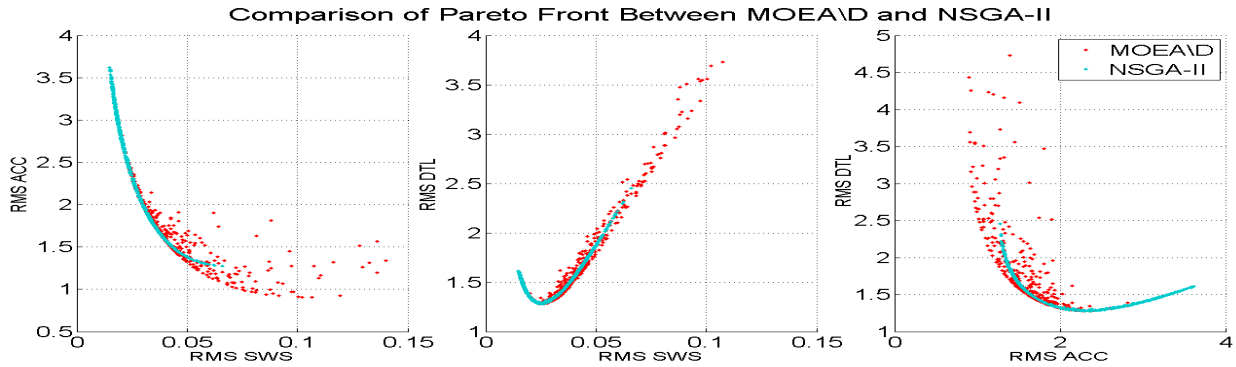
Iteration		10	20	10	20
Population size		500	500	1000	1000
Instance	1	117,4	218,8	276,4	533,7
	2	117,8	223,4	284,0	549,3
	3	116,8	225,2	293,7	550,3
	4	119,0	226,7	293,8	546,5
	5	117,8	227,4	288,4	549,4
Average time, sec		117,8	224,3	287,3	545,8

Table 3: Computational Efficiency of MOEA\D

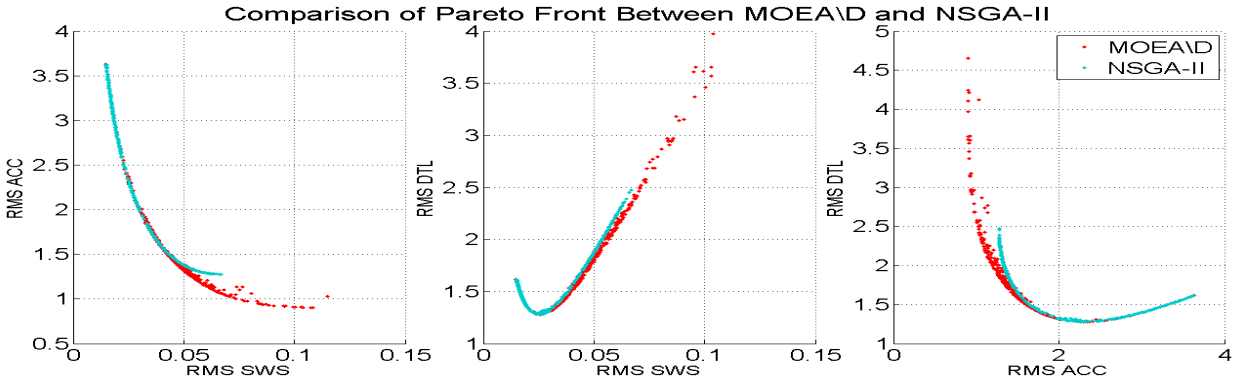
Iteration		10	20	10	20
Population size		500	500	1000	1000
Instance	1	49,8	84,7	96,6	176,9
	2	49,5	85,9	97,2	168,0
	3	48,5	85,0	98,7	190,9
	4	49,4	86,0	98,1	192,4
	5	48,7	87,3	99,0	173,3
Average time, sec		49,2	85,8	97,9	180,3



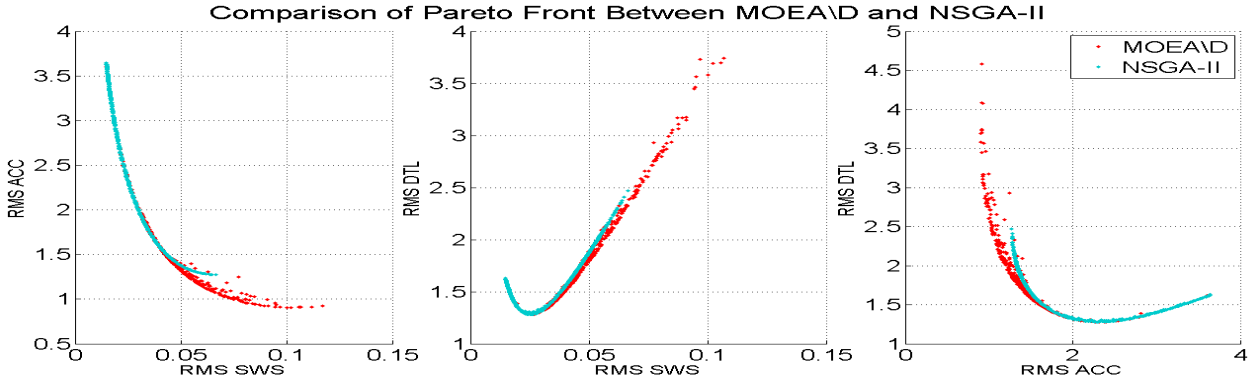
Case 1 (10 Iteration, Population size 500)



Case 2 (10 Iteration, Population size 1000)



Case 3 (20 Iteration, Population size 500)



Case 4 (20 Iteration, Population size 1000)

Figures 5: Plots of Final Populations with MOEA\D and NSGA-II in the Objective Space of SWS, ACC, and DTL

Another aspect of the performance indicator can be observed from the diversity between MOEA\D and NSGA-II. In Figures 5, the termination iteration at 10 suggested that MOEA\D provides a better diversity and

more explorative in reaching the Pareto Front than those of the NSGA-II. This gives MOEA\D the advantage in the preliminary run to indicate the Pareto shape and the range of interest to achieve the objective. When the iteration termination is increased to 20 iterations, the MOEA\D showed larger spread suggesting a better searching capability throughout the Pareto Fronts as compared to those from NSGA-II. The computational efficiency of the algorithm is critical since the ultimate algorithm is critical since the ultimate aim of this research will focus on optimizing high fidelity vehicle model. For example, full vehicle model with increased degree-of-freedom.

Table 4: Comparison of the Computational Efficiency for the Two Competing Optimization Algorithms

Iteration	10	20	10	20
Population size	500	500	1000	1000
NSGA-II	117,8	224,3	287,3	545,8
MOEA\D	49,2	85,8	97,9	180,3
Percentage of MOEA\D faster than NSGA-II, %	239	261	293	302

Optimal Solution for quarter vehicle model

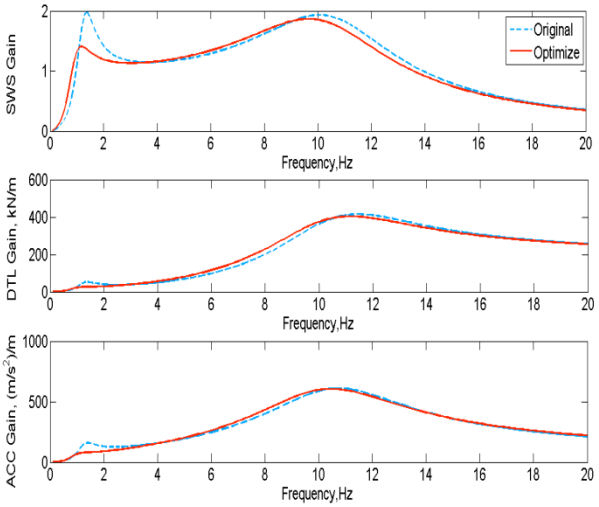
Generally, in designing vehicle suspension, the SWS is chosen as a design constraint. This is because there is a limited clearance between the vehicle body mass and the wheel mass. In this experiment, the design limit of 0.0287m is employed which is similar to the SWS value used in the reference car design (Crolla and Whitehead 2003).

To determine the best range for K_s & C_s, MOEA\D is chosen since earlier results suggest that this algorithm is more computationally efficient. It produces different combinations of K_s and C_s that is capable to achieve the same SWS criteria of 0.0287 m. The optimal combination of K_s and C_s showed reasonable improvement on the ACC and DTL as illustrated in Table 5.

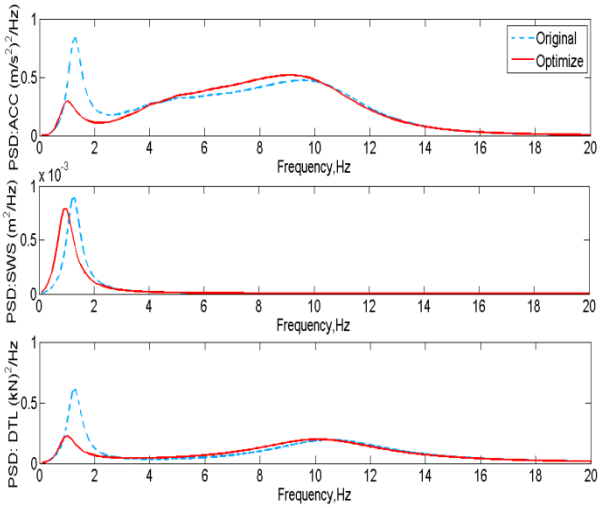
Table 5: Comparison Result between the Original Parameter with Optimize Parameter

Original Reference Vehicle Parameters		Optimized parameters	Improvement, %
K_s , kN/m	22	12,3892	N/A
C_s , kNs/m	1,5	1,6163	N/A
RMS SWS	0,0287	0,0287	0
RMS ACC	2,1152	2,066	2,33
RMS DTL	1,3649	1,3051	4,38

The result of the optimization can be further analyzed by plotting them against its frequency responses as shown in Figures 6 and 7. Here, the optimization process not only improves the overall RMS response of ACC and DTL, but further reduces the natural frequency of vehicle as compare to the reference vehicle. This will further reduce the possibility of excitation to occur at such low frequency on normal road condition.



Figures 6: Bode Plots for SWS, ACC and DTL.



Figures 7: Bode plots for PSD SWS, PSD ACC and PSD DTL

CONCLUSIONS AND FUTURE WORK

In this paper, NSGA-II and MOEA\D are employed to optimize a two DOF QVM. Performance comparison between both algorithms showed that MOEA\D is more computationally efficient and robust than the NSGA-II in finding the Pareto Front. Improvement in ride comfort has been achieved using MOEA\D based on the optimum value of suspension stiffness and damping coefficient. This demonstrates that the proposed method is efficient to be employed in the automotive industry to reduce product development time. In addition, it also provides Pareto Front as a flexible design option for designer to optimize the suspension system in the early

stage of design cycle. Future work will include the usage of this algorithm on multi-body vehicle model (high fidelity model) which involves with increase number of degree-of-freedom and more parameters to simultaneously optimize the ride and handling performances.

REFERENCES

Alkhatib, R., G. N. Jazar, and M.F. Golnaraghi. 2004. "Optimal design of passive linear suspension using genetic algorithm." *Journal of Sound and Vibration* 275.

Baumal, A. E., J. J. McPhee, and P.H. Calami. 1998. "Application of Genetic Algorithms to the Design Optimization of an Active Vehicle Suspension System." *Comput. Methods Appl. Mech. Engrg.* 163.

Crolla, D. A. and J.P. Whitehead . 2003. *Vehicle Dynamics, Control and Suspensions*. Department of Mechanical Engineering. U.K, University of Leeds.

Deb, K., A. Pratap, S. Agarwal and T. Meyarivan . 2002. "Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II." *IEEE Trans Evol Comput* 6, No.2, 182-197.

Els, P. S. and P. E. Uys. 2003. "Investigation of the Applicability of the Dynamic-Q Optimisation Algorithm to Vehicle Suspension Design." *Mathematical and Computer Modelling* 37.

Genta, G. and L. Morello. 2009a. *The Automotive Chassis: Components Design*. Springer.

Genta, G. and L. Morello. 2009b. *The Automotive Chassis: System Design*. Springer.

Gobbi, M. and G. Mastinu. 2001. "Analytical Description And Optimization of the Dynamic Behaviour of Passively Suspended Road Vehicles." *Journal of Sound and Vibration* 245(3), 457-481.

Guglielmino, E., T. Sireteanu, C.W. Stammers, G. Ghia and M. Giuclea. 2008. *Semi-active Suspension Control : Improved Vehicle Ride and Road Friendliness*. Springer.

ISO, B. S. 2001. ISO2631-4:2001. "Mechanical Vibration and Shock - Evaluation of human exposure to whole-body vibration" ISO.

Karamihas, S. M. 2005. *Critical Profiler Accuracy Requirements*. University of Michigan Transportation Research Institute.

Konak, A., D. W. Coit, and A.E. Smith. 2006. "Multi-objective Optimization Using Genetic Algorithms: A tutorial." *Reliability Engineering and System Safety* 91.

Lina, Y. and Y. Zhang. 1989. "Suspension Optimization by a Frequency Domain Equivalent Optimal Control Algorithm " *Journal of Sound and Vibration,* Vol. 133(2), pg. 239-249.

Mastinu, G., M. Gobbi, and C. Miano. 2006. *Optimal Design of Complex Mechanical Systems with Applications to Vehicle Engineering*. Springer.

N.Jazar, R. 2008. *Vehicle Dynamics: Theory and Applications*. Springer.

R. Sharp, D. C. 1987. " Road Vehicle Suspension System Design - A Review." *Vehicle System Dynamics* 16(3).

Reimpell, J., H. Stoll, and J.W. Betzeler. 2001. *The Automotive Chassis*. Butterworth-Heinemann.

Schiehlen, W. 2007. *Dynamical Analysis of Vehicle Systems: Theoretical Foundations And Advanced Application*. Springer Wien New York.

Srinivas N, D. K. 1994. "Multiobjective Optimization Using Nondominated Sorting in Genetic Algorithms." *J Evol Comput* 2(3), 221-248.

Sun, T. C., A. C. J. Luo, and H.R. Hamidzadeh. 2000. "Dynamic Response andRptimization for Suspension Systems with Non-linear Viscous Damping." *ImechE Part K Journal of Multi-body Dynamics* 214, 181-187.

Tamboli, J. A. and S. G. Joshi. 1999. "Optimum Design of a Passive Suspension System of a Vehicle Subjected To Actual Random Road Excitations." *Journal of Sound and Vibration* 219, 193-205.

Thoresson, M. J., P. E. Uys, P.S. Els and J.A. Snyman. 2009a. "Efficient Optimisation of a Vehicle Suspension System, Using a Gradient-based Approximation Method, Part 1: Mathematical Modelling." *Mathematical and Computer Modelling* 50.

Thoresson, M. J., P. E. Uys, P.S. Els and J.A. Snyman. 2009b. "Efficient Optimisation of a Vehicle Suspension System, Using a Gradient-based Approximation Method, Part 2: Optimisation Results." *Mathematical and Computer Modelling* 50.

Zhang, Q. and H. Li. 2007. "MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition." *IEEE Trans Evol Comput,* Vol. 11.

Zhang, Q., W. Liu, and H. Li. 2009. "The Performance of a New Version of MOEA/D on CEC09 Unconstrained MOP Test Instances." IEEE Congress on Evolutionary Computation,Trondheim, Norway.

TEY JING YUEN received his BEng (Hons) from University of Malaya, Malaysia in 2009. He is currently pursuing Ph.D. studies in Mechanical Engineering at the University of Malaya. His main research areas are multi-objective optimization, multi body dynamic simulation and vehicle dynamic of suspension system. His e-mail is: jingyuen_tey85@yahoo.com

RAHIZAR RAMLI received his BSc.(Mech) from the University of Hartford, U.S.A. in 1992. Since then, he worked as a Technical Service engineer in the area of vibration, and condition monitoring. He received his Master (M.Eng.Sc.) degree from the University of Malaya in 1999 specializing in structural dynamics for automotive application. In 2007, he obtained his Ph.D from the University of Leeds, UK in the area of computational mechanics involving vehicle dynamics, semi-active control systems, and durability analysis. His current research interests include experimental and computational mechanics focusing on vibration and acoustics, Finite Element Fatigue Analysis, vehicle dynamics, structural and dynamic optimization. His e-mail address is: rahizar@um.edu.my

A COMPARISON OF POSTURE RECOGNITION USING SUPERVISED AND UNSUPERVISED LEARNING ALGORITHMS

Maleeha Kiran, Chee Seng Chan, Weng Kin Lai
Centre for Multimodal Signal Processing
Mimos Berhad,Technology Park Malaysia
57000 Kuala Lumpur, Malaysia
Emails: {malehaa.kiran;cs.chan;lai}@mimos.my

Kyaw Kyaw Hitke Ali, Othman Khalifa
Department of Electrical & Computer Engineering
International Islamic University Malaysia
53100 Gombak, Selangor, Malaysia
Email: khalifa@iiu.edu.my

KEYWORDS

Posture Recognition, Machine Learning, Artificial Intelligence, Image Processing

ABSTRACT

Recognition of human posture is one step in the process of analyzing human behaviour. However, it is an ill-defined problem due to the high degree of freedom exhibited by the human body. In this paper, we study both supervised and unsupervised learning algorithms to recognise human posture in image sequences. In particular, we are interested in a specific set of postures which are representative of typical applications found in video analytics. The algorithms chosen for this paper are K-means, artificial neural network, self organizing maps and particle swarm optimization. Experimental results have shown that the supervised learning algorithms outperform the unsupervised learning algorithms in terms of the number of correctly classified postures. Our future work will focus on detecting abnormal behaviour based on these recognised static postures.

INTRODUCTION

At present the majority of automated surveillance systems act passively meaning they are usually employed to analyze events after an incident has occurred. However to deter such incidents from occurring, it is imperative that the security system pre-emptively alert security personnel whenever any suspicious behaviour or event is detected. In order to do that, the system should have the capability to analyze human behaviour. One of the key aspects of analyzing human behaviour is correctly recognizing different types of human posture from static images. This is needed as we are interested in recognizing acts which can later turn into suspicious activities. For example if the system can detect the act of climbing or jumping from a static image then appropriate alert can be raised before it becomes an intrusion detection activity. In addition the detection of such an act from a static image can be done in a smaller time period thereby making the system more efficient and responsive. This is in contrast to the detection of suspicious activity such as intrusion detection which comprises of a combination of acts and would require the system to monitor a human for a longer period of time.

Ideally the recognition of human postures should be performed by using only one static camera and in real time as the majority of applications for video surveillance applications make use of one static camera to observe the scene and any analysis is done in real time (Boulay et al., 2003). This would also imply that the computational cost of the system is a crucial aspect of such systems. However the downside is that there is usually a tradeoff between computational complexity and classification precision.

In this paper, our aim is to study and compare the performance of both supervised and unsupervised learning algorithms to recognize a few challenging human posture. In particular, we are interested in a specific set of postures which are representative of typical applications which are found in the area of video analytics. The algorithms chosen for this paper are Self-Organizing Maps (SOM), K-means, Artificial Neural Network (ANN), and Particle Swarm Optimization (PSO). Experimental results have shown that the supervised learning algorithms outperformed the unsupervised learning algorithms in terms of the number of correctly classified postures.

The rest of the paper is structured as follows. Section II discusses the related work in this area. Section III discusses various supervised learning algorithms. Section IV presents the unsupervised learning algorithms. Section V shows the experimental results in terms of posture recognition. Section VI concludes the paper with discussion and future work.

RELATED WORK

One of the earlier human motion analysis systems applied the method of template matching. Bobick and Davis (Bobick and Davis, 1996) proposed a view-based approach to the representation and recognition of action using temporal templates. They made use of the binary Motion Energy Image (MEI) and Motion History Image (MHI) to interpret human movement in an image sequence. First, motion images in a sequence were extracted by differencing, and these motion images were accumulated in time to form MEI. Then, the MEI was enhanced into MHI which was a scalar-valued image. Taken together, the MEI and MHI were con-

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

sidered as a two-component version of a temporal template, a vector-valued image, in which each component of each pixel was some function of the motion at that pixel position. Finally, by representing the templates by its seven Hu-moments, a Mahalanobis distance was employed to classify the action of the subject by comparing it to the Hu-moments of pre-recorded actions. Bradski and Davis (Bradski and Davis, 2002) pick up the idea of MHI and develop timed MHI (tMHI) for motion segmentation. tMHI allow determination of the normal optical flow. Motion is segmented relative to object boundaries and the motion orientation. Hu-moments are applied to the binary silhouette to recognise the pose. In Bobick and Davis (Bobick and Davis, 1996), an action is represented by several feature images. Principle Component Analysis (PCA) is applied for dimensionality reduction. Finally, each action is represented by a manifold in the PCA space. Motion history images can also be used to detect and interpret actions in compressed video data.

Yi et. al. (Yi et al., 2005) present the idea of a Pixel Change Ratio Map (PCRM) which is conceptually similar to the MHI. However, further processing is based on motion histograms which are computed from the PCRM. Weinland et. al. (Weinland et al., 2006) suggest replacing the motion history image by a 4D motion history volume. For this, they first compute the visual hull from multiple cameras. Then, they consider the variations around the central vertical axes and use cylindrical coordinates to compute alignments and comparisons. One of the main problem of template matching approaches are the recognition rate of objects based on 2D image features is low, because of the nonlinear distortion during perspective projection and the image variations with the viewpoint's movement. These algorithms are generally unable to recover 3D pose of objects. The stability of dealing effectively with occlusion, overlapping and interference of unrelated structures is generally poor.

Spagnolo et. al. (Spagnolo et al., 2003) proposed a fast and reliable approach to estimate body postures in outdoor visual surveillance. The sequences of images coming from a static camera is trained and tested for recognition. The system uses a clustering algorithm and therefore manually labelling of the clusters is required after the training stage. The features extracted are the horizontal and vertical histograms of binary shapes associated with humans. After training, the Manhattan distance is used for building clusters and for recognition. The main strengths of their method are high classification performance and relatively low computational time which allows the system to perform well in real time.

Buccolieri et. al. (Buccolieri et al., 2005) used active contours and Artificial Neural Networks (ANN) for their posture recognition system. With regards to the feature extraction, localization of moving objects in the image and human posture estimation are performed. The classification is performed by the radial basis functions neural network. Their approach has some advantages such as low sensitivity to noise, fast processing speed, and the ability to handle some degree of occlusion. However, the system is limited to recognizing only three postures, namely standing, bending and squatting postures.

The literature work surveyed focuses on developing a solution required for posture or action recognition tailored to the requirements of specific problems being faced by the authors. There is little effort devoted to doing a comparative analysis of techniques to help decide which technique is suited for the purpose of posture recognition. This paper seeks to address this gap in the current literature work.

SUPERVISED LEARNING

Artificial Neural Network

ANN is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurones) working in unison to solve specific problems. ANNs, like people, learn by example.

The most basic building block in the neural network is the perceptron. It is characterised by several input lines with weights associated to them. The input is collected and summed and an output is given according to the output function *f*. The output is formally defined as follows:

$$o = f(\sum_{i=1}^{n} w_i \vec{x}_i) \tag{1}$$

where o is the output function and w_i the weight associated with input line i. Assume the feature values of the samples are real numbers, and their labels are either 0 or 1 denoting two different classes. The delta rule adapts the weights w_i according to the following formula:

$$w_i(t+1) = w_i(t) + \eta(y_i(t) - o(t))\vec{x}_i(t) \tag{2}$$

where $y_i(t)$ is the desired response, i.e. the real class label, and η an adjustable parameter called the learning rate which usually has a value between 0 and 1. The *t* is used to indicate the time step. Note that there is no change in the weights if the perceptron produces a correct answer, y(t)-o(t) = 0, in other cases the weights are adapted. This makes it unsuited for solving the XOR problem, which need a nonlinear solution (Minsky and Papert, 1969).

The invention of the back-propagation rule caused a major breakthrough in neural network research. In principle the rule is very simple: calculate the error made by the network and propagate it back through the network layers. This back-propagated error is used to update the weights. The sample $\vec{x}$ is fed to the network and produces an output on the right side $\vec{o}$. The input pattern $\vec{x}$ is propagated through the network in the following way:

$$o_j^{(1)} = f\left(\sum_{k=1}^{N} w_{jk}^{(1)} \vec{x}_k\right) \quad (3)$$

$$o_i^{(2)} = f\left(\sum_{j=1}^{M} w_{ij}^{(2)} o_j^{(1)}\right) \quad (4)$$

where $o_j^{(1)}$ and $o_i^{(2)}$ denotes the output of a hidden unit and an output unit respectively. The variables *N* and *M* denote the number of input units and the number of hidden units. A weight from a unit to another unit is denoted by $w_{ij}^{(1)}$ where *j* is the source of the connection, *i* is the target and *l* the layer. The final output of the network can be written as:

$$o_i^{(2)} = f\left(\sum_{j=1}^{M} w_{ij}^{(2)} f(\sum_{k=1}^{N} w_{jk}^{(1)} \vec{x}_k\right) \quad (5)$$

where $o_j^{(1)}$ has been replaced by Eq. 3 . The output of the network has to be judged using some error criterion. The criterion determines the size of the error to be back propagated. In general, the Mean Squared Error (MSE) criterion is employed:

$$E = \frac{1}{2}\frac{1}{|L|}\sum_{i=1}^{|L|}\sum_{\vec{x}\in L}[y(\vec{x}) - o(\vec{x})]^2 \quad (6)$$

where $y(\vec{x})$ is the desired network output value for the sample $\vec{x}$ under investigation and —L— is the size (cardinality) of the learning set. The objective during the training is to minimise the error function (Eq. 6) by choosing the appropriate weights.

Self-organizing maps

SOM (T.Kohonen, 1995) is a type of ANN that is trained using unsupervised learning to produce a two dimensional, discretised representation of the input space of the training samples, called a map. The map preserves the topological properties of the input space. This makes SOM useful for visualizing low-dimensional views of high-dimensional data, similar to multidimensional scaling. Same to most of the ANN, SOMs operate in two modes: training and mapping. Training builds the map using input examples. It is a competitive process, also called vector quantization. Mapping automatically classifies a new input vector. Associated with each node in this neural network is a weight vector of the same dimension as the input data vectors and a position in the map space. The usual arrangement of nodes is a regular spacing in a hexagonal or rectangular grid.

The SOM describes a mapping from a higher dimensional input space to a lower dimensional map space. The procedure for placing a vector from data space onto the map is to find the node with the closest weight vector to the vector taken from data space and to assign the map coordinates of this node to the input vector. The goal of learning in the SOM is to cause different parts of the network to respond similarly to certain input patterns. This is partly motivated by how visual, auditory or other sensory information is handled in separate parts of the cerebral cortex in the human brain. The weights of the neurons are initialized either to small random values or sampled evenly from the subspace spanned by the two largest principal component eigenvectors. With the latter alternative, learning is much faster because the initial weights already give good approximation of SOM weights.

The network must be fed a large number of example vectors that represent, as close as possible, the kinds of vectors expected during mapping. The examples are usually administered several times. The training utilizes competitive learning. When a training example is fed to the network, its Euclidean distance to all weight vectors is computed.

The neuron with weight vector most similar to the input is called the best matching unit (BMU). The weights of the BMU and neurons close to it in the SOM lattice are adjusted towards the input vector. The magnitude of the change decreases with time and with distance from the BMU. The update formula for a neuron with weight vector $W_v(t)$ is:

$$W_v(t+1) = W_v(t) + \Theta(v,t)\alpha(t)(D(t) - W_v(t)) \quad (7)$$

where $\alpha(t)$ is a monotonically decreasing learning coefficient and *D(t)* is the input vector Duda et al. (2000); Alpaydin (2004). The goal of learning in self-organizing maps is to cause different parts of the network to respond similarly to certain input patterns. The network must be fed a large number of example vectors that represents, as close as possible, the kinds of vectors expected during mapping. The examples are usually administered several times and the training utilizes competitive learning.

UNSUPERVISED LEARNING

K-means Algorithm

As similar to other partitional clustering algorithms, K-means algorithm is generally an iterative algorithm that converge to local optimal (Costa and Cesar, 2001). Employing the general form of iterative clustering, the steps of K-means algorithm are:

1. Using the output from PSO as initial K cluster centroids

2. **Repeat**
   (a) **For** each pattern, z_p in the dataset **do**
   Compute its membership $u(M_k|z_p)$ to each centroid m_k and its weight $w(z_p)$
   end loop
   (b) Recalculate the K cluster centroids, using

$$m_k = \frac{\sum_{\forall z_p} u(m_k|z_p)w(z_p)z_p}{\sum_{\forall z_p} u(m_k|z_p)w(z_p)} \quad (8)$$

until a stopping criterion is satisfied

where $u(m_k|z_p)$ is the membership function which quantifies the membership of pattern z_p to cluster *k*. For K-means algorithm in this paper, the membership and weight function are defined as

$$u(m_k|z_p) = \begin{cases} 1 & \text{if } d^2(z_p, m_k) = \arg\min_k\{d^2(z_p, m_k)\} \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

$$w(z_p) = 1 \quad (10)$$

Particle Swarm Optimisation
Particle Swarm Optimisation (PSO) is a computational paradigm based on the phenomenon of collective intelligence exhibited by the social behaviour of bird flocking and fish schooling (Kennedy and Eberhart, 1995). In a real number space, each individual possible solution can be modelled as a particle that moves through the problem hyperspace. The position of each particle is determined by the vector $x_i \in \Re_n$ and its movement by the velocity of the particle $v_i \in \Re_n$ (Kennedy, 1997), as shown in Eq 11.

$$\overrightarrow{x_i}(t) = \overrightarrow{x_i}(t-1) + \overrightarrow{v_i}(t) \quad (11)$$

The information available for each individual is based on its own experience (the decisions that it has made so far and the success of each decision) and the knowledge of the performance of other individuals in its neighbourhood. Since the relative importance of these two factors can vary from one decision to another, it is reasonable to apply random weights to each part, and therefore the velocity will be determined by

$$\overrightarrow{v_i}(t) = \alpha\overrightarrow{v_i}(t) + \phi_1 r(\cdot)(p_i - \overrightarrow{x_i}(t-1)) + \phi_2 r(\cdot)(p_g - \overrightarrow{x_i}(t-1)) \quad (12)$$

where α called inertia is a parameter within the range [0, 1] and is often decreased over time; ϕ_1 and ϕ_2 are two constants, often chosen so that $\phi_1 + \phi_2 = 4$, which control the degree to which the particle "'follow the herd"' thus stressing exploitation (higher values of ϕ_1); $r(\cdot)$ is a uniformly random number generator function that returns values within the interval (0,1); and g is the particle in i's neighbourhood with the current neighbourhood-best candidate solution.

According to the formulation above, the following procedure can be used for implementing the PSO algorithm (Shi, 2004).

1. Initialize the swarm by assigning a random position in the problem hyperspace to each particle.

Figure 1: Example posture silhoutte from dataset. Top to bottom: Climbing, Fighting, Jumping, Lying Down and Pointing

2. Evaluate the fitness function for each particle.
3. For each individual particle, compare the particle's fitness value with its p_{best}. If the current value is better than the p_{best} value, then set this value as the p_{best} and the current particle's position, x_i, as p_i.
4. Identify the particle that has the best fitness value. The value of its fitness function is identified as g_{best} and its position as p_g.
5. Update the velocities and positions of all the particles using Eq. 11 and Eq. 12.
6. Repeat steps (2) to (5) until a stopping criterion is met (e.g., maximum number of iterations or a sufficiently good fitness value).

EXPERIMENTAL RESULTS

In this section, the comparison between supervised and unsupervised training to posture recognition is presented. Our dataset consist of five challenging postures silhouette, e.g. lying, jumping, fighting, climbing and pointing and a sample dataset is illustrated in Fig. 1. The dataset is collected in such a way that no occlusion between intra body-parts occurs and the camera distance to subject is constant for each posture. Similar constraints have been imposed by other researchers as many techniques and algorithms are at an experimental stage performing very well in tightly controlled laboratory settings as to (Spagnolo et al., 2003; Korde and Jondhale, 2008).

For each of the supervised or unsupervised methods we first perform the pre-processing and feature extraction as (Herrero-Jaraba et al., 2004) in order to remove noise and illumination problems occurring during the data capture phase. The feature extracted from each binary image comprises of a silhouette of the human figure. Secondly the input data (silhouette) is cropped to a 50x50 pixels to compensate the processing time needed. Finally, according to Section III and IV, each supervised and unsupervised algorithms are employed to learn and classify five

different postures. All the algorithms are implemented in MATLAB and the learning and testing data is chosen empirically. The parameters settings for the algorithms can be summarized as below:

- The ANN consists of 2500 input neurons as each image is represented by a vector of 50x50 pixels in size with 100 dataset for each of the 5 posture classes. There are 60 hidden neurons and 5 output neurons, representing one for each of the posture class. The network was trained with the back-propagation using the Levenberg-Marquardt algorithm (Hagan and Menhaj, 1994). This algorithm appears to be the fastest method for training moderate-sized feed-forward neural networks (up to several hundred weights). Meanwhile, the activation function used for the ANN is a log sigmoid function as neural network learn much faster when the activation function is represented by a hyperbolic tangent (Negnevitsky, 2005).

- In K-means, the system is trained with 20 empirically chosen samples per posture and then increased to 50 samples per posture. However, we noticed that increasing the number of samples did not improve the recognition rate significantly therefore we revert back to using 20 samples are selected.

- The 2D lattice in SOM algorithm consists of 40 rows and 40 columns and the epoch is set to 200.

- PSO algorithm was initialised as to (Teng et al., 2010). In each separate experiment, different combinations of static or dynamic acceleration and static or dynamic inertia were used (thus leading to 4 different combinations). In this paper, the number of particle is set to 10 and cluster number is 5 for each experiments.

The system is trained with a dataset of 100 images per posture class. Testing is done a different set of test images (20 per posture class). Each image is classified as belonging to a particular posture class and its recognition accuracy is calculated. The overall effectiveness of these algorithms is determined using this recognition accuracy. This overall accuracy measurement determines how well the algorithms are able to correctly classify postures in a given dataset. The number of correctly classified posture is referred to as the True Positives (TP). Any postures that are not correctly classified are considered False Positives (FP). The overall accuracy is thus calculates as Eq. 13:

$$\text{Overall Accuracy} = \frac{\sum \text{TP for all cluster}}{\text{Total Number of Data}} \quad (13)$$

The recognition results for the dataset using different algorithms are shown in Table 1. From the analysis of the results, the following hypotheses can be made:

Table 1: Average Recognition Accuracy for the Posture Classes

Algorithm	Pointing	Jumping	Lying Down	Fighting	Climbing
SOM	45%	85%	65%	100%	100%
Neural Network	100%	100%	100%	100%	100%
K-means	30%	25%	20%	15%	65%
PSO	33%	5%	23%	4%	39%

- For all postures, the supervised algorithms outperform the unsupervised algorithms in terms of the percentage of correct classification. The main reason is that supervised learning algorithms learn a mapping from x to y, given a training set made of pairs (x_i, y_i), while the unsupervised learning algorithms find similarity in the training data and the resulting cluster should match the intuitive classification of the data. As the posture classes show large intra-class variation, the clusters discovered are not very distinctive for each posture class.

- On average, the supervised approaches were able to identify the more difficult poses of climbing and fighting postures very well as indicated in Table 1. Nevertheless, the recognition of the lying down posture by SOM was not as good as that obtained by ANN even though this particular posture is significantly different from the others.

- ANN was able to achieve 100% recognition accuracy for all the posture categories. The most probable reason for this is because the dataset of images used for training and testing are free from excessive noise and the posture silhouette of the human is easily distinguished in the image. The dataset of images used in experimental and testing work was collected in controlled environment and it was possible to keep it free from noise or any occlusion from other objects in the scene.

- PSO algorithm achieved the worst recognition rate in the experiment. One of the main reason as indicated by, Merwe and Engelbrecht (van der Merwe and Engelbrecht, 2003) lies in utilizing the PSO algorithm's optimal ability which, if given enough time, could generate a more compact clustering results from the low dimensional dataset than the traditional K-means clustering algorithm. However, when clustering large datasets, the slow shift from the global searching stage to the local refining stage causes the PSO clustering algorithm to require many more iterations to converge to an optimum solution than the K-means.

As Table 1 shows, the PSO and K-means algorithms achieved a low recognition rate for posture classes. Hence for these two approaches we decided to revise the experiment and extract a feature set for representing the human posture instead of using only the human silhouette image. For our study, we have selected six

Table 2: Recognition rate for K-means and PSO algorithms after feature extraction

Algorithm	Pointing	Jumping	Lying Down	Fighting	Climbing
K-means	50%	40%	80%	20%	60%
PSO	60%	40%	90%	30%	60%

features from the human body to be used in representing a single sample of silhouette image. These are the location of the head (H), left arm (A^L), right arm (A^R), left leg (L^L), right leg (L^R) and torso (T). The points of the features, $S = ([H_x, H_y], [A_x^L, A_y^L], [A_x^R, A_y^R], [L_x^L, L_y^L], [L_x^R, L_y^R], [T_x, T_y])$ selected represents the feature points as shown in Figure 2 where *x* and *y* are the pixels coordinate.

Figure 2: Feature points of the six body parts

The points obtained from the samples are used in the training as feature set and then for testing with the K-Means and PSO algorithms. Generally, K-Means is able to converge to a stable solution within a limited number of iteration number compared to PSO which will take a longer time to generate an optimal solution. Hence, for effective comparison, we fixed the number of iterations to be 300 for both the algorithms. The number of samples used in the training and testing phase is 10 samples per posture respectively. Table 2 gives the new results of revised experiment.

It can be surmised from the Table 2 that:

- K-means is able to converge to a solution quickly within a small time range. However, it tends to produces lower accuracies compared to PSO as can be observed in the case of fighting and pointing postures.
- Generally, the results of 'Fighting' and 'Jumping' postures are the lowest for both the algorithms. This could be due to the large similarities of the individual posture as compared to other postures. Including feature extraction stage with PSO has significantly improved the recognition rate for all the posture classes especially for the case of 'Fighting' and' Jumping' which saw the recognition rate jump to 30% and 40% respectively. Even K-means saw an improvement in the recognition rate using a feature set.
- The 'Lying Down' posture produces the highest accuracy for both the algorithms (as shown in Table 3) which was only 20% and 23% before the feature set was included. In the 'Lying Down' posture the orientation of the body plays a vital part in its recognition rate as the orientation is horizontal whereas for others it is vertical

Hence having an appropriate feature set plays a vital role to improve the recognition rate of the posture classes especially for K-means and PSO. The feature set can be made more effective if we extract the angle information for the feature set which includes the angle of the head and limbs with respect to the base line of the sample. Another modification could be to compute the distance of the nearest point of the subject to the base line of the image to determine the orientation of the subject.

CONCLUDING REMARKS

Automated surveillance systems are increasingly being used to analyze human behaviour to detect any suspicious activity. Being able to correctly recognize human posture contributes towards the detection of such an activity. We investigated several supervised and unsupervised algorithm to recognize human postures. The system was first trained to recognize the various postures and then tested against them. The results showed that the unsupervised algorithms tend to give low recognition rates as compared to the supervised algorithms. However the recognition rate from these algorithms showed promising improvement once we extracted a feature set from the posture dataset and tested it with them.

Future work will investigate how the recognition accuracy of the pair of lying down and jumping postures, even though quite distinctive, can be further improved. This improvement is needed especially in the case for unsupervised algorithms. The research work reported in this paper is in its preliminary stages but we intend to obtain accuracy results of the learning algorithms using publicly available databases such as CAVIAR. Finally we also plan to investigate a better selection of feature set to represent the postures and the effect this will have on improving the recognition accuracy of the system.

REFERENCES

Alpaydin, E. (2004). *Introduction to Machine Learning*. The MIT Press.

Bobick, A. and Davis, J. (1996). Real-time recognition of activity using temporal templates. In *Proceedings of IEEE CS Workshop on Applications of Computer Vision*, pages 39–42.

Boulay, B., Bremond, F., and Thonnat, M. (2003). Human posture recognition in video sequence. In *IEEE International Workshop on Visual Surveillance, Performance Evaluation of Tracking & Surveillance*, pages 23–29.

Bradski, G. and Davis, J. (2002). Motion segmentation and pose recognition with motion history gradients. *Machine Vision and Applications*, 13(3):174–184.

Buccolieri, F., Distante, F., and Leone, A. (2005). Human posture recognition using active contours and radial basis function neural network. In *IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 213–218.

Costa, L. and Cesar, R. M. (2001). *Shape Analysis and Classification: Theory and Practice*. CRC Press.

Duda, R. O., Hart, P., and Stork, D. G. (2000). *Pattern Classification*. Wiley - Interscience.

Hagan, M. and Menhaj, M. (1994). Training feed-forward networks with the marquardt algorithm. *IEEE Transactions on Neural Networks*, 5(6):989–993.

Herrero-Jaraba, E., Orrite-Urunuela, C., Monzon, F., and Buldain, D. (2004). Video-based human posture recognition. In *Proceedings of the 2004 IEEE International Conference on Computational Intelligence for Homeland Security and Personal Safety*, pages 19–22.

Kennedy, J. (1997). The particle swarm: social adaptation of knowledge. In *IEEE International Conference on Evolutionary Computation*, pages 303–308.

Kennedy, J. and Eberhart, R. (1995). Particle swarm optimization. In *IEEE International Conference on Neural Networks*, volume 4, pages 1942–1948 vol.4.

Korde, S. and Jondhale, K. C. (2008). Hand gesture recognition system using standard fuzzy c-means algorithm for recognizing hand gesture with angle variations for unsupervised users. In *Proceedings of the 2008 First international Conference on Emerging Trends in Engineering and Technology*, pages 681–685.

Minsky, M. L. and Papert, S. A. (1969). *Perceptrons*. Cambridge: MIT Press.

Negnevitsky, M. (2005). *Artificial Intelligence: A Guide to Intelligent Systems*. Pearson Education Ltd.

Shi, Y. (2004). Particle swarm optimisation. *IEEE Neural Network Society*, Feature Article:8–13.

Spagnolo, P., Leo, M., Leone, A., Attolico, G., and Distante, A. (2003). Posture estimation in visual surveillance of archaeological sites. In *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 277–283.

Teng, S. L., Chan, C. S., Lim, M. K., and Lai, W. K. (2010). Hybrid particle swarm optimization for data clustering. In *Proceedings of SPIE on the* 2^{nd} *International Conference of Digital Image Processing, Singapore*, volume 7546, pages 75460E1–75460E6.

T.Kohonen (1995). *Self-Organizing Maps*. Springer Series in Information Sciences.

van der Merwe, D. and Engelbrecht, A. (2003). Data clustering using particle swarm optimization. In *The 2003 Congress on Evolutionary Computation*, volume 1, pages 215–220.

Weinland, D., Ronfard, R., and Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2-3):249–257.

Yi, H., Rajan, D., and Chia, L.-T. (2005). A new motion histogram to index motion content in video segments. *Pattern Recognition Letters*, 26(9):1221–1231.

AUTHOR BIOGRAPHIES

MALEEHA KIRAN received her Master degree in Computer and Information Engineering from International Islamic University, Malaysia in 2009. She is currently working as a Researcher in the Centre for Multimodal Signal Processing, MIMOS Berhad. Her research interests include image processing and machine learning.

CHEE SENG CHAN received his Ph.D. (Artificial Intelligence) from the University of Portsmouth, U.K. in 2008. Currently, he is a Senior Researcher in Centre of Multimodal Signal Processing, MIMOS Berhad, Malaysia. He previously held research appointment at the University of Portsmouth, U.K. His research interests include computer vision, machine learning, fuzzy qualitative reasoning, cognitive science; esp.: human motion recognition in images and video, biomedical image analysis, compositionality, graphical models and perceptual organization. He is a member for both IET and IEEE.

WENG KIN LAI holds Ph.D. (Engineering) from the University of Auckland, New Zealand. He has been with MIMOS Berhad for over 16 years and is currently the Principal Researcher of the research team developing solutions for video analytics. Weng Kin is a senior member of IEEE , a Fellow of IET and Member of the Governing Board of APNNA. He has served three full terms as the Hon Secretary of IET Malaysia local network, and is currently their Hon Treasurer. His current research interests are in machine intelligence and pattern recognition.

KYAW KYAW HITKE ALI holds B.Eng (Hons) in Electronics-Computer and Information Engineering from International Islamic University Malaysia (ranked 1^{st} out of 620 engineering students). Apart from co-authoring one journal article and a number of conference papers, he has also won several awards in research exhibitions at both national and university levels.

OTHMAN KHALIFA holds a Ph.D. (Digital Image Processing) from Newcastle University, U.K. He has worked in industry for eight years and is currently a Professor and Head of the department of Electrical and Computer Engineering, International Islamic University Malaysia. His areas of research interest are communication systems, information theory and coding, digital image/video processing, coding and compression, wavelets, fractal and pattern recognition. He is a senior member of IEEE.

HYDROMETEOROLOGIC SOCIAL NETWORK WITH CBR PREDICTION

Tomáš Kocyan,
Jan Martinovič,
Andrea Valičková
VŠB - Technical University of Ostrava
FEECS, Department of Computer Science
17. listopadu 15
Ostrava - Poruba
Email: tomas.kocyan@vsb.cz,
jan.martinovic@vsb.cz,
andrea.valickova.st@vsb.cz

Boris Šír,
Michaela Hořínková
VŠB - Technical University of Ostrava
Institute of geoinformatics
17. listopadu 15
Ostrava - Poruba
Email: sir.boris@vsb.cz,
michaela.horinkova@vsb.cz

Veronika Říhová
VŠB - Technical University of Ostrava
Institute of geological engineering
17. listopadu 15
Ostrava - Poruba
veronika.rihova@vsb.cz

KEYWORDS

Case-Based Reasoning, Disaster Management, Case-Based Prediction, Information Retrieval, Episode-Based Reasoning, Social network

ABSTRACT

Human activities are contributing to more frequent natural extremes and climate change, which also come from the atmosphere, water or the Earths crust. With the increasing development of infrastructure, the impacts of these changes and extremes leave more perceivable damage and increasing loss of lives and property. With the use of modern resources and technology we are able to minimize the impact of these extreme phenomena. There are in fact two main approaches - professional and non-professional - both meant from the aspect of data collection and information processing itself. The advantages of social networks have been increasingly utilized during the natural disasters as a way of communicating important information. This article describes the aim of our research - to create a hybrid system which would both enable the collection of data from the professional and non-professional public as well as communicate with other types of systems, to utilize and then process the data and use the data to predict new dangers. The principle is based on collecting data (knowledge, experience, etc.) from both main approaches to disaster management (professional and nonprofessional) and then applying this information to achieve new solutions. The practical application of a DIP system shows that it can be used for describing the risk of future natural disasters and enables us to deduce the threat imposed by them. Analyzing this data can help create new solutions in the fight to minimize the damage incurred by these disasters.

INTRODUCTION

Climate change as well as the occurrence of natural extremes, whose sources come from the atmosphere, water or even the Earths crust, are the results of the natural variability of the atmosphere and the evolution of the Earth. These continual changes, caused not only by natural processes, are increasingly influenced by human activities.

In natural ecosystems, these changes and extremes, including their occurrences and effects, are part of their natural development. But with the increasing development of infrastructure these changes and extremes leave more perceivable damage, i.e. loss of lives and property.

Ecosystems and human society are both getting gradually equipped to adapt to recent and current climate. But there are fears that further adaptation to accelerating changes will be much more difficult. In a short time it can have big consequences for the fundamental values of life, the food or water system or public health, especially in many underdeveloped and economically disadvantaged countries. For almost all of the world there is an increased risk of extreme weather, with subsequent increased risk of disasters associated with them. As stated by the World Meteorological Organization, over the past 30 years, nearly 7500 natural disasters worldwide took the lives of more than two million people. Of that 72.5 percent of the disasters were caused due to weather, climate change or water hazards. In the context of the ongoing process of global warming and global climate change, the question arises to what extent these changes affect meteorological and hydrological extremes causing floods.

Floods are one of the most significant natural extremes in the Czech Republic, and are largely the result of anthropogenic factors, as well as meteorological and physical-geographical factors. Anthropogenically bur-

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

dened landscape is losing its ability to maintain stability and dynamic balance.

We may not be able to control the winds and rains yet, but in the modern era we fortunately have new tools and kinds of technology, which give us the opportunity to minimize the impact of these extreme phenomena. For example, by providing relevant and comprehensive information for decision support, i.e. creating forecasts for the situation, with the aim of limiting the adverse effects of natural phenomena and their consequences through modern computer and internet technologies. These activities can be supported by both professionals and nonprofessionals in the field, as well as by various levels of public safety administrations, and may reveal solutions for dealing with future disasters.

In fact there are two main approaches to disaster management meant from the aspect of data collection and information processing itself. On the one hand there are some systems which enable the entry of data relative to disaster management and make it possible to trace it afterwards (NEDIES, 2010). These systems are very general and have almost no specialization. They are at the level of chronicles or encyclopedias. There are even systems, which, through the combination of GPS and mobile phone applications, allow for the entry of information about the events right from the location (Ohya et al., 2007). Often these are only data storages which are publicly accessible but with no other use.

On the other hand, there are models individually specialized for each natural phenomenon - floods (Vondrak el al., 2008), landslides etc., possibly for prescriptions and methodologies of how to integrate more miscellaneous models and to create joint interface such as OpenMI (Gregersen et al., 2007). Nevertheless, these systems are not publicly accessible.

Our aim was to create a hybrid system which would both enable the collection of data from the professional and non-professional public, and would be capable of communicating with the other types of systems, to utilize and process their data, and on the basis of this collected data would be capable of predicting new dangers.

Through research it was found out that advantages of social networks have been more and more utilized during natural disasters as a way of communicating important information (Palen et al., 2007). That information can be broadcasted quicker than by way of other news media. Faster communication of information could be the key point for protecting or even saving the lives of the people living in the area affected by a disaster.

It could seem that information provided by people in social networks may not be exact or reliable, but in the fact it is just the opposite. News coming from the administration authorities and big news media are often deliberately distorted (Palen et al., 2007). But it is not only because of that that social networks can help during the crises; they can help to bring people together through the flow of intensive information, and people can solve the problems together and also better resist the catastrophes and recover from their effects (Palen et al., 2007). Besides social networks such as Facebook, Twitter, MySpace and others that are used to that purpose, there is also the emergence of social networks specifically designed for crisis situations.

The first example of them is the social network IGLOO (IGLOO, 2007) which together with its members coming from more than 200 countries and the global connection of varied organizations aiming to solve complicated problems. Recently this network has connected more than 200,000 research workers, academics and specialists from various spheres including education and administration worldwide. Thus, it improves communication and co-operation which results in the better effectiveness of the crisis management and of Rotary's reaction to the natural disasters.

Another social network is the Gustav Information Center, which was created in the days when Hurricane Gustav was approaching the Gulf of Mexico. This social network provided people with the necessary information to assist in organizing the help of volunteers and also in the evacuation of people before the storm. The network included links to other resources, as well as lists of volunteers, evacuation routes, blogs, photos and videos from the Gulf Coast, and many more resources to help during the hurricane (Edwards et al., 2009).

Microblogging and the Microsoft Vine warning system are other tools which will help to predict disasters such as Hurricane Katrina, earthquakes, pandemics or to manage critical and emergency situations of any kind (MS Vine, 2010). People will choose the field of certain problems and then they will be informed by way of short status messages and security alerts by using either a Vine desktop client or via email. The client will be able to link-up with Facebook and Twitter and will also have the opportunity to monitor the location of his or her relatives on a map background in a similar manner as Google Latitude function in mobile Google Maps for smart phones. Vine is currently available only in America. It should be a sort of system of last communication, which people would use to communicate with family when the telephone or mobile network fails (MS Vine, 2010).

SYSTEM OVERVIEW

In the following paragraphs we will describe a system prototype we have made which currently runs in the testing regime and which is already capable of receiving the inputs from the users. That system is based on the Case-Based Reasoning methodology (Aamodt et al., 1994) and it is described below.

System architecture

An essential element of the structure is characterized as a vertical cross-section of the entire architecture which consists of these six layers: The *user* enters a *natural phenomenon* (from a particular *territory* that has caused

some *damage* or *impairments*) into the system. The way the situation is resolved then becomes more *reliable*.

Based on this collected data, the system is capable of deriving solutions for new cases via Case-Based Reasoning methodology. The accuracy of estimation will be in direct proportion to the increase in the number of cases already correctly entered into the system.

Phenomena

One of the key elements of the system is natural phenomena. Based on a specific phenomena's strength power in given territory, the system is able to search existing cases of similar situations from which potential damages and consequences may be predicted. We have decided that the phenomena in the system will be divided into a hierarchy and stored in the database in order to develop a system modularly and independent of the phenomena. This structure will clarify individual categories making them easier to work with.

Based on consultations with experts, we created a tree of categories. To distinguish between the strength of individual phenomena, we have divided each into several degrees of intensity. These degrees of intensity were determined based on consultations with experts or with the aid of tested methods for gauging the strength of a phenomenon.

A weighted vector is a part of every end category of phenomena and identifies the importance of individual aspects when comparing two cases. Each event includes three weighted vectors that form a hierarchical structure. At the micro level, there is a vector for the surface according to the specifications of Corine Land Cover (Bossard et al., 2000). Seven components with a strong emphasis on the specification of the surface are in the middle part, namely:

influence of the location of areas of interest, influence of Corine Land Cover as a complex, influence of river networks, influence of the slope, influence of the orientation, influence of altitude, influence of neighboring territories.

The highest weighted vector (or vector for the whole case) then determines: weight of strength significance, weight of phenomenon duration, weight for similarity (on which the phenomenon operates).

Territory

Exposure of phenomena to a certain type and composition may naturally have the effect of natural and environmental disasters. As already mentioned, this system will serve to prevent and mitigate these disasters. The aim of this section is to find an efficient method for describing a territory.

The territory is separated into two parts. The first part (raster), determines the composition of a given area or, more specifically: the percentage of individual components of the surface of a given territory. The second part (vector), simply describes the river network, which plays a major role in the operation of any element.

The Corine Land Cover method, which offers various types of surfaces and their locations, is used to describe the territory.

The approximate structure of water is characterized by two indicators - nodes and flows.

Nodes: nodes are places where the river branches, runs, or breaks at an angle exceeding the limit angle.

Flows: flows are used as vectors starting in FNODE and orientation to TNODE. FNODE and TNODE are labels of individual nodes.

The territory described by a user will be expressed as a vector containing the specific parameters for a given area (the vector describing the territory), and will be presented in the following format.

The first part of the vector is location, to be determined using the S-JTSK coordinate system (CUZK, 2010). The second part of the vector is created by using Corine data, which is made up of individual components of Corine Land Cover, and expressed in percentages (or values in the range of 0-1). These values are identified for the particular field of interest as well as its surroundings. This may affect the territory in which there has been a phenomenon. The third part is the vector river network. As is the case with Corine data, the river network consists of two parts - the area of interest and its surroundings.

River inflow and outflow, river junctions, river branching, and river segments are all searched in a given territory. The slope, orientation, and digital terrain model is used for a more detailed description of the field. These properties constitute the last three attributes of our vector (see Fig. 1).

Solutions, the consequences and damage

The solution is a set of measures and other actions bound to a specific phenomenon, leading to the minimizing of consequences or damage. The term "solution" is therefore used to present sandbags, for example, which are used to prevent the flow of water during floods. Since our goal was to create an intelligent system, we did not settle for a mere determination of whether or not a given solution was used in a specific case.

This is why we define the structure as illustrated in Figure 2.

The solution is composed of indicators assigning its jurisdiction to the phenomenon and mainly outlines potential values that may be assigned during this step. If we define the above example with sandbags as a solution (field values), we also determine the recommended height for stacking such bags.

The solutions defined as above are only general rules for a situation (the mere abstraction). To be able to convert this information into a "tangible" form, we define a particular solution that represents the specific action in specific circumstances. The main media information is made up of attributes: the success of solution and index value. Since the minimum and maximum for

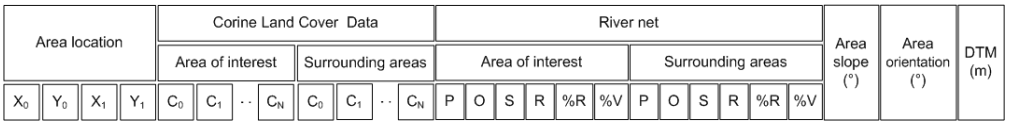


Figure 1: Data structure of the vector

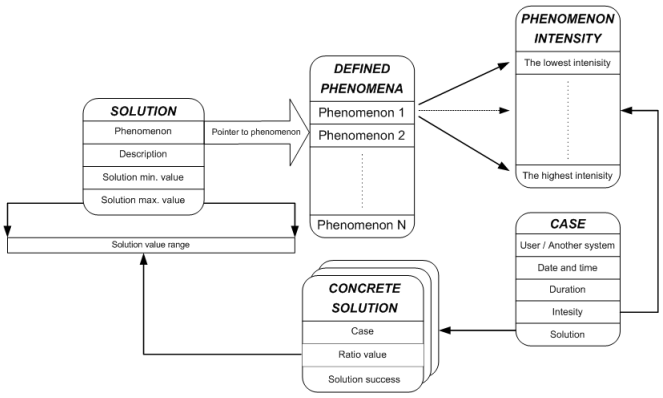


Figure 2: Block diagram of follow-up data on the subsequent derivation of a solution

each solution is defined, the following formula is sufficient: $RealValue = Solution_{min} + RatioValue * (Solution_{max} - Solution_{min})$.

The value ratio was introduced in order to unify the scale of all phenomena, thus providing an improved overview. Maximum and minimum values may obviously provide inconsistent values, thus, the introduction of new, specific solutions to exceed the scale interval is automatically extended and its values are converted.

Consequences and damage. A formal definition of consequences and damages is no longer necessary. Now unwanted phenomena occurring after exposure to a natural element, despite all efforts and measures taken to prevent it, becomes the priority.

Derivation and solution. The system DIP which we develop is not only a warehouse of recorded cases of the past, but is able to intelligently on the basis of experience to derive a solution for the situation which currently threatens. Selecting the most appropriate action is carried out using the methodology of CBR.

Activity (score) of the searched phenomenon can be summarized as follows:

1. The user chooses a list of phenomenon that directly threatens the field of action, intensity, estimated duration and other parameters, described in the scenarios above.

2. The system then monitors all cases, which relate to similar situations, with corresponding information describing the territory.

3. The system monitors all solutions used for the given phenomenon.

4. The deductible matrix is created and evaluates a list of appropriate consequences and damage.

5. The deductible matrix is created and evaluates a list of appropriate measures.

A detailed approach of individual items is described in the following paragraphs.

Observations of similarities. The first step for deriving an appropriate solution is the calculation of similarity (Watson, 1997) found in sample cases which the user entered. This number is expressed in the interval $CaseSimIndex \in \langle 0, 1 \rangle$, where 0 indicates the maximum diversity and the number 1 symbolizes the identity. The similarity is now calculated on two levels as defined by the phenomenon and its weight vector.

The landscape and its surface structure are derived at the lowest level of similarity, creating the attribute $LandSimIndex \in \langle 0, 1 \rangle$ unit scope. This value, combined with the duration and intensity of the phenomenon, determines the final congruity of the entire case $CaseSimIndex(CSI)$

$$CSI = SimilarityVector \times WeightedVector, \quad (1)$$

where $SimilarityVector$ consists of the following members:

$LandSimIndex$**:** similarity of the territory,

$WeightedVector$**:** similarity of the length of exposure to the phenomenon.

An important part of the formula is $WeightedVector$ where its components determine how important different parameters of vector similarity $SimilarityVector$ are

for the calculation. This enables us to easily determine which components can be ignored and which we need to highlight, thus speeding up the calculation.

Creation of deductive matrix. Once we have traced cases and evaluated their similarity, we can create a deductive matrix that will help us to derive specific solutions. The matrix has a form which is shown in Figure 3.

Rows form our searched cases, while columns define all the solutions used for the phenomenon. The last line of the matrix presents the ideal case.

The calculation of the ideal values. Each particular solution is a system characterized by the proportionate value "success" which is defined as the number $SolutionSuccess \in \langle 0, \infty \rangle$. A value of less than 1 indicates that the solution was insufficient, whereas a higher index indicates an unnecessary waste of resources. Ideally calculated values then become the basis upon which the entire algorithm works - seeking solutions for crisis situations.

Derivation of the ideal solution is carried out as follows:

1. For each specific solution in the current column of a deductive matrix:
   (a) Calculate the ideal value and solution for the given event.
   (b) Add to the calculated value the potential impact of a phenomenon's duration.
   (c) Add to the calculated value the potential impact of exposure to a phenomenon.
   (d) The system derives a weight for the similarity of the earth's surface to surface similarities of both cases and the weight vector.
   (e) The system derives a calculation for the total weight of the final solution.
2. The system calculates the recommended weight average using collected data and adds the item to a list of recommended measures for the situation.

The calculation of probable consequences. Even after measures are evaluated, it is still necessary to alert the user of the consequences of this type of phenomenon and to what extent it is likely that this situation will affect the user. Derivation is equivalent to the calculation of the recommended solutions of a deductive matrix; the calculation is simplified by the fact that consequences are monitored only as a binary value (whether it happened or not). The result is the number of operations in the interval $AfterEffectProbability \in \langle 0.1 \rangle$, where we obtain the probability of % after we multiply the number by 100.

CASE STUDY

In the system there are inserted data which you can see in the table 1 in a very simplified sample. The table shows a few columns of the projection of many attributes inserted into our case database.

The sample consists of geographical location (latitude, longitude), time specification of the phenomenon (start time and duration), phenomenon details (type and intensity) and a group of affected objects.

An actual version of the DIP system user interface is a simplification of the original DIP design which was presented in (Martinovic et al., 2009). Deployment into the real conditions shows that users are not able to describe situations by specifying several constants and generally by a set of numbers.

The number of details users had to specify was too high and we think that they were intimidated by it. So we decided to adapt input form the capabilities of an amateur community.

The results of our efforts can be seen in the following screenshots. The Figure 4 shows the main application window where you can see inserted events symbolized by pushpins on the map. Events may by filtered and sorted in several ways, such as location or time.

The user can click on one of these pushpins to display a detailed information window of a selected event. The event can be modified, or additional information can be added in this window (see Fig. 5).

However, the user can include his or her own event description (see Fig. 6) with various details of the events, including photos or other additional content.

CURRENT WORK

Tests show (Martinovic et al., 2009) the system provides satisfactory results for the cases where individual events have no time succession, and therefore this succession may not be essential. In that case we are satisfied with the basic techniques of information retrieval and statistics (Bjarne et al., 2003). During the practical application of a DIP system, the need to process not only a single moment but whole time intervals representing individual situations arose. Nonetheless, some information cannot be described by a single time moment; possibly this single moment would give almost no evidence.

Our aim is to provide users with a freely accessible tool which would be capable of helping the public with a progressive description of an event which would last longer and would be inserted into the system. It should be satisfied in the simplest way. These situations will be progressively put together from the individual snapshots and each user will be allowed to enter some parameters, and thus gradually help to describe the whole situation.

We do believe that the social network which will arise within the system will achieve to sufficiently motivate the user to enter details of the events. The exact information in the numerical form, verbal description as well as illustrative pictures will be welcome, both coming directly

	SOLUTION S_1		SOLUTION S_N
CASE C1	Concrete solution C_1S_1		Concrete solution C_1S_N
CASE C2	Concrete solution C_2S_1		Concrete solution C_2S_N
CASE CM	Concrete solution C_MS_1		Concrete solution C_MS_N
IDEAL CASE	**Derived concrete solution 1**		**Derived concrete solution N**

Figure 3: Appearance of deductive matrix

Figure 4: Main application window

from the users or, for example, automatically from other systems. For longer lasting phenomena (meant longer than one moment) the term episode was introduced by (Sanchez-Marre et al., 2005). Each of these episodes consists practically of 2...n cases (so called snapshots) which has the form of film tape consisting of single photos. At least one parameter needs to be fulfilled on each snapshot (see the Fig 7).

It is evident that the denser the series of shots is, the more exact picture about the situation we will have. Via the suitable interpolation or approximation of the points (see the Fig. 8) we are capable of getting close to a real picture of the situation (respectively about the progress of an individual phenomenon) which is inserted into the system.

This extension of the situations' descriptions leads to a change in the requirements of saving the data and especially their extraction from the database. Thus CBR methodology for these kinds of situations is replaced by its extension called Episode Base Reasoning (Sanchez-Marre et al., 2005), which operates with only these episodes. In the case of successfully obtaining the episodes, the knowledge from the field of time series could be used to add precision to the prediction. At the same time we could thus discover a suitably derived parameter for better indexing and tracing of the episodes.

DISCUSSION

Besides the comments directly inserted by the users, the system could be suitably extended by the automatic download of additional information. Data could come

Table 1: Case sample data

Case name	Lat. Long.	Duration Date	Ph.type Ph.strength	Aff.buildings Crisis staff int.	Aff.util.lines Aff.comm.
Flood in Ceske Budejovice city,	48.98	5 days	3	false	true
the Malse River flooded	4.47	7.8.2002 17:00	4	true	true
Huge rainstorm, the Granicky	48.87	2 hours	5	false	false
brook flash-flooded.	15.95	2.5.2004 20:00	2	true	true
The town of Znojmo,	48.86	3 days	3	true	true
part of the town flooded.	16.05	28.3.2006 17:00	3	true	true
Dam break in Horce n. Moravou,	49.64	3 hours	3	false	true
local inundations.	17.21	2.4.2006 7:00	3	true	true
Senov u Noveho Jicina, intensive	49.59	3 hours	5	true	false
rainstorm, flooded houses.	18.02	24.6.2009 20:00	4	true	true
...	...	...	...	...	...
...	...	...	...	...	...
The town of Bezkov, intensive	48.86	3 dny	3	true	true
snowmelt, brook flooded.	16.05	28.3.2006 17:00	3	true	true

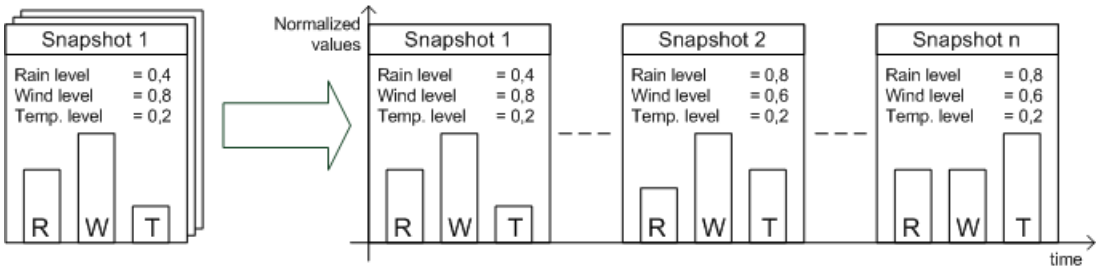


Figure 7: Snapshots collection mapped to a timeline

from measurement and observation gauges as well as from other systems. Suitable parameters could be, for example, precipitation depth, temperatures etc. coming not only from professional gauges but also from the amateur meteorological stations which often publish their measurement in accessible sources.

As was stated in the Introduction, the goal of the work was the development of a hybrid system combining the possibilities of the two disaster management approaches discussed above. It will be accomplished by the connection of the DIP system with the FLOREON$^+$ system (Unucka et al., 2009; Vondrak el al., 2008; Martinovic et al., 2008).

Information about the events manually inserted into DIP by its users will be continually and without the necessity of the operation updated by FLOREON$^+$ system, which could be thus understood like another - automatic - user of the DIP system. FLOREON$^+$ is a complex and modular system for hydrologic and environmental modelling (Unucka et al., 2009). System FLOREON$^+$ partly disposes of automated hydro-meteorological data collecting, partly of the automated computational cascade of rainfall-runoff and hydrodynamic models. The hydro-meteorological data is collected from the network of gauges professionally run by the Czech hydrometeorological institute and the Povodi Odry state enterprise (the River Odra basin management state enterprise) and are used as the inputs of the hydrologic models. On the side of meteorological inputs, data about precipitation depth is most important, followed by data of air temperature and snow pack (thickness and water equivalent). The time step of the obtained precipitation depth and temperature data is 1 hour. On the side of hydrological inputs the data about hourly discharges at the hydrologic gauges are obtained as well. This discussed data is used within the FLOREON$^+$ system as inputs of the calculation of the runoff response to causal rainfall (possibly runoff caused by snowmelt) or as the reference data to optimize the model parameters (to model calibration). The outputs of these models (hydrographs, also in the profiles over and above the professionally monitored gauges and gauges of professional hydrologic prediction) together with the observed discharges are then used as the inputs to the hydrodynamic models which solve the water routing in the riverbeds, and possibly outside the riverbeds during the floods. One of their outputs is the spatial localisation of potential flood lakes. The whole cascade including the postprocessing of the models' outputs and the web visualisation of the results is the constituent of a fully automated server solution and

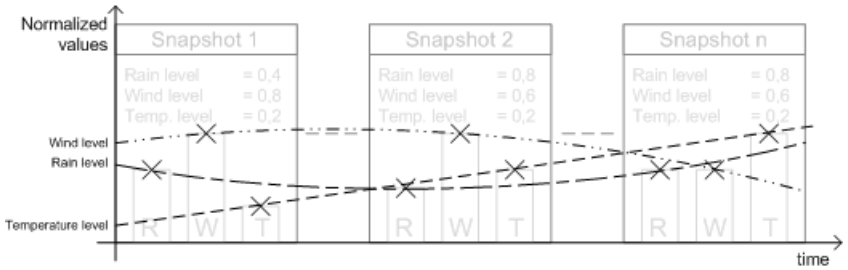


Figure 8: Interpolated shapshots values

Figure 5: Event details window

Figure 6: New event details inserting

nowadays its outputs are used as a basis for decision-making within the disaster management. In the future the FLOREON$^+$ system outputs should be among others used as an automatic user of DIP system, which is intended for the level of the near real-time information (update of hydro-meteorological data measured at the gauges, signalisation of achievement or exceedance of the flood emergency degrees etc.) as well as on the level of hydrologic situation prediction (hydrographs and flooding predictions, prediction of achievement or exceedance of the flood emergency degrees etc.).

The classic imperfection of the professional gauges' networks is their spatial density, which can be insufficient for some purposes (e.g. estimation of convective rainfall effects in the landscape etc.). A possible way to solve these problems can be the use of the precipitation radar measurement outputs. At the level of the gauge network the solvers team works on the automatic collection of the data measured by amateur meteorologists. Organized efforts in linking the hobbyist and the nation's demand for weather data is not new (e.g. (Bristow et al., 2005; COOP, 2000)). These meteorology enthusiasts could supplement the professional gauge network and they could serve as another user with automatic data update to the DIP system and its social network. On top of that they could provide input data for the hydrologic models of FLOREON$^+$ system.

It is evident that data obtained by the amateur meteorological stations measurements are tainted by some amount of uncertainty which could often be quite significant. Measurement of the meteorological factors at the professional level follows the strictly given rules and

standardised procedures and the gauges are calibrated. Localization of these gauges is not random. Amateur meteorological stations can hardly fulfil these conditions. However this is not a problem that could impede their implementation into the DIP system. The information coming from these kinds of users can very suitably and effectively supplement the knowledge of actual conditions, particularly during spatially limited events and possibly localised between the points of the professional observation network. In the question of their potential use within FLOREON$^+$ system the solutions team is aware of the necessity of the homogenisation of this data coming from two different measurement sources (professional and amateur). Thus the use of the amateur data will first need to be carefully and theoretically elaborated and practically tested besides the official use and outputs of FLOREON$^+$ system.

ACKNOWLEDGEMENT

We acknowledge the support of projects SP/2010196 Machine Intelligence and SP/2010101 Integration of GIS and numerical models for analyzing the vulnerability and operational crisis management in relation to selected natural and anthropogenic hazards.

REFERENCES

Aamodt, A. (1994). Case-Based Reasoning: Foundational Issues. Methodological Variations & System Approaches, AI Communications, Vol. 7 Nr. 1, March 1994.

Bedient, P.B., Huber, W.C. and Vieux, B.C. (2007). Hydrology and floodplain analysis. 4th edition, Prentice Hall, London, 795 p., 2007.

Bjarne, K. Hanse and Riordan, D. (2003). Fuzzy Case-Based Prediction of Cloud Ceiling and Visibility. Dalhousie University, Halifax, NS, Canada, 2003.

Bossard, M., Feranec, J. and Otahel, J. (2000). CORINE land cover technical guide. Addendum 2000.

Bristow, S.R., Leiker, K., DeBlois, R. (2005). Emergency preparedness and tourism: both win with help of the amateur meteorologists. In Proceedings of the 2005 Northeastern Recreation Research Symposium. Bolton Landing, NY, USA, 135-139.

COOP - Cooperative Observer Program 2000.
URL: http://www.coop.nws.noass.gov (23rd Feb. 2010).

CUZK, Czech office for surveying, mapping and cadastre
URL: http://www.cuzk.cz/ (23rd February 2010).

Edwards, C. (2009). RESILIENT NATION, London: Demos 2009

Gregersen, J.B., Gijsbers, P.J.A. and Westen, S.J.P., (2007). OpenMI : Open Modelling Interface. Journal of Hydroinformatics, 9 (3), 175-191.

IGLOO (2007). URL: www.igloosoftware.com/company (22nd February 2010)

Ishioka, T. (2003). Evaluation of Criteria for Information Retrieval. The National Center for university Entrance Examinations, Japan, 2003.

Jonov, M., Unucka, J. and Zidek, D. (2008). The comparison of two floods in the Ole catchment - the possibilities of hydrological forecasting with the use of radar products. Fifth European Conference on Radar in Meteorology and Hydrology ERAD 2008, Helsinki Finland, 2008.

Martinovic, J., Stolfa, S., Kozusznik, J., Unucka, J. and Vondrak, I. (2008) Floreon – the system for an emergent flood prediction. In ECEC-FUBUTEC- EUROMEDIA, Porto, April 2008.

Martinovic, J., Kocyan, T., Unucka, J., Vondrak, I. (2009). FLOREON$^+$: Using Case-Based Reasoning in System for Flood Predictions. Disaster Management and Human Health Risk: Reducing Risk, Improving Outcomes, ISBN: 978-1-84564-202-0.

Microsoft Vine, social networking for crisis management, URL: www.techshout.com/internet/2009/30/microsoft-vine-social-networking-tool-for-crisis-management/ (22nd Feb. 2010)

Microsoft Vine, URL: http://www.vine.net/default.aspx (22nd February 2010)

NEDIES - Natural and Environmental Disaster Information Exchange System. URL: http://nedies.jrc.it (23rd Feb. 2010)

Ohya, M., Asada, J., Harada, Matsubayashi, R., Hara, M., Takata, R. (2007). Disaster information-gathering system using cellular phones with a global positioning system. Ministry of Land, Infrastructure and Transport, Toshitaka KATADA, Gunma University.

Palen, L., Vieweg, S., Sutton, J., Liu, S.B., Hughes, A. (2007) Crisis Informatics: Studying Crisis in a Networked

Sanchez-Marre, M., Cortes, U., Martinez, M., Comas, J. and Rodriguez-Roda, I. (2005). An Approach for Temporal Case-Based Reasoning: Episode-Based Reasoning. Springer Berlin / Heidelberg, ISSN 0302-9743.

Sun, Z., Finnie, G. and Weber K. (2001). Integration of abductive CBR and deductive CBR. Sch. of Inf. Technol., Bond Univ., Gold Coast, Qld., 2001.

Unucka, J., Martinovic, J., Vondrak, I., Rapant, P. (2009). Overview of the complex and modular system FLOREON+ for hydrologic and environmental modeling. In River Basin Management V. C.A. Brebbia (Ed.), WIT Press, Southampton, UK.

Vondrak, I., Martinovic, J., Kozusznik, J., Unucka, J. and Stolfa, S. (2008). FLOREON - the System for an Emergent Flood Prediction.. 22nd EUROPEAN Conference on Modelling and Simulation ECMS 2008, Nicosia Cyprus, 2008.

Watson, I. (1997). Applying Case-Based Reasoning: Techniques for Enterprise Systems. Morgan Kaufman, 1997.

Watson, I. (1999). Case-based reasoning is a methodology not a technology. AI-CBR, University of Salford, Salford M5 4WT, UK

STRUCTURAL COMPRESSION OF DOCUMENT IMAGES WITH PDF/A

Sergey Usilin
Innovations and High Technology Department
Moscow Institute of Physics and Technology
9, Institutskii lane, Dolgoprudny, Russia, 141700
usilin.sergey@gmail.com

Dmitry Nikolaev
Institute for Information Transmission Problems,
Russian Academy of Sciences
19, Bolshoy Karetny lane, Moscow, Russia, 127994
dimonstr@iitp.ru

Vassili Postnikov
Institute for System Analysis,
Russian Academy of Sciences
9, 60-letiya Oktyabrya ave., Moscow, Russia, 117312
vassili.postnikov@gmail.com

KEYWORDS

Document image processing, data compression, morphological operations, color saturation

ABSTRACT

This paper describes a new compression algorithm of document images based on separating the text layer from the graphics one on the initial image and compression of each layer by the most suitable common algorithm. Then compressed layers are placed into PDF/A, a standardizated file format for long-term archiving of electronic documents. Using the individual separation algorithm for each type of document makes it possible to save the image to the best advantage. Moreover, the text layer can be processed by an OCR system and the recognized text can also be placed into the same PDF/A file for making it easy to perform cut and paste and text search operations.

INTRODUCTION

Scanning is the most popular method of document conversion to electronic forms nowadays. However, images resulted from high quality scanning have large size and are not effective in electronic archives. Common techniques of image compression are not applicable because documents may contain monochrome text and full-color graphics at the same time. Lossless image compression algorithms which are effective for monochrome texts are ineffective for full-color graphics, while lossy image compression algorithms show good results for color images, but can corrupt text information.

A combined approach can be used to compress document images. The idea of this approach consists in extracting structural blocks in the image, combining these blocks into layers (i.e. image "separation" into textual, graphic and other layers), and compressing each layer using the most appropriate technique.

One of the structural document compression methods is implemented in the DjVu format (Lizardech 2005, Haffner et al. 1998). But in spite of high ratios of the document image compression, the DjVu-format has essential disadvantage: this format is not standardizated therefore its usage for creating electronic archives is complicated. Besides, using the same separation technique for all image types is not always worthwhile and may lead to significant corruption of the document.

This paper proposes a new method of structural compression document images into PDF/A (ISO 2005, Rog 2007), a standardizated file format. The method contains a set of image separation algorithms typical for images of different document types.

FORMAT PDF/A

PDF/A is a file format for long-term archiving of electronic documents. It is based on the PDF Reference Version 1.4 from Adobe System Inc. and is defined by ISO 19005-1:2005. It ensures the documents can be reproduced the exact same way in years to come. All of the information necessary for displaying the document in the same manner every time (all content, fonts, colors and etc.) is embedded in the file. These features of PDF/A make it possible to use it as a major file format in the electronic archives.

STRUCTURAL COMPRESSION ALGORITHM

In this section we describe briefly a scheme of the proposed structural compression algorithm (Figure 1). The algorithm is meant for compression of document images (images of financial documents, books and magazines pages, manuscripts, etc.). For each type of document there exists a unique separation algorithm. Therefore, first of all, it is necessary to determine the type of the source image and choose the appropriate separation algorithm. Using the chosen algorithm we separate the source image into text and graphics layers. According to the method architecture the text layer has only parts of the source image such that correspond to real text information. Hence it is easy to recognize this

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

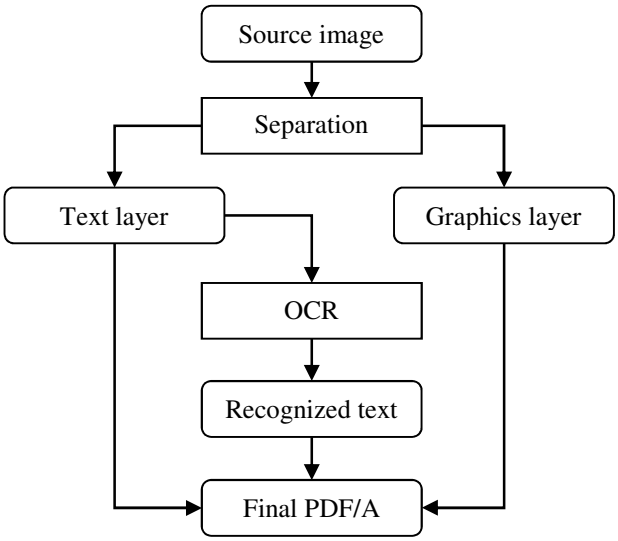


Figure 1. The scheme of the structure compression algorithm

layer by any OCR (optical character recognition) system. The obtained recognized text can be useful for seeking documents in electronic archives. Then both text and graphics layers as well as the recognized text are placed into a PDF/A file in a special way.

Thereby the main parts of the proposed algorithm are the image separation technique and the way of placing obtained layers and recognized text into PDF/A. These parts are described in the following sections.

IMAGE SEPARATION

Besides existence of text information, each document type has its own features. For example, financial documents usually have a lot of stamps and signatures; illustrated magazines can contain a complex multicolor background; books often include formulas, schemes, and diagrams. Therefore the proposed algorithm provides a unique separation method for each type of document. We would like to consider the separation algorithms for two practically important document types: books and financial documents.


Задачи 1 — 4 собраны из старых номеров «Кванта». Их решали младшие школьники, которые уже стали младшими и старшими научными сотрудниками.

3. Десять человек пришли в гости в галошах. Уходили они по одному, и каждый надевал произвольную пару галош, в которую он мог влезть (т. е. не меньшего размера, чем его собственная). Какое наибольшее число людей не смогло надеть галоши?
(«Квант» №5 за 1978 г.)

4. Из спичек сложено слово «ТОЛЯ»:

Figure 2. An example of a common book page image

An image separation of a common book page

A book page usually contains black text on white background and, possibly, schemes, diagrams, etc. (Figure 2). The areas with text and graphics usually are not intercrossed in the books. The second key feature of the book makeup is usage of the fonts of similar linear sizes.

Taking into account these features, we will build the algorithm of book page image separation.

Step 1. Binarize the initial image using one of global binarization methods, for example, Otsu method (Otsu 1979, Nikolaev 2003, Trier and Taxt 1995). As the image contains black text on white background, such binarization should not significantly corrupt text information.

Step 2. Apply mathematical morphology (Gonzalez and Woods 2002, Hasan and Karam 2000) to join each word in the text to connected components (Figure 3). By w and h denote typical width and height of the characters correspondingly. Let us note that the letter spacing is approximately equal to the stem width and the word spacing is comparable with the character width. Therefore words can be joined to connected components by applying mathematical morphology with a rectangular structuring element $w \times 1$.

This operation is not required large computation power because only one simple morphological operation (opening) is needed. Furthermore, using van Herk's algorithm (Van Herk 1992) allows computing any simple morphology operations with a rectangular structuring element for time independent of the size of primitive.

Step 3. Build a heights histogram of the found connected component (Figure 3). As characters in the

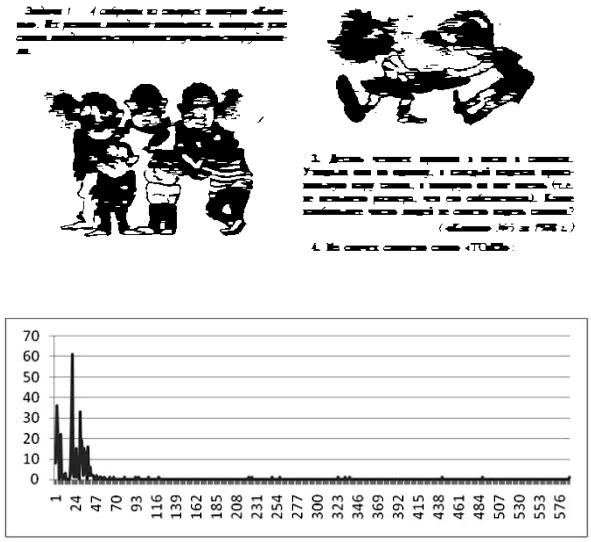


Figure 3.Connected components and a histogram of their heights

text on the image have similar size the connected components corresponding to the words shall be of similar height and create one or more maximums at the histogram (Fletcher and Kasturi 1988). Due to this it is possible to determine typical heights of the letters h_{font} and therefore find the areas of the source image corresponding to text blocks and areas corresponding to graphics and split the source image into the text and the graphics layers.

As we use fast algorithms of mathematical morphology with a rectangular structural element, the text blocks are strongly required to be aligned relatively axes of image coordinates. Therefore we will use the Fast Hough Transform (Nikolaev et al. 2008) to unskew the input image before applying morphological filtrations.

An image separation of a financial document

The above features of the book page image are not typical of the color images of financial documents (invoices, receipts, contracts, etc.) because the graphic elements (seals, signatures, manuscript notes) are often laid over the textual blocks (Figure 4). Therefore, it is unreasonable to use the above algorithm. We will build a separation algorithm based on color characteristics of the image. Color saturation of black text and white background is close to zero point, while color saturation of the blue seals and signatures is high. Taking these features into account we will build the following algorithm of separation for financial document image.

Step 1. Calculate color saturation (Gonzalez and Woods 2002) of each pixel on the source image using the formula:

$$s = \max(r, g, b) - \min(r, g, b)$$

Here r, g and b are red, green and blue components of the pixel concerned. It is obvious that s can take value between 0 and 255.

Step 2. Build a saturation histogram $y = \log N_x$, where N_x – number of the pixels whose saturation equals x (Figure 5).

Figure 4. An example of financial document

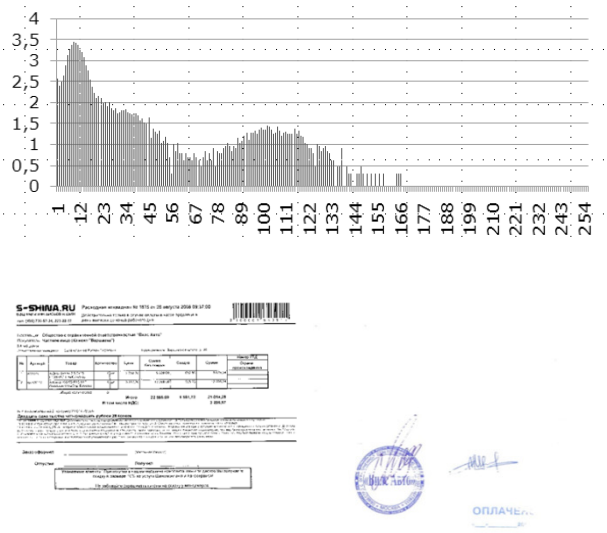


Figure 5. A saturation histogram and text and graphics layers after separation

Step 3. It is easy to notice two classes in this histogram: the first class is formed by pixels with low value of color saturation and the second one – by pixels with high value of color saturation. That is the first class corresponds to background and text areas in the source image and the second class corresponds to graphics areas. Find the threshold t^* between these classes by Otsu method (Otsu 1979).

Step 4. If the obtained threshold value t^* is quite small (i.e. if t^* less than some in advance defined minimal value t_{min}) than the source image contains only black text and has no graphics. So the source image coincides with the text layer and the graphics layer is empty. Otherwise generate the following separation algorithm. For each pixel (x, y) calculate its saturation s and if $s < t^*$ then (x, y) is a part of text layer. Otherwise $s \geq t^*$ then (x, y) is a part of graphics layer.

An example of separation of a financial document image is shown in Figure 5.

PLACING LAYERS IN PDF/A

After the image is separated the text layer is binarized by Otsu binarization method. Two popular lossless algorithms, CCITT Group 4 and JBIG2 (ISO, 1993), were considered. In this case (compression of monochrome image with only text information) we chose the last one because JBIG2 typically generates files one third to one fifth the size of CCITT Group 4.

The graphics layer is down-scaled to 100 dpi and compressed by the JPEG (ISO, 1994) with middle compression quality. Such compression makes it possible to reach small size without significant distortion of graphics.

As it has been mentioned in the brief description of the algorithm, the text layer can be easily recognized by any OCR system. We used OpenOCR engine which is an

omnifont multilingual open source system (OpenOCR.org). We split the recognized text into the words and place separately each word into PDF/A file using invisible style. In this case correct layout which could be left out during OCR processing will be established automatically by PDF parser program.

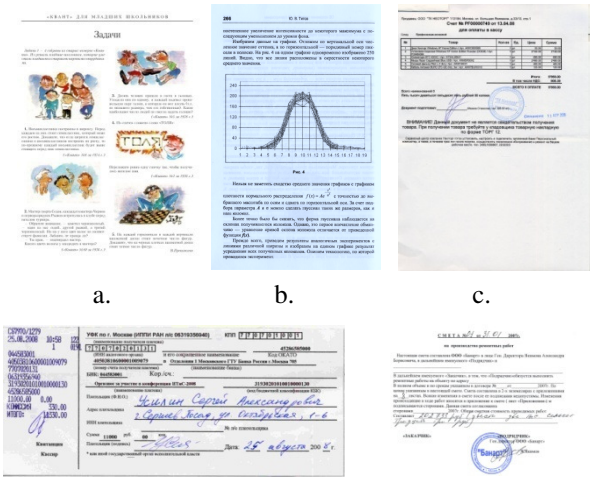
CONCLUSION

In the paper a new structural compression algorithm of document images is considered. Thanks to the using PDF/A as output format the compressed image can be used in electronic archives. Using different schemes of image separations makes it possible to save the document image face and achive the highest possible compression ratio. Including recognized text to the PDF/A file makes it easy to find and copy information in the documents.

Typically proposed algorithm compress document images at 300 dpi to 50-150 KB (approximately 3 to 10 times better than JPEG for a similar level of subjective quality). The following table presents the compression results of several images shown in Figure 6.

Table 1. Compression results

Image	JPEG	DJVU	PDF/A
Figure 6a	718 KB	50 KB	94 KB
Figure 6b	373 KB	29 KB	51 KB
Figure 6c	642 KB	23 KB	52 KB
Figure 6d	120 KB	20 KB	40 KB
Figure 6e	292 KB	17 KB	29 KB



a. b. c.

d. e.

Figure 6. Examples of different document images

Significantly difference in sizes between DJVU and PDF/A files is explained by the fact that besides the useful information (images and text) PDF/A file contains also some service information (fonts, color profiles, metadata and others).

In case of automated compression of a set of different type documents, automatic selection of the appropriate separation algorithm is necessary. This task is solved with the help of methods of preliminary determination of document type.

The considered technique of structural compression is implemented as a program and is embedded in Magnitogorsk Iron and Steel Works Open Joint Stock Company (MMK) workflow system.

REFERENCES

Fletcher, L.A. and R. Kasturi. 1988. “A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.10, No.6, 910-918.

Gonzalez, R.C. and R.E. Woods. 2002. *Digital Image Processing, second edition (2nd Edition)*. Prentice Hall.

Haffner, P.; L. Bottou; P.G. Howard; P. Simard; Y. Bengio; and Y. Le Cun. 1998. “Browsing through High Quality Document Images with DjVu”. In *Proceedings of the Advances in Digital Libraries Conference*. IEEE, Washington, DC, USA, 309-319.

Hasan, Y.M.Y. and L.J. Karam. 2000. “Morphological Text Extraction from Images”. *IEEE Transactions on Image Processing*, Vol.9, No.11, 1978-1983.

International Organization for Standardization. 1994. *Information technology – Coded representation of picture and audio information – Progressive bi-level image compression (ISO/IEC 11544)*.

International Organization for Standardization. 2005. *Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A) (ISO 19005-1)*.

Lizardech. 2005. *Lizardtech DjVu Reference*.

Nikolaev, D.P. 2003. “Segmentation-based binarization method for color document images”. In *Proceedings of 6th Open Russian-German Workshop on Pattern Recognition and Image Understanding*, 190-193.

OpenOCR.org. *http://en.openocr.org/*.

Organization for Standardization. 1994. *Information technology – Digital compression and coding of continuous-tone still images: Requirements and guidelines (ISO/IEC 10918-1)*.

Otsu, N. 1979. “A threshold selection method from gray-level histograms”. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol.9, No.1 (Jan), 62-66.

Rog, J. 2007. *PDF Guidelines: Recommendations for the creation of PDF files for long-term preservations and access*.

Trier, P.D. and T. Taxt. 1995. “Evaluation of binarization methods for document images”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.17, No.3, 312-315.

Van Herk, M. 1992. “A Fast Algorithm for Local Minimum and Maximum Filters on Rectangular and Octagonal Kernels”, *Pattern Recognition Letters*. Elsevier Science Inc, NY, USA, 517-521.

Nikolaev D.P., S.M. Karpenko, I.P. Nikolaev and P.P Nikolayev. 2008. “Hough Transform: Underestimated

Tool in the Computer Vision Field". In *Proceedings of 22nd European Conference on Modelling and Simulation (ECMS 2008)*. Dudweiler, Germany, 238-243.

AUTHOR BIOGRAPHIES

SERGEY A. USILIN was born in Sergiev Posad, Russia. He studied mathematics and physics, obtained his Master degree in 2009 from Moscow Institute of Physics and Technology (MIPT). Since 2009 he has been a Ph.D. student at Moscow Institute of Physics and Technology. His research activities are in the areas of image processing and object detection. His e-mail address is usilin.sergey@gmail.com.

DMITRY P. NIKOLAEV was born in Moscow, Russia. He studied physics, obtained his Master degree in 2000 and Ph.D. degree in 2004 from Moscow State University. Since 2007 he has been a head of the sector at the Institute for Information Transmission Problems, RAS. His research activities are in the areas of computer vision with primary application to colour image understanding. His e-mail address is dimonstr@iitp.ru and his Web page can be found at http://chaddt.net.ru/Lace/cv_ip/index.html

VASSILI V. POSTNIKOV was born in Sverdlovsk, USSR. He studied mathematics and physics, obtained his Master degree in 1990 from Moscow Institute of Physics and Technology (MIPT). His Ph.D. thesis was named "Structural documents identification and recognition". He obtained his Ph.D. degree in 2002 in the Institute for System Analysis (ISA RAS). Since 2006 he was deputy head of the Cognitive technologies department at MIPT and leading scientist at ISA RAS. His research activities are in the areas of document image analysis, processing and understanding. His e-mail address is vassili.postnikov@gmail.com.

A SIMULATION MODEL FOR THE WHOLE LIFE CYCLE OF THE SLIME MOLD *DICTYOSTELIUM DISCOIDEUM*

Matthias Becker
FG Simulation and Modeling
University Hannover
Welfengarten 1, 30167 Hannover, Germany
Email on www.sim.uni-hannover.de

KEYWORDS

biological simulation, Dictyostelium discoideum, slime mold

ABSTRACT

Slime molds are fascinating organisms, they can either live as an organism consisting out of a single cell or they can form a multi-cellular organism. Therefore from the biological point of view, the slime molds are studied in order to understand the evolutionary step from a single cell organism to a multi-cellular organism. Studies have shown that the behavior of cooperating single cell organisms exhibits synergistic emergent intelligence, for example finding shortest paths. Just recently, simulation and experiments with a real slime mold (*Physarum polycephalum*) have been used for traveling salesman like problems.

In this work we present a simulation model for the slime mold *Dictyostelium discoideum*. Different to other studies, here the whole life-cycle is modeled and simulated. Very detailed behavioral patterns and parameters are modeled and as result a simulation model is obtained, that shows a behavior very close to the living slime mold. This result is consolidated by extensive verification experiments. As consequence, this model can be used to further study the mechanism of cooperation of single cells, mechanisms of synergy and emergence, and additionally this model offers the possibility to develop more slime mold inspired algorithms.

MOTIVATION

Slime molds are fascinating organisms, that basically are living as single celled amoeba. If the environmental living conditions are bad (mainly if there is not enough food), then the individual amoeba act together in order to ensure the survival of at least some part of the population. In such situation the slime mold forms a social organism (pseudoplasmodium), where the amoeba act as *one* multicellular organism, that moves and reproduces.

Therefore from the biological point of view, slime molds are studied in order to understand the evolutionary step from a single celled organism to a multi-cellular organism. Studies have shown that the behavior of cooperating single celled primitive organisms exhibit synergistic emergent intelligence, for example finding shortest paths. Just recently, simulation and experiments with a real slime mold (*Physarum polycephalum*) have been used for traveling salesman like problems (16; 17). Thus from the computer science point of view, it is interesting to study and learn from the natural mechanisms, that lead to emergent intelligent behavior of a whole system that consists of simple parts.

For this reason we developed a simulation model for the slime mold *Dictyostelium discoideum*. Different to other studies, here the whole life-cycle is modeled and simulated. Very detailed behavioral patterns and parameters are modeled and extensive verification has been done in order to ensure that the resulting model is valid. As consequence, our model can be used in the future to further study the mechanism of cooperation of single cells, mechanisms of synergy and emergence, and to develop more slime mold inspired algorithms exhibiting self-X properties.

In the following the natural life cycle and the most important mechanisms of movement and reproduction of *Dictyostelium discoideum* are summarized. Then it is explained, which of the features are essential for describing the behavior of *Dictyostelium discoideum* and as consequence have been used in order to build a good simulation model. The simulation model is then verified by doing extensive experiments and comparing behavior and quantitative measures of simulation model and real *Dictyostelium discoideum*.

DICTYOSTELIUM DISCOIDEUM BASICS

The study of *Dictyostelium discoideum* has a long history and also the study of collective movement has been focus of research. Special interest was given to the interaction between different amoeba: How do relatively simple single cells communicate? And which mechanisms let them get together and act as a *one* multi-cellular organism?

Early studies about the basic behavior and mechanisms of *Dictyostelium discoideum* can be found in (1; 2; 3; 4). In (2) evidence is given for the chemotaxis bringing the amoeba together.

Summaries and a whole picture of *Dictyostelium discoideum* can be found in recent literature, such as (5) or on a dedicated web page (6).

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

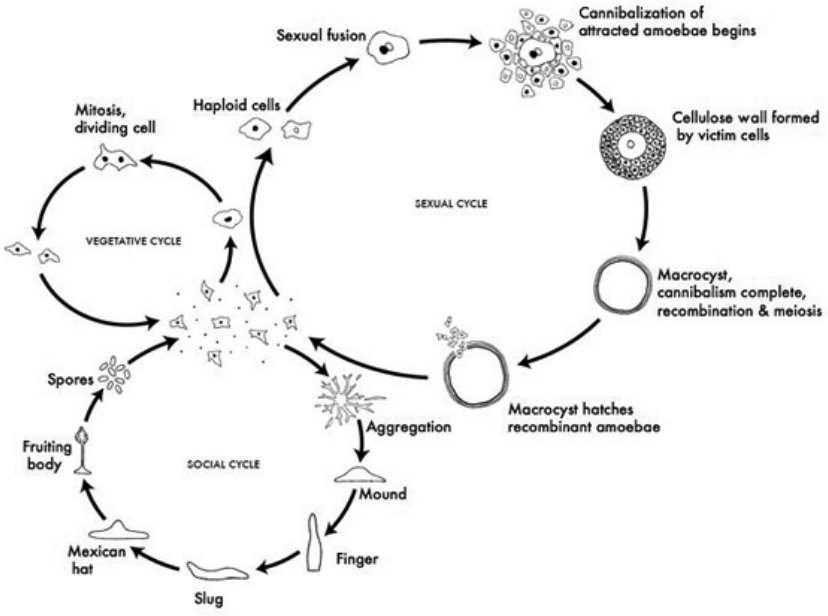


Figure 1: Life-cycle of *Dictyostelium discoideum*(creative commons license, by David Brown and Joan E. Strassmann)

Life-cycle of *Dictyostelium Discoideum*

In the biological taxonomy *Dictyostelium discoideum* belongs to the single celled *myxomycota* (= slime-molds), there to the *acrasiomcycota* (cellular slime-molds).

The normal appearance is a single celled amoeba with a size of 8 - 12 μm. Chromosome two (out of six, with ca. 8000 to 10000 genes) suggests that *Dictyostelium discoideum* has closer relationship to animals than to plants. The movement of *Dictyostelium discoideum* is a reaction to chemical stimuli (chemotaxis), the stimulus normally being the excrement of bacteria, the main food source of *Dictyostelium discoideum*. This attraction can span over 5 μm if many bacteria are present. A second stimulus are chemicals that are emitted by other amoeba that are hungry. If none of those stimuli is present the movement is directed by the next source of light and if the light is diffuse or non-existent, then the movement of *Dictyostelium discoideum* is random.

For the movement *Dictyostelium discoideum* builds pseudopodia.

The reproduction is mainly asexual by cell division (although also the possibility of sexual reproduction is possible, however rarely). Between two divisions an amoeba has to eat approximately 1000 bacteria, the final division needs a duration of approximately three minutes.

The life-cycle consists of three phases (see Fig. 1 for a detailed illustration):

- growth and reproduction
- interphase (phase of hunger)
- aggregation phase

If enough food is present *Dictyostelium discoideum* eats and reproduces and moves towards more food. If even by movement no more food can be found the amoeba's state changes into 'hunger' and emits cAMP (cyclic adenosine monophosphate, discovered by (2) in 1947 and identified 1967 by (3)).

Amoeba that are not hungry do not react on cAMP very much. While being hungry the reaction sensitivity to cAMP increases and with a small delay the cAMP concentration is amplified by hungry cells that, additionally to being hungry, sense cAMP.

By this stimulus, the amoeba enters the aggregation phase, which is characterized by all hungry amoeba coming together and that ends with building a fruiting body and finally releasing the spores. The minimum population density for aggregation is 400 amoeba per mm^2.

The aggregation phase can be divided in five subphases (see Fig. 2):

- Streaming
- Mound
- Slug
- Mexican hat
- Fruiting body/Sorokarp

The aggregation phase culminates the spore out of the fruiting body. Spores are spread out and carried by wind and other external factors, fall down and start growing new populations of *Dictyostelium discoideum*, the life-cycle starts again.

Figure 2: *Dictyostelium discoideum* from aggregation to spore out Copyright granted by, M.J. Grimson and R.L. Blanton, Biological Sciences Electron Microscopy Laboratory, Texas Tech University.

STATE OF THE ART

As early as in the 70s, researchers tried to find a mathematical description of the behavior of *Dictyostelium discoideum* such as in (8). Later on, especially the spiral geometry of the aggregation patterns have been of interest (9). In (7) a model for the individual as well as the collective movement of *Dictyostelium discoideum* has been developed.

However, each of the previous works only simulates *one* aspect of the life-cycle, neither work includes a simulation model for the *whole* life-cycle of *Dictyostelium discoideum*. This is what we present in the remainder.

SIMULATION MODEL

Our simulation model consists of a basically two-dimensional grid where amoeba can move around. In the grid also information about light and wind, food (bacteria) and chemical stimuli (cAMP signaling hunger, folic acid signaling food) is stored. The level of geometrical detail is one μm. The level of detail concerning the timing is derived from the speed of *Dictyostelium discoideum* which is two μm/min. Thus the smallest time duration is an interval of 30 seconds which we chose as cycle time of the simulation.

Vegetative Phase

In this phase, *Dictyostelium discoideum* moves around, searches food and reproduces, when enough food has been consumed. Since in the grid information about food and the corresponding chemical signal is stored, an amoeba can smell the direction of the nearest food and move in that direction. If no food is near, the movement is random.

If food is met, a waiting time corresponding to the time of phagocitosis (eating) is implemented. Since the real phagocitosis needs approximately 90 seconds, it corresponds to three simulation cycles. After that, the food disappears from the grid and the food counter of the amoeba having eaten is increased.

If enough food has been consumed (mean value for that is 1000 bacteria), a cell division is started. That means that the amoeba is divided into two amoeba being placed in the grid.

After a certain number of cycles (240 cycles corresponding to approx. two hours for real *Dictyostelium discoideum*) with no food the state will be changed to hungry and after a certain delay (1 hour thus 120 cycles) *Dictyostelium discoideum* will not move towards food but is sensitive to cAMP and moves into the direction of the highest cAMP concentration instead of searching for food. However, if food is meet by chance, it will be consumed.

Mound Phase

As mentioned before, hungry amoeba stream together, following the cAMP signal of other hungry individuals. Through the pressure from all sides to the center, the amoeba are lifted in the third dimension, which is also modeled here.

If the majority of surrounding hungry amoeba have been streaming to one place the so-called mound phase is reached. That means that a mound-like accumulation is built up. If enough amoeba are together or after a certain time, the next phase is started: the slug phase.

Slug Phase

In this phase, the gathered amoeba start to move and act as one multicellular organism. They start moving, mainly towards the next light source, in order to find the best place for the spore out. It is called slug phase, because the moving accumulation of amoeba looks like a microscopic slug (snail without house).

The collective movement is mainly steered by the cAMP concentration while the position of the individuals inside the slug is determined by the DIF (Differencing Inducing Factor (15)) concentration (this is important when different clones of *Dictyostelium discoideum* are studied), which also starts the building of the spore stalk, when a certain limit is reached.

Mexican Hat Phase

When a certain level of DIF concentration is reached the amoeba stop moving. The pressure of the outer amoeba lets the stalk grow into the height. The cells in the middle start building the stalk, harden and die. Other cells wander up the stalk and build the fruiting body. Only these cells reproduce by spore out. Spore out is like an explosion, the cells are distributed around the stalk nearly randomly, additionally moved by the wind.

VERIFICATION

In order to prove that our simulation model is accurate and represents the behavior of the real *Dictyostelium discoideum*, we conducted a thorough verification of the

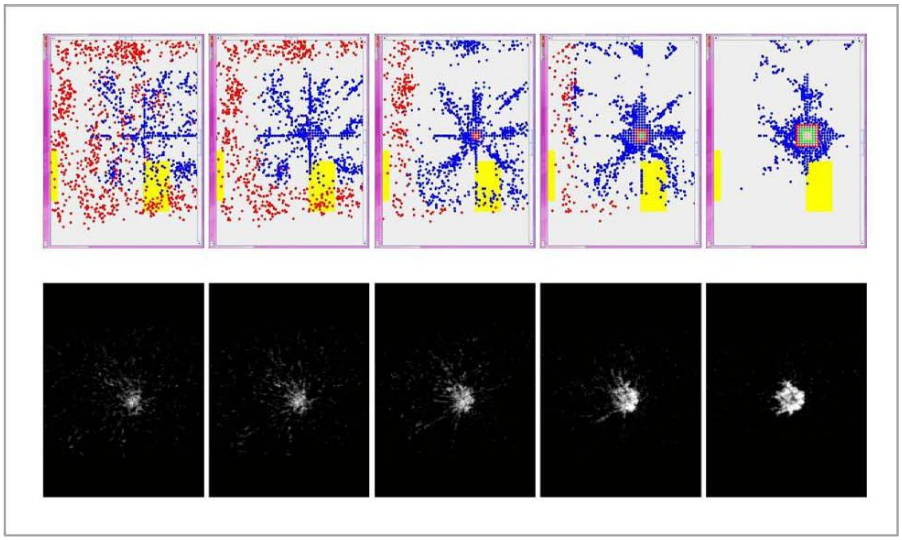


Figure 3: Verification of the aggregation phase

simulation model. Only then, the model can be used for further studies.

First we do optical verification, that is we compare by sight the behavior of the real and the artificial slime mold.

See Fig. 3 for a comparison of the aggregation phase of the simulation (upper pictures) and pictures of the real *Dictyostelium discoideum*. Note that also the timing of the set of pictures is comparable. One can see that the simulation shows the same behavior, that is the same aggregation patterns (streaming like) at the same points in time as the real movement and streaming patterns of *Dictyostelium discoideum*.

At the end of the streaming/aggregation phase *Dictyostelium discoideum* builds a mound (see Fig. 4). Afterwards it moves slug-like into direction of better environmental conditions, that is towards light (see Fig. 5). (In all pictures: (not) hungry cells are blue (red), the light source is yellow, the third dimension is also color coded in the center of the mound).

Quantitative Verification

After comparing the real and simulated behavior mainly by sight, we now do detailed measurements of the timing of the development of *Dictyostelium discoideum*.

The most imminent quantitative properties characterizing a population of *Dictyostelium discoideum* are the durations of the different phases. In the nature the aggregation takes 9 to 13 hours depending on the size of the population, environmental conditions and distance of individuals.

In the verification scenarios 2310 aggregation cycles have been simulated and the durations of the phases have been measured and the mean value has been calculated and verified by the Chi Square test.

It can be seen in table 1 that the simulation results are well within the natural durations, also regarding the confidence intervals.

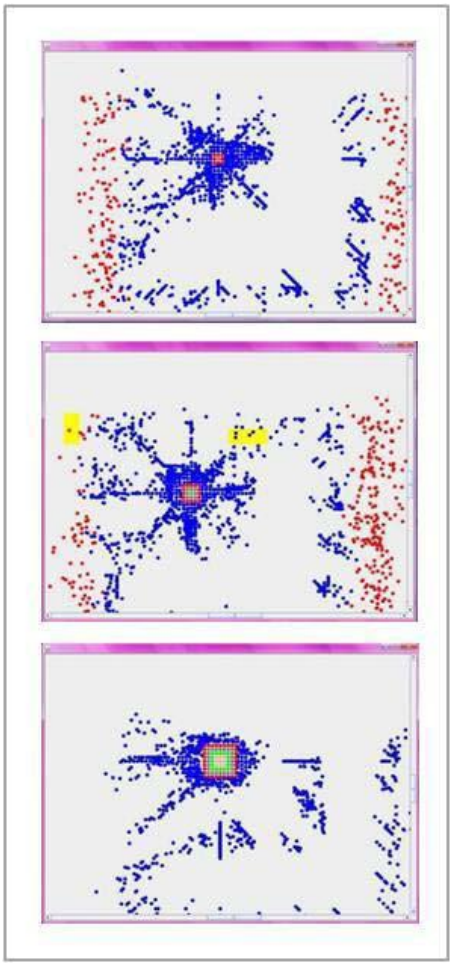


Figure 4: Simulation of forming the mound

Table 1: Verification of phases of the life-cycle

	Aggregation	Slug	Mexican Hat
Mean duration/min	696	174	238
Mean duration/h	11.6	2.9	3.96
95%-confidence interval (min)	694-699	207 - 213	237 - 240
Mean natural value from literature	9-13 hours	up to 5 hours	4-5 hours

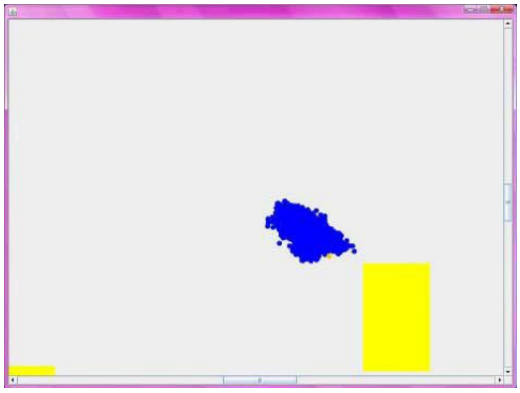


Figure 5: Simulation of slug movement towards light

The durations of the phases in a natural environment differ depending on the detailed environmental conditions. Considering the natural deviations of the timing of the phases, the simulated lengths of the phases is always within the natural bandwidth of the time values.

So as result we can state, that with high probability our simulation model captures the qualitative as well as the quantitative behavior of the whole life-cycle of *Dictyostelium discoideum*.

CONCLUSION

In this work we showed a simulation model for the whole life cycle of *Dictyostelium discoideum*. The verification shows that the results are well within the natural parameters. *Dictyostelium discoideum* has gained increasing interest as model organism for biologists as well as computer scientist in order to study the mechanisms of synergy and emergent intelligent behavior out of simple single parts. With our model we hope to better understand these mechanisms and derive new algorithms for the networked world that make things easier to inter-operate. In future work we will use the the model to study mechanisms of group selection and also try to derive computational algorithms from the behavior of *Dictyostelium discoideum*.

REFERENCES

[1] Konjin, Theo M. Effect of Bacteria on Chemotaxis in the Cellular Slime Molds. Journal of Bacteriology, vol. 99, August 1969: 503-509.

[2] Bonner, John T. Evidence for the formation of cell aggregates by chemotaxis in the development of the slime mold Dictyostelium discoideum. Journal of Experimental Zoology, 1947: 1-26.

[3] Konijn, Theo M., J. G.C. van De Meene, John T. Bonner, and David S. Barkley. The acrasin activity of adenosine-3',5'-cyclic phosphate. PNAS, 1967.

[4] Anderson, Alexander R. A., M. A. J Chaplain, and Katarzyna A. Rejniak. Single-cell-based models in biology and medicine. Birkhuser Verlag, 2007.

[5] Campbell, Neil A., und Jane B. Reece. Biologie, Pearson,6. edition 2005.

[6] http://dictybase.org/

[7] Palsson, Eirikur, and Hans G. Othmer. A model for individual and collective cell movement in Dictyostelium discoideum. PNAS, vol. 97, 12. September 2000.

[8] MacKay, Steven A. Computer Simulation of Aggregation in Dictyostelium discoideum. Cell Science 33, 1978.

[9] Foerster, P., S. C. Mller, and B. Hess. Curvature and spiral geometry in aggregation patterns of Dictyostelium Discoideum. Journal 'Development'. 109:11-16, 1990.

[10] Queller, David C., Kevin R. Foster, Angelo Fortunato, and Joan E. Strassmann. Cooperation and conflict in the social amoeba Dictyostelium Discoideum.

[11] Campbell, Neil A., and Jane B. Reece. Biologie, Pearson Studium, 2005.

[12] Strassmann, Joan E., and David C. Queller. Altruism among Amoebas. Natural History, September 2007: 24-29.

[13] Strassmann, Joan E., Yong Zhu, and David C. Queller. Altruism and social cheating in the social amoeba Dictyostelium Discoideum. Nature, vol 408, 21. December 2000: 965-967.

[14] Foster, Kevin R., Angelo Fortunato, Joan E. Strassmann, und David C. Queller. The costs and benefits of being a chimera. 2002.

[15] Strassmann, Joan E., und David C. Queller. Altruism among Amoebas. Natural History, September 2007: 24-29.

[16] Atsushi Teroa, Ryo Kobayashia and Toshiyuki Nakagakib. Physarum solver: A biologically inspired method of road-network navigation. Physica A: Statistical Mechanics and its Applications 2000: number 1, vol. 363, pp 115-119

[17] Wolfgang Marwan. Amoeba-Inspired Network Design Science 22 January 2010: Vol. 327. no. 5964, pp. 419 - 420

AUTHOR BIOGRAPHIES

MATTHIAS BECKER got his diploma in computer science from University Würzburg in 1996 and his Ph.D. from University Bremen in 2000. Since then he is researcher and lecturer at University Hannover in the fields of Petri Nets, simulation of discrete event systems, heuristic optimization and biological systems

Thanks to **Alexandra Nölle** for doing an excellent job of coding the simulation, finding the right parameters and doing the tedious work of running the large number of simulations for verification.

CONSTRUCTING CONTINUOUS-TIME CHAOS-GENERATING TEMPLATES USING POLYNOMIAL APPROXIMATION

Hidetaka Ito, Shinji Okamoto, Kosuke Fujimoto, and Akira Kumamoto
Department of Electrical and Electronic Engineering
Faculty of Engineering Science, Kansai University
3-3-35 Yamate, Suita, Osaka, 564-8680 Japan
Email: h.ito@kansai-u.ac.jp

KEYWORDS

Dynamical Systems, Intelligent Systems, Chaos, Vector Field, Polynomial Approximation

ABSTRACT

Aiming at developing a methodology for constructing continuous-time chaotic dynamical systems as flexible pattern generators, this paper discusses a strategy for binding desired unstable periodic orbits into a chaotic attractor. The strategy is comprised of the following two stages: constructing an interim "chaos-generating template", and deforming the template according to the specifically desired orbits.

INTRODUCTION

Chaos is a fascinating phenomenon arising from nonlinearities in dynamical systems, and investigations on its applications to intelligent and flexible systems have appeared in many fields. An important aspect of chaos is that chaotic attractors embed or "host" an infinite number of unstable periodic orbits (UPO's) bifurcated from prechaotic states (Ott, 2002). Among them, some distinctive orbits can be used for characterization or control purposes. For example, a variety of chaos control methods (Ott et al., 1990; Pyragas, 1992; Zhang et al., 2009) can stabilize UPO's embedded in chaotic attractors, enlarging the operation range and/or enhancing the functionality of the system.

In more recent years, the synthesis of chaos from various approaches (Chen and Ueta, 2002; Zelinka et al., 2008; Muñoz-Pacheco and Tlelo-Cuautle, 2009) has been an active direction of research along the line of exploiting chaos. Primary concerns of these efforts include statistical and topological characteristics (e.g., invariant measure, Lyapunov spectrum, novel scrolling behaviors) that would be important in designing chaos-based information processing and communication applications.

While the present study share some common motivation with the above-mentioned studies, we have put more focus on the geometrical shape and dynamical properties of UPO's themselves from the viewpoint of the adaptive generation of periodic behaviors. Here our intention lies in extending the functionality of (stable) periodic pattern generators based on function approximation of vector fields, e.g., polynomial approximation (Okada and Nakamura, 2002) and neural network learning (Kuroe and Miura, 2006).

In this paper, we consider continuous-time chaotic attractors as a container of UPO's (patterns) where they can be stabilized, entrained, or targeted by external inputs into the dynamical system. In what follows, we propose a design strategy for binding desired UPO's into a chaotic attractor governed by a polynomial vector field. The strategy is comprised of the following two stages: constructing an interim "chaos-generating template", and deforming the template according to the specifically desired orbits.

POLYNOMIAL VECTOR FIELDS

We consider polynomial dynamical systems of the form $\dot{x} = f(x)(x \in R^N)$ where the vector field $f(x)$ is represented as

$$f(x) = [f_1(x) \cdots f_N(x)]^T = \Phi(a_{(p_1 p_2 \cdots p_N)})\theta(x), \quad (1)$$

$$\theta(x) = [x_1^\ell \ \cdots \ x_N^\ell \ \ x_1^{\ell-1}x_2 \ \cdots \ 1]^T. \quad (2)$$

Here, ℓ denotes the maximum degree of the polynomials, and the matrix Φ is comprised of the coefficients $a_{(p_1 p_2 \cdots p_N)}$ of the polynomials. For example, with 3rd-order polynomials, the first element of a two-dimensional vector field is represented as

$$\begin{aligned} f_1(x) \ = \ & a_{1(30)}x_1^3 + a_{1(21)}x_1^2x_2 + a_{1(12)}x_1x_2^2 \\ & + \ a_{1(03)}x_2^3 + a_{1(20)}x_1^2 + a_{1(11)}x_1x_2 \\ & + \ a_{1(02)}x_2^2 + a_{1(10)}x_1 + a_{1(01)}x_2 + a_{1(00)}. \end{aligned} \quad (3)$$

In our design process of the dynamical system, the coefficients are obtained by least-square fitting. To this end, we set up target vectors $f(\eta_i)$ $(i = 1, 2, \cdots, L)$ at design points η_i on and in the vicinity of the target orbits, and construct matrices

$$F \ = \ [f(\eta_1) \ \ f(\eta_2) \ \cdots \ f(\eta_L)], \quad (4)$$

$$\Theta \ = \ [\theta(\eta_1) \ \ \theta(\eta_2) \ \cdots \ \theta(\eta_L)]. \quad (5)$$

With these matrices, the least-square solution for Φ is given by

$$\Phi(a_{(p_1 \ p_2 \ \cdots \ p_N)}) = F\Theta^{\#} \quad (6)$$

where $\Theta^{\#}$ is the Moore-Penrose pseudo inverse of Θ.

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

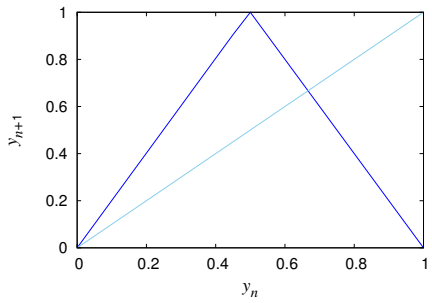


Figure 1: The target one-dimensional map.

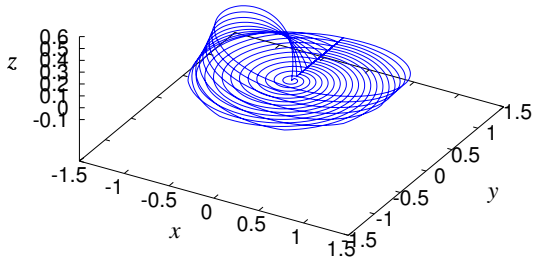


Figure 2: The target continuous-time chaotic attractor corresponding to the map in Fig.1.

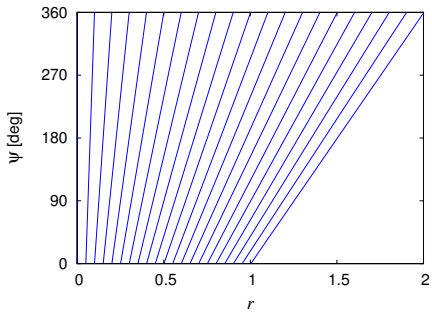


Figure 3: The stretching characteristics of the target attractor.

Design of stable periodic orbits using the above formulation and its applications to robotics can be found in (Okada and Nakamura, 2002).

CHAOS-GENERATING TEMPLATES

When we consider accommodating multiple UPO's that largely overlap in the state space of a single dynamical system, haphazard approaches to placing unstable and stable manifolds can easily fail. For example, unintended stable periodic orbits may emerge from the conflict of desired instabilities, which leads to the loss of transitivity among the orbits. Therefore we here consider binding UPO's according to a typical chaos-generating mechanism of stretching and folding.

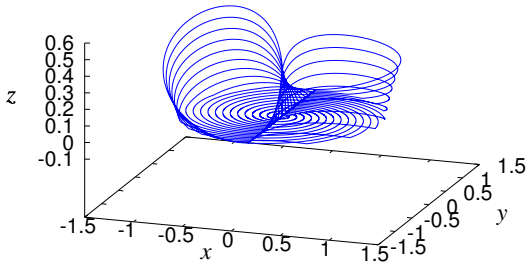


Figure 4: The target chaotic attractor embedding three period-1 UPO's.

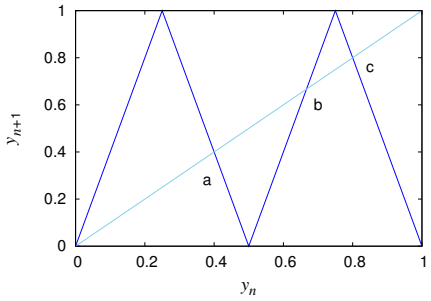


Figure 5: The target one-dimensional map corresponding to Fig. 4.

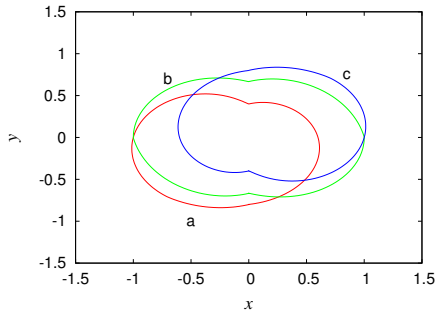


Figure 6: Embedded three period-1 UPO's.

First, as a simplest example, we consider the one-dimensional map (Fig. 1) for the y-coordinate of the n-th crossing on the Poincaré section $\Sigma = \{x, y, z | (x = 0, y \geq 0)\}$. While we have a wide freedom of choice of continuous-time trajectories leading to this map, we here adopt the systematically designed (explained below) bundle of orbits shown in Fig. 2 as a target chaotic attractor for the polynomial approximation. This target attractor embeds a single period-1 UPO corresponding to the fixed point of the map.

The procedure for constructing the target attractor in Fig. 2 is as follows: We first draw the set of orbits on the polar-coordinate $r\psi$ plane as shown in Fig. 3. In this

figure, the individual trajectories starting from $(r, \psi) = (r_0, 0^\circ)$ is given by

$$r = r_0 \left(1 + \frac{\psi}{360^\circ}\right) \tag{7}$$

$(z = 0)$, and the stretching between neighboring trajectories takes place at the rate of twice per rotation. Next, in the region where r becomes greater than 1.0, we reassign the values of r and z as

$$\begin{cases} r = 1 + \left[r_0 \left(1 + \dfrac{\psi}{360^\circ}\right) - 1.0)\right] \cos\dfrac{\psi}{2}, \\ z = \left[r_0 \left(1 + \dfrac{\psi}{360^\circ}\right) - 1.0)\right] \sin\dfrac{\psi}{2}, \end{cases} \tag{8}$$

giving the characteristic of one folding per rotation. Finally, transforming into the Cartesian coordinates, we obtain the target chaotic attractor shown in Fig. 2. The obtained target attractor, precisely a target invariant set since we have not specified the attracting properties, is more gently constructed than many known attractors in the sense that stretching and folding do not take place in a particular region of the state space.

The above strategy for embedding a single period-1 UPO can be extended to embed, or bind, several period-1 UPO's. For example, if we reorganize, or reduce, two rotations along the above target attractor into one rotation, we obtain the target chaotic attractor shown in Fig. 4. From the corresponding one-dimensional map in Fig. 5, we see that the attractor now embeds three period-1 UPO's whose continuous-time trajectories are shown in Fig. 6. Now we will use this attractor as a template for designing a set of three coexisting period-1 UPO's.

ASSIGNING ATTRACTING PROPERTIES

While we intend to use UPO's instead of stable periodic orbits to enhance functionality of the embedded orbits, it is desirable that the chaotic invariant set that binds the UPO's should be an attractor rather than a chaotic saddle. Thus we assign attracting properties by setting appropriate velocity vectors in the vicinity of the target attractor.

This can be done in various ways, and we here propose a method that we consider to be relatively easy to control. In this method, we first place two planar sets of design points at $z = \pm d$ in the $r\psi z$ space (see Fig. 7). At these design points, we set target vectors whose horizontal components reproduce the target vectors on the $z = 0$ plane, and vertical components define the degree of transverse stability of the target attractor (see Fig. 8). These three layers are then folded according to the procedure described in the preceding Section.

POLYNOMIAL REALIZATION OF THE VECTOR FIELDS ALONG THE TEMPLATES

Now we apply polynomial approximation to obtain a functional form for the vector field along the target chaotic attractor (chaos-generating template). Figure 9

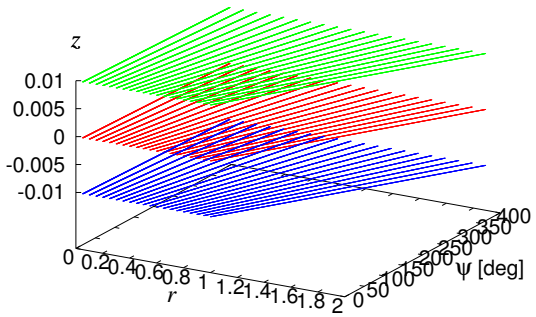


Figure 7: The horizontal components of target vectors on two planes $z = \pm d$ reproducing the target vectors on the $z = 0$ plane.

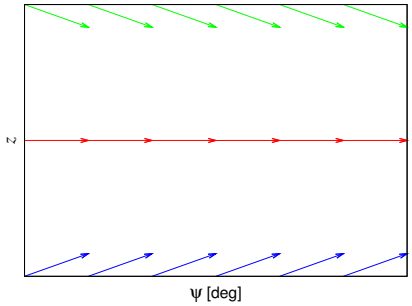


Figure 8: The vertical components of the target vectors on two planes $z = \pm d$ defining the degree of transverse stability of the target attractor.

shows the obtained chaotic attractor of the polynomial $(\ell = 8)$ system constructed from the target attractor in Fig. 4. Also shown in Fig. 10 is the corresponding (underlying) return map that represents y_{n+1} versus y_n where each mapping starts from $(0, y_n, 0)$ and ends at $(0, y_{n+1}, z_{n+1})$. The obtained map implies the existence of the intended three period-1 UPO's, which are confirmed by a numerical search and shown in Fig. 11.

DEFORMING THE TEMPLATES

In order to utilize the UPO's for specific applications, we need to deform the UPO's embedded in the templates according to the desired dynamical patterns. This deformation can be performed either before or after the polynomial approximation depending on the overall implementation. In both cases, the deformation needs to be topology-preserving to guarantee the existence of an inverse deformation that is necessary for maintaining the chaotic dynamics and the feedback path.

As an example, we consider deforming the xy component of the UPO b in Fig. 11 into a circle $x^2 + y^2 = (0.8)^2$. To this end, we construct an appropriate transformation $A(r, \psi)$ that is piecewise linear for r, and apply this transformation to the whole target chaotic attractor. Figure 12 shows the deformed target attractor, and Fig. 13 the chaotic attractor of the resulting polynomial

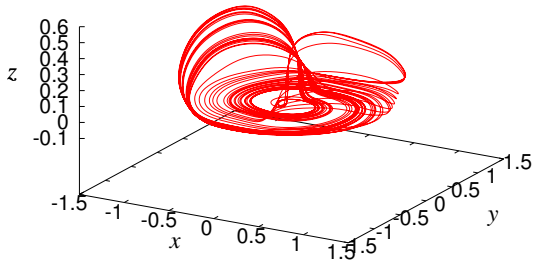


Figure 9: The chaotic attractor of the constructed system.

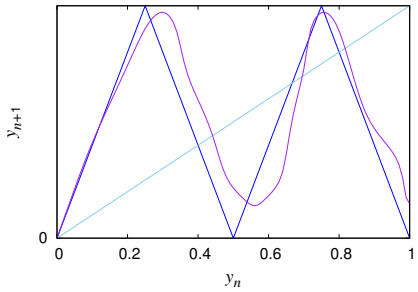


Figure 10: The return map corresponding to the chaotic attractor in Fig. 9.

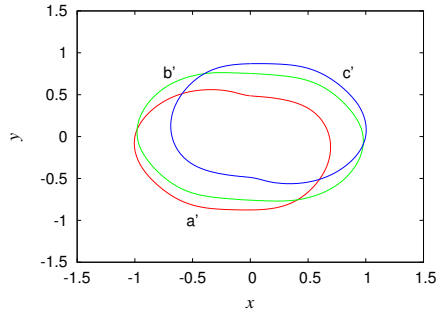


Figure 11: Three period-1 UPO's emdedded in the constructed attractor.

($\ell = 8$) system. The UPO's in the template (the target attractor before deformation), in the deformed template, and in the constructed attractor are compared in Fig. 14, showing that the obtained UPO (blue) is in good agreement with the desired one (red).

CONCLUDING REMARKS

We have successfully embedded three period-1 orbits while keeping them unstable but confined in an attractor. While the instability of the orbits may permit better operational flexibility, the attractor provides some degree of robustness against fluctuations. In addition, the current approach can also be regarded as an attempt to

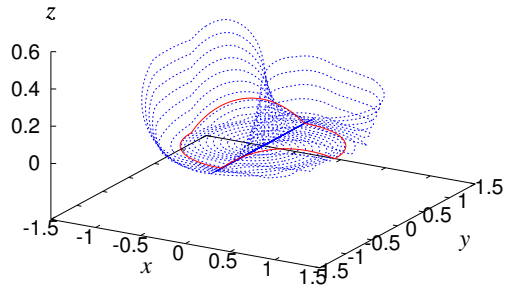


Figure 12: The deformed target attractor.

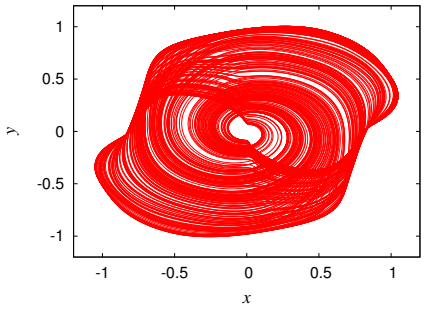


Figure 13: The chaotic attractor of the constructed system for the target in Fig.12.

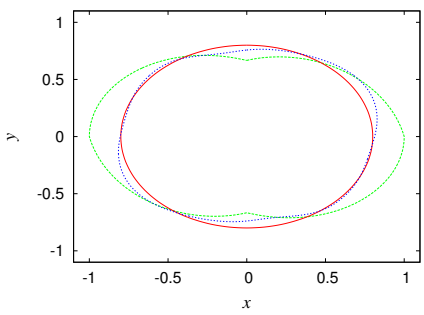


Figure 14: Comparison of UPO's in the template (green), in the deformed template (red), and in the constructed attractor (blue).

bridge the topological classification of chaotic attractors (Tsankov and Gilmore, 2004) with the design of dynamical systems, and therefore, consideration on the choice of appropriate templates will be important for enhancing design flexibility.

REFERENCES

Chen, G. and T. Ueta (eds.). 2002. *Chaos in Circuits and Systems*. World Scientific.

Kuroe, Y. and K. Miura. 2006. "Generation of oscillatory trajectories with specified stability degree using recurrent neu-

ral networks". In *Proceedings of the 2006 International Joint Conference on Neural Networks* (Vancouver, BC, Canada), 3478-3485.

Muñoz-Pacheco, J.M. and E. Tlelo-Cuautle. 2009. "Automatic synthesis of 2D-n-scrolls chaotic systems by behavioral modeling". *Journal of Applied Research and Technology*, 7(1), 5-14.

Okada, M. and Y. Nakamura. 2002. "Polynomial design of dynamics-based information processing system". In *Proceedings of the Second Joint CSS/RAS International Workshop on Control Problems and Automation* (Las Vegas, Nevada), 91-104.

Ott, E. 2002. *Chaos in Dynamical Systems, 2nd ed*. Cambridge University Press.

Ott, E.; C. Grebogi; and J.A. Yorke. 1990. "Controlling chaos". *Physical Review Letters*, 64(11), 1196-1199.

Pyragas, K. 1992. "Continuous control of chaos by self-controlling feedback". *Physics Letters A*, 170(6), 421-428.

Tsankov, T.D. and R. Gilmore. 2004. "Topological aspects of the structure of chaotic attractors in R^3". *Physical Review E*, 69, 056206.

Zelinka, I.; G. Chen; and S. Celikovsky. 2008. "Chaos synthesis by means of evolutionary algorithms". *International Journal of Bifurcation and Chaos*, 18(4), 911-942.

Zhang, H.; D. Liu; and Z. Wang. 2009. *Controlling Chaos*. Springer.

AUTHOR BIOGRAPHIES

HIDETAKA ITO received the B.E., M.E., and Ph.D. degrees in electrical engineering from Kyoto University in 1988, 1991, and 1996, respectively, and the MSc degree in satellite engineering from the University of Surrey in 1991. Since 1994, he has been with Kansai University, where he is currently an associate professor. His research interests include nonlinear dynamics, scientific computing, and intelligent computing.

SHINJI OKAMOTO received the B.E. degree in electrical engineering and computer science from Kansai University in 2009, and is currently a graduate student working on his M.E. degree. His research interests include dynamical system design and scientific computing.

KOSUKE FUJIMOTO received the B.E. and M.E. degrees in electrical engineering and computer science from Kansai University in 2008 and 2010, respectively. His research interests include dynamical system design and scientific computing.

AKIRA KUMAMOTO received the B.E., M.E., and D.E. degrees in electrical engineering from Kyoto University in 1968, 1971, and 1983, respectively. Since 1971, he has been with Kansai University, where he is currently a professor. His research interests include intelligent computing and software, network-based information systems, and human information processing.

ON RELIABILITY OF SIMULATIONS OF COMPLEX CO-EVOLUTIONARY PROCESSES

Peter Tiňo
School of Computer Science
The University of Birmingham
Edgbaston, Birmingham B15 2TT, UK
Email: P.Tino@cs.bham.ac.uk

Siang Yew Chong
School of Computer Science
The University of Nottingham, Malaysia Campus
Jalan Broga, Semenyih 43500, Malaysia
Email: Siang-Yew.Chong@nottingham.edu.my

Xin Yao
School of Computer Science
The University of Birmingham
Edgbaston, Birmingham B15 2TT, UK
Email: X.Yao@cs.bham.ac.uk

KEYWORDS

Evolutionary game theory, evolutionary computation, co-evolution, shadowing lemma.

ABSTRACT

Infinite population models of co-evolutionary dynamics are useful mathematical constructs hinting at the possibility of a wide variety of possible dynamical regimes - from simple attractive fixed point behavior, periodic orbits to complex chaotic dynamics. We propose to use the framework of shadowing lemma to link such mathematical constructs to large finite population computer simulations. We also investigate whether the imposition of finite precision computer arithmetic or the requirement that population ratios be rational numbers does not leave the infinite population constructs and theories irrelevant. We argue that if the co-evolutionary system possesses the shadowing property the infinite population constructs can still be relevant. We study two examples of hawk-dove game with Boltzmann and (μ, λ) selection. Whereas for Boltzmann selection there is a strong indication of the shadowing property, there is no shadowing in the case of (μ, λ) selection.

INTRODUCTION

Games have been used to model some important complex real-world problems in politics, economics, biology (Axelrod, 1984) and engineering (Nisan et al., 2007). They capture intrinsic properties of such problems through the specification of rules that constrain strategies to certain behaviors (legal moves), goals for strategies to meet (to win the game), and rewards under finite resources (payoffs). However, games that successfully abstract real-world problems are often analytically intractable. In such cases, alternative heuristic approaches, such as evolutionary algorithms (EAs), are used to solve these problems in practice.

Games are particularly important in the development of a class of EAs known as co-evolutionary algorithms (CEAs) (Chellapilla and Fogel, 1999). Unlike classical EAs that require an *absolute* quality measurement of solutions to guide the population-based, stochastic search process, in CEAs, the solution quality can be estimated only with respect to its performance against a (usually) small sample of test cases. In cases where an absolute quality measurement is not available, CEAs can still solve the problem by making use of some form of *strategic* interactions between competing solutions in the population to guide the search (Chong et al., 2008).

Despite early success of CEAs in solving games, there are well-documented failures leading to poor performance of CEAs under certain conditions. One example is the *overspecialization* of evolved strategies that performs well only against specific opponents rather than a large number of different opponents (Darwen and Yao, 2002; Chong et al., 2008). As such, there are strong motivations for more in-depth theoretical studies on CEAs to understand precisely how and why CEAs can solve the problem of games.

One theoretical framework that is naturally suited for analysis of CEAs is evolutionary game theory (EGT) (Smith, 1982). The classical game theory setting involves a rational individual (player) that has to choose between different strategies, one that maximizes its payoff when interacting against another player in a game (which in turn, also maximizes its own payoff). In contrast, the EGT setting involves an infinitely large population of players that are allowed to use a set of predefined strategies. These strategies are *inheritable* and all players compete for payoffs that decide their average reproductive success (Sigmund and Nowak, 1999). Different constructions (e.g., different games, different selection mechanisms etc.) will lead to different frequency-dependent population dynamics (Hofbauer and Sigmund, 2003a). As such, EGT provides the means with which one can study precisely the conditions that affect the outcome or success of some strategies over others in the population under evolutionary process.

However, there are few studies that apply EGT in the

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

analysis of CEAs. For example, a simple EGT setting of the hawk-dove game that involves interactions between two strategies has been used to investigate the evolutionary process of CEAs under various conditions: the study in (Fogel et al., 1997) investigated the impact of finite population, while the study in (Ficici et al., 2005) investigated the impact of selection mechanisms.

While evolutionary dynamics typically involves rather simple dynamical scenarios, co-evolution can lead to a rich variety of dynamical behaviors, including chaos (see, e.g., (Ficici et al., 2005)). Such complex dynamics are usually (and conveniently) explored under the assumption of infinite populations, where the entities of interest are, e.g., the ratios of individuals adopting a particular strategy. Given that for some game settings, computer generated orbits of co-evolutionary dynamics, under the assumption of infinite populations, reveal intricate dynamical patterns (including chaotic dynamics), we ask two important questions regarding the information content in such simulations:

1. How informative are the observed chaotic trajectories, given that the computer precision is limited? In chaotic dynamics, nearby trajectories get locally exponentially separated, so round-off errors will inevitably lead to trajectories very different from the 'ideal' infinite population ones the equations are supposed to describe.

2. Often, we are ultimately interested in dynamical behaviors of potentially large, but finite populations, using infinite population formulations as convenient conceptual constructs. How informative are then the complex infinite population trajectories about the dynamics of large, but finite populations modeled on the computer?

We propose to address these questions in the context of shadowing lemma (see e.g. (Katok and Hasselblatt, 1995)). For simplicity of presentation, we only will consider co-evolution of games with two pure strategies. We consider in more detail some settings of the co-evolutionary dynamics of the two-strategy hawk-dove game with (μ, λ) and Boltzmann selection mechanisms. We will show that there is indeed a possibility of quite intricate infinite population co-evolutionary dynamics. We will also study whether such complex infinite population dynamical patterns can have direct relevance for large scale finite population computer simulations.

EVOLUTIONARY GAME THEORY

We first describe the standard EGT framework that makes the following assumptions:

1. An infinitely large population of players, each of which has a finite set of *pure strategies* to choose from.

2. *Complete mixing* - every player interacts with all players in the population. Each player accumulates payoff depending on the outcome of the games.

3. Players reproduce in proportion to their cumulative payoffs. Reproduction is asexual and without variation, i.e., players generate clones as their offspring.

In this paper, we consider a simple EGT setting similar to that of (Ficici et al., 2005) with a game involving two players (for illustration). Each player has a finite set of pure strategies to choose from, i.e., S_i for the first player and S_j for the second player. The game is *symmetric*, i.e., both players have the same set of strategies to choose from ($S_i = S_j$). The payoff (game outcome) for the first player is g_{ij} when the player chooses strategy $i \in S_i$ while the opponent chooses $j \in S_j$ (the payoff for the second player is g_{ji}).

For the case where there are only two *pure strategies* (i.e., $\{X, Y\} \in S$), the payoff matrix for the first player (row) in a game against the second player (column) can be constructed as follows:

	X	Y
X	a	b
Y	c	d

(1)

where each entry gives the respective payoff for the chosen pair of strategies. For example, the first player receives the payoff b when it chooses strategy X while its opponent chooses Y.

Each player in the population chooses only one of the two pure strategies. Let p be the proportion of players in the population choosing X. $1 - p$ is the proportion of players in the population choosing the other pure strategy Y. We can compute the cumulative payoffs for both strategies, w_X and w_Y, in the form of a pair of linear equations:

$$\begin{aligned} w_X &= ap + b(1-p) \\ w_Y &= cp + d(1-p). \end{aligned} \quad (2)$$

We consider games with the payoff structure satisfying $a < c$ and $b > d$. For any such game, there is one population state known as *polymorphic equilibrium*

$$p_{EQ} = \frac{d-b}{a-c+d-b} \quad (3)$$

in which the cumulative scores for both strategies are the same ($w_X = w_Y$) (Ficici et al., 2005). Interpreting the population as a mixture of strategies s, (i.e., uses pure strategies X and Y with probability $p_X = p$ and $p_Y = 1 - p$, respectively), the state p_{EQ} is a *Nash equilibrium*, whereby the mixture of strategies s is its own *best reply* (i.e., if a player uses s, the highest payoff that the opponent can obtain is when it also uses s). When both players use s, they are in Nash equilibrium since neither has the incentive to deviate unilaterally to use other strategies.

Hawk-Dove Game

Although there are many games that satisfy the constraints of $a < c$ and $b > d$, in this paper we consider the classical game setting of the hawk-dove game. The setting involves interactions of two distinct behaviors (pure strategies), *hawk* and *dove*, competing for gains (G) upon winning under the costs (C) of injury.

Hawks are aggressive and two hawks will fight until one retreats with an injury. Interactions between hawks would lead to the payoff expectation of $(G-C)/2$ given a probability of $1/2$ for injury. Doves, in contrast, will avoid a fight and perform threatening postures until both retreats without injury. In such a case, they share the gain with a payoff of $G/2$. Any interaction between a hawk and a dove will lead to the dove retreating immediately. The hawk will take the full gain (payoff of G) while the dove has zero gain (payoff is 0), with no cost on injury incurred to both parties.

The payoffs for hawk-dove interactions are summarized in the following payoff matrix:

	Hawk	Dove
Hawk	$(G-C)/2$	G
Dove	0	$G/2$

(4)

When the cost of injury is greater than the gain in winning the game (i.e., $G < C$), the hawk-dove game satisfy the constraints of $a < c$ and $b > d$. The polymorphic equilibrium can be easily obtained from Equation 3: $p_{EQ} = G/C$. Note that adding a constant value w to each entry in the payoff matrix (4) ensures that the cumulative payoffs for strategies are non-negative without changing the nature of the population dynamics (Ficici et al., 2005).

Replication and Selection Pressure

In the classical EGT, the population dynamics are described by the *replicator equation* that governs how the *frequency* (proportion) of strategies in the population changes with time t of the evolutionary process. For the case of the game specified by the payoff matrix (1), the replicator equation under a selection mechanism based on the proportion of cumulative payoffs is given by:

$$f(p) = \frac{pw_X}{pw_X + (1-p)w_Y}$$

where $f(p)$ is the frequency of strategy X in the population in the following generation $(t+1)$, given that its frequency in the current population is p.

In addition to the classical fitness-proportional selection, there is a variety of alternative selection mechanisms. In this paper, we consider two well known selection mechanisms: (μ, λ) and Boltzmann selection.

The (μ, λ) **selection** is usually associated with the selection operator that is used in a class of EAs known as *evolution strategies*. Each of μ parents generates k offspring, which results in $\lambda = k\mu$ offspring. In the context of an EGT setting of CEAs operating with infinite population, we are only concerned with the ratio $\gamma = \mu/\lambda$. The replicator equation can then be obtained as (Ficici et al., 2005):

$$f(p) = \begin{cases} 1 & \text{if} \quad p < p_{EQ} \text{ and } p \geq \gamma, \\ p/\gamma & \text{if} \quad p < p_{EQ} \text{ and } p < \gamma, \\ 1 + (p-1)/\gamma & \text{if} \quad p > p_{EQ} \text{ and } p > 1-\gamma, \\ 0 & \text{if} \quad p > p_{EQ} \text{ and } p \leq 1-\gamma, \\ p_{EQ} & \text{if} \quad p = p_{EQ}. \end{cases} \tag{5}$$

In **Boltzmann selection** the fitness that determines the selection of an individual is scaled as $w_{scaled} = e^{\beta w}$, where w is the original fitness. The *inverse temperature* parameter $\beta = 1/T$ determines the selection pressure. Selection of an individual is proportional to the scaled fitness w_{scaled}. For a CEA with an infinite population, the replicator equation with Boltzmann selection is given by (Hofbauer and Sigmund, 2003b; Ficici et al., 2005):

$$f(p) = \frac{pe^{\beta(pa-pb+b)}}{pe^{\beta(pa-pb+d)} - pe^{\beta(pc-pd+d)} + e^{\beta(pc-pd+d)}}. \tag{6}$$

THE SHADOWING PROPERTY

Given a game and a particular selection mechanism (that together imply the corresponding map f) and an initial condition p_0, the map f generates an orbit $p_n = f(p_{n-1})$, $n = 1, 2, ...$ If instead of the true iterands p_n we observed p_n corrupted by a bounded noise (for example due to rounding errors or finite population effects), but still used the dynamics f, we would obtain a *pseudo-trajectory* $\{\tilde{p}_n\}_{n \geq o}$,

$$|\tilde{p}_0 - p_0| < \delta, \;\; |f(\tilde{p}_{n-1}) - \tilde{p}_n| < \delta, \;\; n \geq 1,$$

where $\delta > 0$ is the range size of the bounded noise. Such a pseudo-trajectory is often referred to a δ-pseudo-trajectory.

Given an $\epsilon > 0$, we say that a trajectory $\{q_n\}_{n \geq 0}$ ϵ-shadows another trajectory $\{g_n\}_{n \geq 0}$, if $\{q_n\}_{n \geq 0}$ stays within the ϵ-tube around $\{g_n\}_{n \geq 0}$:

$$|g_n - q_n| < \epsilon, \quad n \geq 0.$$

The Shadowing lemma tells us that (remarkably) even for the most complex and locally exploding chaotic maps, under some circumstances, the corrupted pseudo-trajectories are informative: For *any* $\epsilon > 0$, there exists a $\delta > 0$, such that for *every* δ-pseudo-trajectory $\{\tilde{p}_n\}_{n \geq 0}$ there is a true (uncorrupted) trajectory $\{q_n\}_{n \geq 0}$ under f that ϵ-shadows the pseudo-trajectory $\{\tilde{p}_n\}_{n \geq 0}$:

$$|\tilde{p}_n - q_n| < \epsilon, \;\; q_{n+1} = f(q_n), \;\; n \geq 0.$$

Hence, even though one would be tempted to assume that, under chaos, trajectories that are disrupted at every point by a bounded noise cannot possibly represent anything real, in fact such trajectories can be closely shadowed by a true trajectory.

What are the conditions that guarantee the shadowing property? Virtually all studies of the shadowing property have been performed in the framework of continuous and smooth dynamical systems. If a system is (uniformly) hyperbolic on an invariant set Ω, i.e., (loosely speaking) at each point $p \in \Omega$ the (linearized) system has only local contracting and expanding subspaces that get consistently translated by f into the local contracting and expanding subspaces at $f(p)$, then the system will have the shadowing property (Hayesand and Jackson, 2005). In general, establishing that a dynamical system is hyperbolic can be rather difficult, but it can be shown that a smooth one-dimensional system acting on Ω is hyperbolic iff for all $p \in \Omega$, there exists an $n(p) > 0$, such that the derivative $(f^{n(p)})'(p)$ of $f^{n(p)}$ at p is (in absolute value) larger than 1. Setting $p_1 = p$ and $p_n = f(p_{n-1})$, $n = 2, 3, .., n(p) - 1$, (by chain rule) this translates into the requirement that

$$|\prod_{n=1}^{n(p)} f'(p_n)| > 1.$$

In what follows we concentrate on co-evolutionary dynamics of the hawk-dove game with Boltzmann and (μ, λ) selections. Through representative case studies, we first show that such infinite population systems are indeed inherently capable of generating very complex dynamical scenarios. We then discuss whether studying infinite population models of such co-evolutionary systems can be informative about the dynamics one would observe in large population size simulations.

In this contribution we use the framework of shadowing lemma for two purposes:

1. *To investigate whether the complex simulated co-evolutionary trajectories under the infinite population assumption represent anything real.* The computer arithmetic operates with finite precision and can only yield pseudo-trajectories that cannot be *a-priori* guaranteed to represent any true trajectory of the given complex system. If the co-evolutionary system has the shadowing property then one can be assured that the observed pseudo-trajectories are shadowed by true ones generated by the underlying system.

2. *To investigate whether the complex dynamical co-evolutionary patterns under infinite populations indicate complex dynamics in large finite population computer simulations.* For large population size, the effects of finite population size on the strategy ratios p can be considered bounded noise. Furthermore, the larger the population, the smaller the range size of the noise. In a system with shadowing property, finite population pseudo-trajectories are shadowed by the complex trajectories from the original infinite population model.

Shadowing under Boltzmann selection

As a case study, we use the following setting of the hawk-dove game and Boltzmann selection that can lead to complex chaotic population dynamics: $G = 7$, $C = 12$, $\beta > 3.8$ (Ficici et al., 2005).

To show that for a range of selection pressures $\beta = 1/T$, the hawk-dove game with Boltzmann selection leads to complex dynamics, we estimated the topological entropy h_0 of the system (Balmforth et al., 1994). Topological entropy quantifies the exponential growth rate in the number of periodic points of the system. In particular, for each continuous, piece-wise monotone map f (such as the replicator map under Boltzmann selection (6)), it holds $h_0(f) \leq \kappa(f)$, where $h_0(f)$ is the topological entropy of f and $\kappa(f)$ is the exponential growth rate of the number of periodic orbits of f (Katok and Mezhirov, 1998). Hence, for complex systems with positive topological entropy (calculated with log base 2), the number of periodic points of period m grows as $\approx 2^{h_0 m}$. For selection pressures given by $\beta = 3.8$, 5.0, 5.7 and 6.5 the estimated topological entropies were $h_0 = 0.281$, 0.842, 0.905 and 0.949, respectively, indicating intricate dynamical organization of the infinite population system at higher selection pressures. But are the computer generated trajectories informative about the true co-evolutionary dynamics of this game under the Boltzmann selection? Furthermore, do such trajectories tell us anything about the dynamical complexity one might expect when running simulations with large finite populations? To address these questions we turn to the shadowing lemma.

The system (6) is a smooth dynamical system over the unit interval $[0, 1]$. For higher selection pressures the map f has two critical points where the derivative vanishes:

$$c_i = \frac{1}{2} + (-1)^i \sqrt{\frac{1}{4} - \frac{2}{\beta C}}, \;\; i = 1, 2.$$

Construct the set of all pre-images of the critical points under f;

$$\mathcal{E} = \{p \in [0, 1] \mid f^n(p) \in \{c_1, c_2\} \text{ for some } n \geq 0\}.$$

If the f-invariant set $\Omega = (0, 1) \setminus \mathcal{E}$ is a hyperbolic set under f, then the system (6) possesses the shadowing property on Ω. Showing analytically that Ω is hyperbolic is very difficult due to the (rather involved) nonlinearities in f. Nevertheless, we can obtain at least an indication of hyperbolicity of Ω by numerically checking whether for each p from a fine grid of points, there exists a natural number $n(p)$ such that $|(f^{n(p)})'(p)| > 1$. In Figure 1 we show the empirically determined fold numbers $n(p)$ for $G = 7$, $C = 12$ and $\beta = 5.0$. The figure indicates that f is hyperbolic (and so has the shadowing property) on Ω. Of course, such indication should be taken with a pinch of salt due to possible numerical inaccuracies in the neighborhood of critical points.

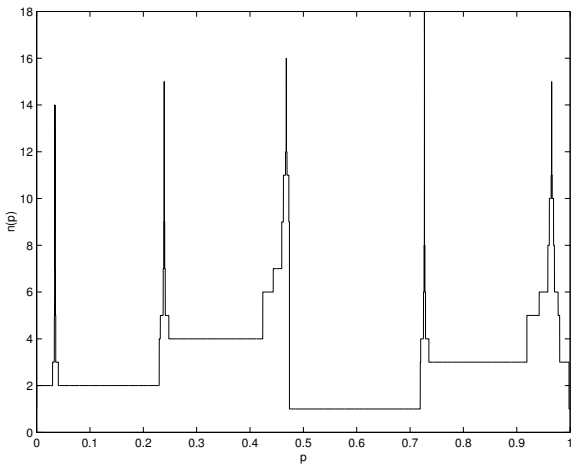


Figure 1: Hawk-dove game with $G = 7$, $C = 12$ and selection pressure $\beta = 5.0$. Shown is the fold number $n(p)$, such that $|(f^{n(p)})'(p)| > 1$, as a function of p.

Shadowing under (μ, λ) selection

Consider the hawk-dove game with the cost of injury twice the gain in winning, $C = 2G$, and a (μ, λ) selection with $\lambda = 2\mu$ (each of μ parents generates two offspring). The replicator equation (5) reads:

$$f(p) = \begin{cases} 2p & \text{if} \quad p \in [0, 1/2) \\ 1 + 2(p-1) & \text{if} \quad p \in (1/2, 1] \\ 1/2 & \text{if} \quad p = p_{EQ} = 1/2. \end{cases} \tag{7}$$

The map f acts as left-shift on binary representations of $p \in [0, 1/2) \cup (1/2, 1)$. As such, it can generate a wide variety of dynamical behaviors, dictated by the distribution of digits in the binary expansion $[p]_2$ of the initial condition p. The possible dynamics include periodic orbits of arbitrary periods, a-periodic and 'chaotic' orbits (arising from irrational initial conditions).

The map (7) is discontinuous. There is virtually no literature on shadowing property in discontinuous systems. Some discontinuous systems have the shadowing property, while others do not. It turns out that the difference between the well studied map $r(p) = 2p \bmod 1$ in chaotic dynamics and the map $f(p)$ in (7) - the existence of the additional equilibrium $p_{EQ} = 1/2$ for f - is crucial. Since $r(p)$ is a smooth expanding map on a smooth manifold (unit circle) (Katok and Mezhirov, 1998), it has the shadowing property. However, as we show below, this is not the case for the system (7).

For some small $\delta > 0$, consider a δ-pseudo-trajectory that gets within the δ neighborhood of $p_{EQ} = 1/2$. Then, for *arbitrary* number of time steps $m \geq 1$ the pseudo-trajectory can stay in p_{EQ}. After that, the trajectory can continue with a high, or low value of p, depending on to which side of $p_{EQ} = 1/2$ it shifts. More formally, consider a pseudo-trajectory $....\tilde{p}_{n-2}\tilde{p}_{n-1}\tilde{p}_n$ with

$$|f(\tilde{p}_n) - p_{EQ}| < \delta.$$

We can then set $\tilde{p}_{n+1} = p_{EQ}$ and let the pseudo-trajectory stay in p_{EQ} for m time steps: $\tilde{p}_{n+1} = \tilde{p}_{n+2} = ... = \tilde{p}_{n+m} = p_{EQ}$. The next element of the pseudo-trajectory will be

$$\tilde{p}_{n+m+1} = f(p_{EQ}) + \nu = p_{EQ} + \nu,$$

where $|\nu| < \delta$ and $\nu \neq 0$.

It is impossible to closely shadow such pseudo-trajectories with real trajectories of (7). Given a small $\epsilon > 0$, for a real trajectory to ϵ-shadow the equilibrium p_{EQ} for an arbitrarily large number of steps m, it must oscillate around p_{EQ} within the ϵ-neighborhood of p_{EQ}. Since the action of $f(p)$ is the left shift of the binary representation $[p]_2$ of p, this is not possible for arbitrarily small ϵ. In fact, this is only possible for $\epsilon \geq 1/6$, e.g., the orbit is given by the initial condition $0.101010...$ and oscillates between $[2/3]_2 = 0.101010...$ and $[1/3]_2 = 0.010101....$

It follows, that the general shadowing property does not hold for this setting of the game and (μ, λ) selection. However, on a reduced set of initial conditions, one can argue that a 'shadowing-like' property holds at least for a particular interpretation of the computer round-off error (Sharkovsky and Chua, 1993): If the computer can guarantee only M exact binary digits, then iterative application of f on $0.x_1x_2x_3...$ will lead to the pseudo-trajectory $\mathcal{O}$: $0.x_1x_2x_3...x_{M-1}x_M$, $0.x_2x_3...x_{M-1}x_My_1$, $0.x_3...x_{M-1}x_My_1y_2$, ... $0.y_1y_2...y_M$, $0.y_2...y_{M+1}$, ..., for some $y_j \in \{0, 1\}$, $j \geq 1$. Consider now the set $\mathcal{E}$ of all pre-images under f of $p_{EQ} = 1/2$,

$$\mathcal{E} = \{p \in [0, 1] \mid f^n(p) = p_{EQ} \text{ for some } n \geq 0\}.$$

The real trajectory of (7) starting in $0.x_2x_3...x_{M-1}x_My_1y_2...$ from the f-invariant set $\Omega = (0, 1) \setminus \mathcal{E}$ will ϵ-shadow the pseudo-trajectory $\mathcal{O}$ with $\epsilon = 2^{-M+1}$.

The set $\mathcal{E}$ contains all $p \in [0, 1]$ whose binary expansion $[p]_2$ contains any finite word over the alphabet $\{0, 1\}$ (including the empty word), followed by digit 1, followed by the right-infinite sequence of 0's:

$$\mathcal{E} = \{p \in [0, 1] \mid [p]_2 = .\{0, 1\}^*1000...\}.$$

The set $\mathcal{E}$ is dense in $[0, 1]$, since for any $p \in [0, 1]$ and arbitrarily small $\epsilon > 0$, there will be a $q \in [0, 1]$, such that $|p - q| < \epsilon$ and $[q]_2$ has an infinite tail of 0's. Analogously, it is easy to show that the set Ω is dense in $[0, 1]$ as well. However, Ω is much larger than $\mathcal{E}$ - in fact, while $\mathcal{E}$ is countable, Ω is uncountable since it contains infinite expansion rational numbers and all irrational numbers in $[0, 1]$. Under f, the set Ω contains seeds for a wide variety of dynamical regimes, including periodic and 'chaotic' orbits.

DISCUSSION AND CONCLUSION

Infinite population models of co-evolutionary dynamics are useful mathematical constructs hinting at the possibility of a wide variety of possible dynamical regimes

- from simple attractive fixed point behavior, periodic orbits to complex chaotic dynamics. We have used the framework of shadowing lemma (form the theory of complex dynamical systems) to link such mathematical constructs to large finite population computer simulations. It has also been investigated whether the imposition of finite precision computer arithmetic does not leave the infinite population constructs and theories irrelevant.

If the co-evolutionary system possesses the shadowing property, finite population pseudo-trajectories from large scale computer simulations are shadowed by trajectories from the original infinite population model. This can be used, e.g., to calibrate time consuming large population computer simulations into regimes of interesting complex dynamics revealed by the infinite population mathematical models. The effects of finite population size on co-evolutionary dynamics under stochastic replicator dynamics was studied in (Ficici and Pollack, 2007). The focus of our approach is different. We ask - can one expect that as the population size grows the complex dynamical features of infinite population co-evolutionary dynamics will shadow features observed in large scale finite population simulations?

As examples, we have dealt with some settings of the two-strategy hawk-dove game with (μ, λ) and Boltzmann selection mechanisms. It has been shown there is indeed a possibility of quite intricate infinite population co-evolutionary dynamics and that this dynamics can have direct relevance for large scale finite population computer simulations under the smooth dynamics driven by Boltzmann selection. However, the discontinuous nature of the transition map under the (μ, λ) selection can prevent the co-evolutionary dynamics to posses the shadowing property. Hence, one cannot expect the infinite population trajectories to have direct relevance for large-population simulations. On the other hand, it is not all bad news: the computer generated trajectories of *infinite* population models can be shadowed by the real *infinite* population dynamics.

To the best of our knowledge, this paper represents the first attempt to address the issues of relevance of infinite population co-evolutionary dynamics simulations within the context of shadowing properties of complex dynamical systems.

Acknowledgements

This work was partially supported by an EPSRC grant (No. GR/T10671/01) on "Market Based Control of Complex Computational Systems".

REFERENCES

Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books, New York.

Balmforth, N. J., Spiegel, E. A., and Tresser, C. (1994). Topological entropy of one-dimensional maps: Approximations and bounds. *Phys. Rev. Lett.*, 72(1):80–83.

Chellapilla, K. and Fogel, D. B. (1999). Evolution, neural networks, games, and intelligence. *Proceedings of the IEEE*, 87(9):1471–1496.

Chong, S. Y., Tiňo, P., and Yao, X. (2008). Measuring generalization performance in co-evolutionary learning. *IEEE Transactions on Evolutionary Computation*, 12(4):479–505.

Darwen, P. J. and Yao, X. (2002). Co-evolution in iterated prisoner's dilemma with intermediate levels of cooperation: Application to missile defense. *International Journal of Computational Intelligence and Applications*, 2(1):83–107.

Ficici, S. and Pollack, J. (2007). Evolutionary dynamics of finite populations in games with polymorphic fitness-equilibria. *Journal of Theoretical Biology*, 247(13):426–441.

Ficici, S. G., Melnik, O., and Pollack, J. B. (2005). A game-theoretic and dynamical-systems analysis of selection methods in coevolution. *IEEE Transactions on Evolutionary Computation*, 9(6):580–602.

Fogel, D. B., Fogel, G. B., and Andrews, P. C. (1997). On the instability of evolutionary stable strategies. *Biosystems*, 44:135–152.

Hayesand, W. and Jackson, K. (2005). A survey of shadowing methods for numerical solutions of ordinary differential equations. *Appl. Numer. Math.*, 53(2):299–321.

Hofbauer, J. and Sigmund, K. (2003a). Evolutionary game dynamics. *Bulletin of the American Mathematical Society*, 40(4):479–519.

Hofbauer, J. and Sigmund, K. (2003b). Evolutionary game dynamics. *Bull. Amer. Math. Soc.*, 40:479–519.

Katok, A. and Hasselblatt, B. (1995). *Introduction to the modern theory of dynamical systems*. Cambridge University Press.

Katok, A. and Mezhirov, A. (1998). Entropy and growth of expanding periodic orbits for one-dimensional map. *Fundamenta Mathematicae*, 157:245–254.

Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V. V., editors (2007). *Algorithmic Game Theory*. Cambridge University Press, Cambridge, UK.

Sharkovsky, A. and Chua, L. (1993). Chaos in some 1-D discontinuous maps that appear in the analysis of electrical circuits,. *IEEE Transactions on Circuits and Systems I-Fundamental Theory and Applications*, 40(10):722–731.

Sigmund, K. and Nowak, M. A. (1999). Evolutionary game theory. *Current Biology*, 9:R503–R505.

Smith, J. M. (1982). *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, UK.

AUTHOR BIOGRAPHIES

PETER TIŇO is Senior Lecturer with the School of Computer Science, the University of Birmingham, UK. His main research interests include probabilistic modeling and visualization of structured data, statistical pattern recognition, dynamical systems, evolutionary computation, and fractal analysis.

SIANG YEW CHONG is an Assistant Professor with the School of Computer Science, University of Nottingham, Malaysia Campus, and a member of the Automated Scheduling, Optimization and Planning (ASAP) Research Group, University of Nottingham, UK. His major research interests include evolutionary computation, machine learning, and evolutionary game theory.

XIN YAO is professor of computer science in the School of Computer Science, the University of Birmingham, and the Director of the Centre of Excellence for Research in Computational Intelligence and Applications (CERCIA). Prof. Yao is a Fellow of IEEE, a Distinguished Lecturer of the IEEE Computational Intelligence Society, and a Distinguished Visiting Professor at the Nature Inspired Computation and Applications Laboratory (NICAL) of University of Science and Technology of China, Hefei, China. His main interests include evolutionary computation (evolutionary optimization, evolutionary learning, evolutionary design), neural network ensembles and multiple classifiers, meta-heuristic algorithms, data mining and computational complexity of evolutionary algorithms.

Robot Soccer - Strategy Description and Game Analysis

Jan Martinovič, Václav Snášel, Eliška Ochodková
Lucie Žołtá, Jie Wu,
VŠB - Technical University of Ostrava,
FEECS, Department of Computer Science
Ostrava, Czech Republic
Email:jan.martinovic@vsb.cz, vaclav.snasel@vsb.cz,
eliska.ochodkova@vsb.cz, l.zolta@seznam.cz,
defermat2008@hotmail.com

Ajith Abraham
Machine Intelligence Research Labs - MIR Labs,
USA, http://www.mirlabs.org
ajith.abraham@ieee.org

KEYWORDS

Robot Soccer, Cluster Analysis, Strategy, Prediction

ABSTRACT

The robot soccer game, as a part of standard applications of distributed system control in real time, provides numerous opportunities for the application of AI. Real-time dynamic strategy description and strategy learning possibility based on game observation are important to discover opponent's strategies, search tactical group movements and synthesize proper counter-strategies. In this paper, the game is separated into physical part and logical part including strategy level and abstract level. Correspondingly, the game strategy description and prediction of ball motion are built up. The way to use this description, such as learning rules and adapting team strategies to every single opponent, is also discussed. Cluster analysis is used to validate the strategy extraction.

INTRODUCTION

A typical example of a distributed control system with embedded subsystems is the control of robot soccer game (FIRA, 2010). The game can be described as double eleven autonomous mobile robots (home and visiting players), which are situated at the field with size of 220×180cm. Robot soccer is a challenging platform for multi-agent research, including real-time image processing and control, path planning, obstacle avoidance and machine learning. The robot soccer game presents an uncertain and dynamic environment for cooperating agents. (Bezek, 2005; Bezek et al., 2006; Tucnik et al., 2006) describe multi-agent strategic modeling of robot soccer. (Berger and Lämmel, 2007; Fard et al., 2007) give a very broad overview for using Case Based Reasoning techniques for this challenging problem.

Dynamic role switching and formation control are crucial for a successful game (Sng et al., 2002). The entire game can be divided into a number of partial tasks (Kim et al., 2004; Stankevich et al., 2005), such as evaluation of visual information, image processing, hardware and software implementation of distributed control system, hard-wired or wireless data transmission, information processing, strategy planning and control of robots. Complex control tasks can often be solved by decomposing them into hierarchies of manageable subtasks (Whiteson et al., 2005).

Because of the attraction of robot soccer, many interesting approaches were developed to improve robot soccer. An abstract description of the game was developed in (Obitko and Snasel, 2004; Smid et al., 2004; Srovnal et al., 2007), together with the ways to use this description. In (Zhao et al., 2006), it discussed the strategy based on opponent information in the Robot-Soccer Game. We adopted this representation to control robots and learn game strategies.

In our approach, the game is separated into logical and physical parts. The logical part includes the strategy selection, calculation of robot movement and adaptation of rules to the opponent's strategy. The physical part includes robot actual movement on the game field and recognition of the opponent movement. The logical part is independent of the physical part because we can calculate movements of the opponent robots as well as movements of own robots.

By separating the game into two parts, the logical part is independent of the field size and the resolution of the camera used in visual information system. In the logical part, the game is represented as an abstract grid with a very high resolution, which ensures a very precise position of robots and ball. However, this very detailed representation of the game field is not suitable for strategy description. Too many rules are required to describe robot behavior. Therefore, a strategy grid, which has a much lower resolution than an abstract grid, is used. This simplification of reality is sufficient because it is not necessary to know the exact position of robot. It is sufficient to know its approximate position for strategy realization (Fig. 1). When the physical part is used, we only need to transform the coordinates from the abstract grid into coordinates based on the game field size and camera resolution.

The rest of the paper is organized as follows. Section 2 describes possible game strategies. Using the abstract grids and game strategies, it is explained how to learn rules that describes specific game strategies. Our

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

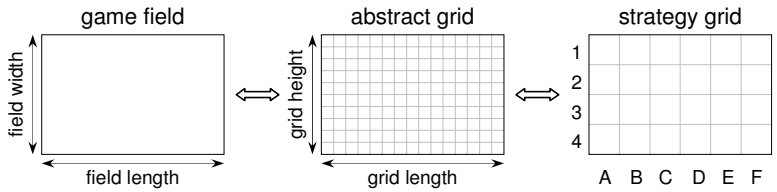


Figure 1: Inner game representation.

approach to predict the movement of ball is shown in section 3. It is based on the grids, respectively on strategy level and abstract level, which is helpful to determine the game strategy and control the robot formation. Section 4 discusses the strategy extraction based on cluster analysis. Section 5 concludes with the discussion of the presented approach.

GAME STRATEGY

The game strategy can be dynamically changed based on game progress (i.e. the history and the current position of the players and the ball (Veloso and Stone, 1998)).

In this section, we describe our approach for learning game strategy from observation. Similar approach we can see in (Ros and Arcos, 2007; Ros et al., 2007). Our objective is to learn an abstract strategy. The main steps of the learning process are:

- Transformation of observations into abstract grids.
- Transformation of observations into strategy grids.
- Learning a strategy based on the observed transitions in the strategy grid.

The movement of a particular robot is determined by the current game class and situation, and by the robot's role. For example, the goalkeeper's task is to prevent the opponent from scoring a goal. In most cases, its movements are limited along the penalty zone and near the goal line. The preferred movements are in the direction of the goal line. The goalkeeper's movements ensure that the ball will be kicked from the defence zone.

We adopt this definition of strategy (Johnson and Scholes, 2001): "Strategy is the direction and scope of an organization over the long-term, which achieves advantage for the organization through its configuration of resources within a challenging environment..."

Strategy application for one movement of players is computed in following steps:

1. Get coordinates of players and ball from camera
2. Convert coordinates of players into strategic grid
3. Convert ball and opponents' positions into abstract and strategic grids
4. Choose goalkeeper and attacker, exclude them from strategy and calculate their exact positions.
5. Detect strategic rule from opponents' and ball positions
6. Convert movement from strategic grid to physical coordinates
7. Send movement coordinates to robots

Each strategy is stored in one file and currently consists of about 15 basic rules.

```
.Strategy "test"
.Algorithm "Offensive"
.Author "Vaclav Snasel"
.Date "19.12.2008"
.Size 6 4
.PriorityMine      100 100 100 100 100
.PriorityOpponent  50   50  50  50  50
.PriorityBall      50

.Rule 1 "Attack1"
.Mine c2  c3  b1  b4
.Opponent d2  d3  e1  e4
.Ball d3
.Move d2  c3  b1  b4

.Rule 2 "Attack2"
.Mine d2  c3  b1  b4
.Opponent d2  d3  e1  e4
.Ball c3
.Move d2  c4  c1  b4

.Rule 3 "Attack3"
.Mine d2  c4  c1  b4
.Opponent d2  d3  e1  e4
.Ball c3
.Move d2  c4  b2  b3
```

Furthermore the file contains following metadata:

- Information about the name of strategy
- The algorithm to strategy choosing
- The author responsible for current strategy
- The date of last modification
- The size of strategic grid (e.g. Fig. 1)
- Strategic rules

Each strategic rule consists of five records:

- The rule ID and description (e.g. Rule 1 "Attack1"),
- the coordinates of our players in strategic grid (e.g. .Mine c2 c3 b1 b4),
- the coordinates of opponent's players in strategic or abstract grid (e.g. .Opponent d2 d3 e1 e4),

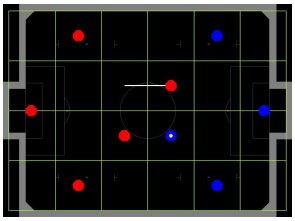


Figure 2: Rule Attack1.

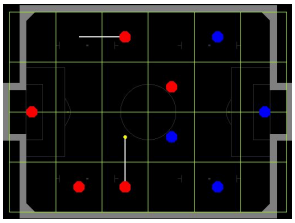


Figure 3: Rule Attack2.

- the ball coordinates in abstract or strategic grid (e.g. .Ball d3)

- strategic or abstract grid positions of the move (e.g. .Move d2 c3 b1 b4).

```
// algorithm for rule selection
// Game.Mine      -- actual positions
// Game.Opponent  -- actual positions
// Game.Ball      -- actual position
maxWeight = 0
SelectRule = 0
foreach r in Rule
{
  ruleTmp = r.Mine
  SumToMine = 0
  foreach p in Game.Mine
  {
    s = nearest position in ruleTmp to p
    w = 1 / (distance(s, p) + 1)
    w = Strategy.PriorityMine * w
    SumToMine = SumToMine + w
    remove s from ruleTmp
  }
  weight = SumToMine

  ruleTmp = r.Opponent
  SumToOpponent = 0
  foreach p in Game.Opponent
  {
    s = nearest position in ruleTmp to p
    w = 1 / (distance(s, p) + 1)
    w = Strategy.PriorityOpponent * w
    SumToOpponent = SumToOpponent + w
    remove s from ruleTmp
  }
  weight = weight + SumToOpponent

  ToBall = 1 / (distance(Game.Ball, r.Ball) + 1)
  ToBall = Strategy.PriorityBall * w
  weight =  weight + ToBall
  if weight > maxWeight
  {
    maxWeight = weight
    SelectRule = r
  }
}
return SelectRule
```

A basic strategy item is a movement in strategy grid. An example may be the following rule:

if $(M_1, M_2, M_3, M_4, M_5)$ is close to $(c2, c3, b1, b4)$
and $(O_1, O_2, O_3, O_4, O_5)$ is close to $(d2, d3, e1, e4)$
and B is close to $(d3)$
then $(M_1, M_2, M_3, M_4, M_5)$ "go to" $(d2, c3, b1, b4)$

By observing opponent's strategy, a new set of rules can be written without the necessity of a program code modification. Furthermore, there is a possibility of automatic strategy (movement) extraction from the game in progress. There are two main criteria in the selection process rules. The selection depends on opponent coordinates, own coordinates and ball position. The strategy file contains rules, describing three possible formations suggesting danger of the current game situation. The opponent's team can be in offensive, neutral or defensive formations. Furthermore, we need to consider ball position risk. Generally, opponent is not dangerous if the ball is near his goal. The chosen rule has minimal strategic grid distance from the current rule.

Optimal movements of our robots are calculated by applying a minimal distance from strategic grid position. The goalkeeper and attacking player, whose distance is closest to the ball, are excluded from strategic movement and their new position is calculated in exact coordinates. To summarize, strategy management can be described in the following way:

- Based on incoming data from the vision system, calculate abstract and strategy grid Coordinates of the players and the ball.

- The abstract grid is then used to decide which player has under the ball control.

- This player is issued a "kick to" command that means that it has to try to kick the ball to a given strategy grid coordinates.

- All other players are given (imprecise) "go to" coordinates. These coordinates are determined by the current game strategy and are determined for each robot individually. The goalkeeper is excluded from this process since its job is specialized, and does not directly depend on the current game strategy.

PREDICTION OF BALL MOTION

Obviously, the prediction of ball motion is very important to the robot soccer. The prediction includes ball track, collision and rebound. Strategically, it is helpful to select proper strategy and regulate robot formation. Tactically, it is contributing to break through, intercept or steal the ball at right position right moment. To sum up, the prediction is a necessary step for defensive and offensive in robot soccer.

In the game representation, there are two kinds of grids, strategy grid and abstract grid. The strategy grid exists in our representation because it is not necessary to know the robot exact position in strategy hierarchy. Similarly, the strategy grid can be used to predict the ball

motion in strategy level, because it is enough to predict the ball approximate position for strategy realization. On the other hand, for the prediction, the abstract grid is necessary to ensure the precision. Therefore, the ball motion should be predict on two levels – strategy and abstract.

PREDICTION ON STRATEGY LEVEL

About the prediction, it could be very simple if the ball velocity is known. However, for the ball motion prediction on strategy level, it is not necessary to predict the ball exact position and we just need to predict the ball approximate position at next frame in the strategy grid. Therefore, this prediction could be further simplified.

Figure 4 shows the ball lying in a strategy grid. The velocity of ball can be decomposed into x-component and y-component. In Fig. 4, there are eight potential positions for the ball at next moment. But the final position depends on the relative magnitude of x-component and y-component. That means we just need to compare those two velocity components with each other, then the new position of ball can be predicted. For the case shown in Fig. 4, because the x-component is much greater than y-component, the ball will move into the left grid at next moment. By this means, the prediction is further simplified to boolean operation, which would reduce the operation time. It's very important to the strategy selection.

In addition, there are two noticeable cases in strategy level prediction. The first one is the ball lies close to boundary, which means potential collision and rebound to the wall. In this case, the number of possible new positions is less than eight, so it's simpler again. The second one is the ball is likely to collide with robot during movement. Regardless of which party robot is, the robot will certainly catch the ball. Therefore, in this case, the ball will lie in the grid in which the colliding robot lies.

PREDICTION ON ABSTRACT LEVEL

When the ball is caught by robot, the motion track of ball is under the control of robot, and in this case, it's not necessary to predict. When the ball is free, the ball would move along a line. Generally, the ball is usually under the control of robot. That means the ball would move freely only for a short time, so it's reasonable to neglect the acceleration, then the motion track can be depicted by simple linear model.

For the prediction on abstract level, it's necessary to predict a very precise position of ball, so the grid must be presented with a very high resolution. To solve the new position of ball, coordinate system should be built according to the grid (see Fig. 5). The track equation of ball can be expressed as

$$\begin{cases} x &= x_0 + v_x t \\ y &= y_0 + v_y t \end{cases}$$

where x and y are respectively predictive coordinates, x_0 and y_0 are initial coordinates, v_x and v_y are two components of ball velocity. The initial coordinate should be re-read from visual information system at the moment of every touch, collision or rebound to something before the ball move freely again.

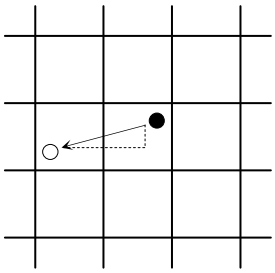


Figure 4: Ball in strategy grid.

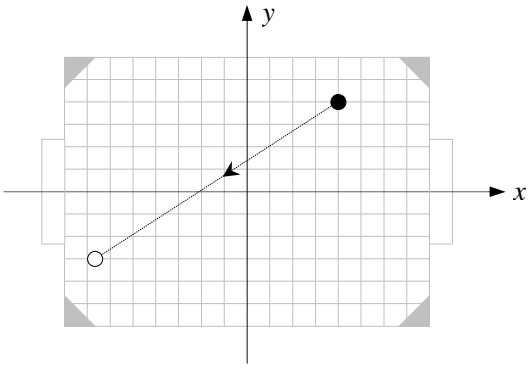


Figure 5: Ball in abstract grid.

Rebound track is very important to the prediction of ball motion. In fact, we developed a method to calculate the detailed position of ball, but we have to simplify it because of the time urgency. The rebound mentioned here only refers to the wall rather than to the robot, because the robot will catch the ball and the rebound to robot would happened scarcely.

Given the collision to the wall is elastic collision, so the post-rebound velocity has the same magnitude to the pre-rebound velocity except orientation. In other words, the velocity could be decomposed into two quadrature-components, the new velocity could be gotten by sign-changing one component while maintain another one. According to the new velocity, the ball rebound track could be predicted. For example, if the ball move forward along the line shown in Fig. 5, it will collide to the left of the wall and then rebound. After rebound, the ball track equation can be expressed as

$$\begin{cases} x &= x'_0 - v_x t \\ y &= y'_0 + v_y t \end{cases}$$

GAME ANALYSIS

A log, which describes in detail the course of the game, is created during the game. Log contains hundreds of lines to describe each game situation (my robots and opponent's robots positions, ball position, what rule was used, etc.). Records in the game log, which can be seen in Ex-

ample 1, contain real and grid positions. We obtain more than a thousand such records (depending on the performance of your computer) during one game.

Example 1

Entry (one line) from the log-file (starting position)

$14; 0; 0; 4; 3; -36, 67; -30; 3; 2; -36, 67; 30; 3; 3;$
$-73, 33; -60; 2; 1; -73, 33; 60; 2; 4; -105; 0; 1; 3;$
$36, 67; -30; 4; 2; 36, 67; 30; 4; 3; 73, 33; -60; 5; 1;$
$73, 33; 60; 5; 4; 105; 0; 6; 3;$

■

By analyzing this log, we can derive the opponent's strategy or optimize our strategy. First, it is necessary to reduce the cardinality of logs to a "reasonable number" of representative situations (records). To reduce records, we used clustering using minimum spanning tree. It was therefore necessary to create a graph in which vertices represent records in the log (game situation) and edges represent similarities between them.

One question is how to compare two game situations. Therefore, each record was transferred to a matrix (map), which is a simplified description of the game situation. Each matrix in this new log consists of three maps of the course in the grid coordinate system (see Table 1).

The first part of the matrix shows the field distributed grid in the ratio 4×6 and the position of the ball. Second part (six columns) represents my positions and last six columns represent opponent's positions. Number one in the matrix represents the presence of the object (ball, my robot, opponent robot).

Thus described game situation can be better compared by e.g. the following relationship used in Equation 1, which sums up the differences between the various elements of the matrix. This method is used here only as an illustrative method because it is not completely accurate, since only considers whether or not there was a change (Example 2).

$$\sum_{i,j=0}^{i\leq r, j\leq s} |A[i,j] - B[i,j]|, \qquad (1)$$

where A, B are compared matrices, r is the number of rows and s number of columns.

Example 2

Comparison of three situations (matrix simplified to vector)

$$\begin{array}{lcccccc} situation_1: & 0 & 0 & 0 & 1 & 0 & 0, \\ situation_2: & 0 & 0 & 1 & 0 & 0 & 0, \\ situation_3: & 1 & 0 & 0 & 0 & 0 & 0, \end{array}$$

$situation_1 - situation_2 = |0-1| + |1-0| = 2,$
$situation_1 - situation_3 = |0-1| + |1-0| = 2.$

In both cases, there was a change toward the $situation_1$. Change is, however, expressed by the same coefficient of similarity, 2, although the $situation_3$ differs from the $situation_1$ more than the $situation_2$ differs from the $situation_1$. ■

For a more precise comparison, the binary matrix was transferred to a matrix "with less contrast"'. Elements with value 1 were increased by a constant c and their vicinity by the value $(c-1)$ etc. (see Table 2 for $c = 1$).

It is now more accurate comparison, as we can see in Example 3. The difference between $situation_1$ and $situation_2$ is less than the difference between $situation_1$ and $situation_3$ now.

Example 3

Comparison of three situations (matrix simplified to vector)

$$\begin{array}{lcccccc} situation_1: & 0 & 0 & 1 & 2 & 1 & 0, \\ situation_2: & 0 & 1 & 2 & 1 & 0 & 0, \\ situation_3: & 2 & 1 & 0 & 0 & 0 & 0, \end{array}$$

$situation_1 - situation_2 = |0-1| + |1-2| + |2-1| + |1-0| = 4,$
$situation_1 - situation_3 = |0-2| + |0-1| + |1-0| + |2-0| + |1-0| = 7.$ ■

Now it was possible to reduce the quantity of records. As mentioned above, we will work with complete graph with vertices representing records and edges representing similarities between them (edges are valued with a number indicating the similarity (the more similar the smaller value and 0 means identical records)). Identical records were written out when creating the graph. From this graph the minimum spanning tree was obtained with the Kruskal's algorithm and by removing the most expensive edges from the spanning tree clusters were created. As a representative element (we named it as centroid) of the cluster, we selected the vertex with the highest degree. The resulting centroids will be used to develop estimates of the opponent's strategy. Each centroid contains information about ball, of my players and opponents. In order to generate the strategy of the opponent, we need to add the following information about shift in the strategic move. This shift is obtained from the original log. In it we find the row i, which represents the selected centroid. Shift can be determined from the $i + k$-th following line, where k will define the jump in the log.

CONCLUSION

The main goal of the control system is to enable a real-time response. The method we described provides fast control. This is achieved by using rules those are fast to process. We have described a method of game representation and a method of learning game strategies from observed movements. The movements can be observed from the opponent's behavior or from the human player's behavior. We believe that the possibility of learning the game strategy that leads to a fast control is critical for

Table 1: Matrix (starting position): position of the ball - of my robots - of the opponent robots

0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0
0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	1	0	1
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0

Table 2: The revised matrix for $c = 1$ (starting position)

0	0	0	0	0	0	1	3	2	1	0	0	0	0	1	2	3	1
0	0	1	1	1	0	2	4	4	2	0	0	0	0	2	4	4	2
0	0	1	2	1	0	3	4	4	2	0	0	0	0	2	4	4	3
0	0	1	1	1	0	2	4	2	1	0	0	0	0	1	2	4	2

success of robot soccer game. Likely to the chess playing program, the database of successful game strategies can be stored in the database and can be used for subsequent matches (Sadikov and Bratko, 2006).

ACKNOWLEDGEMENT

We acknowledge the support of project SP/2010196 Machine Intelligence.

REFERENCES

Barrios-Aranibar, D. and Alsina, P. J. (2005). *Recognizing Behaviors Patterns in a Micro Robot Soccer Game.* HIS IEEE, Pages: 463–468

Berger, R. and Lämmel, G. (2007). *Exploiting Past Experience Case-Based Decision Support for Soccer Agents.* KI 2007: Advances in Artificial Intelligence, LNCS 4667, Pages: 440–443.

Berry, M. W. and Browne, M. (1999). *Understanding Search Engines*. SIAM Society for Industrial and Applied Mathematics, Philadelphia.

Bezek, A. (2005). *Discovering strategic multi-agent behavior in a robotic soccer domain.* AAMAS 2005: Pages: 1177–1178.

Bezek, A., Gams, M. and Bratko, I. (2006). *Multi-agent strategic modeling in a robotic soccer domain.* AAMAS 2006: Pages: 457–464.

Fard, A. M., Salmani, V., Naghibzadeh, M., Nejad, S. K. and Ahmadi H. (2007). *Game Theory-based Data Mining Technique for Strategy Making of a Soccer Simulation Coach Agent.* ISTA 2007: Pages: 54–65.

FIRA robot soccer, http://www.fira.net/ (01-04-2010)

Johnson, G. and Scholes, K. (2001). *Exploring Corporate Strategy: Text and Cases.* FT Prentice Hall.

Kim, J., Kim, D., Kim, Y. and Seow, K. (2004). *Soccer Robotics.* Springer Tracts in Advanced Robotics, Springer-Verlag.

Obitko, M. and Snášel, V. (2004). Ontology Repository in Multi-Agent System. Artificial Intelligence and Applications - Volume I & II. Calgary: Acta Press, vol. 1, Pages: 853–858.

Ros, R. and Arcos, J. L. (2007). Acquiring a Robust Case Base for the Robot Soccer Domain. IJCAI 2007: Pages: 1029–1034.

Ros, R., de Mántaras, R. L., Arcos, J. L. and Veloso, M. M. (2007). *Team Playing Behavior in Robot Soccer: A Case-Based Reasoning Approach.* ICCBR 2007: Pages: 46–60.

Sadikov, A. and Bratko, I. (2006). *Learning long-term chess strategies from databases.* Machine Learning (2006) 63:3 Pages: 329–340.

Smid, J., Obitko, M. and Snášel V. (2004). *Communicating Agents and Property-Based Types versus Objects.* In SOFSEM 2004 - Theory and Practice of Computer Science. Prague: MATFYSPRESS, Pages: 154–163.

Sng, H.L., Gupta, G.S. and Messom, C.H. (2002). *Strategy for Collaboration in Robot Soccer.* The First IEEE International Workshop on Electronic Design, Test and Applications, Pages: 347–251.

Srovnal, V. , Horák, B. , Snášel, V., Martinovič, J., Krömer, P. and Platoš, J. (2007). *Strategy Description for Mobile Embedded Control Systems Exploiting the Multi-agent Technology.* International Conference on Computational Science (2), Pages: 936–943.

Stankevich, L., Serebryakov, S. and Ivanov, A. (2005). *Data Mining Techniques for RoboCup Soccer Agents.* AIS-ADM 2005: Pages: 289–301.

Tučník, P., Kožaný, J. and Srovnal, V. (2006). *Multicriterial Decision-Making in Multiagent Systems.* Computational Science ICCS 2006, LNCS 3993, Pages: 711–718.

Veloso, M. and Stone, P. (1998). *Individual and collaborative Behaviours in a Team of Homogeneous Robotic Soccer Agents.* Proceedings of International Conference on Multi-Agent Systems, 1998, Pages: 309–316.

Zhao, X., Zhang, J., Li, W. and Li, Y. (2006). *Research on Strategy of Robot Soccer Game Based on Opponent Information.* IEEE Machine Learning and Cybernetics, Pages: 230–234.

Whiteson, S., Kohl, N., Miikkulainen, R. and Stone, P. (2005). Evolving Soccer Keepaway Players Through Task Decomposition. Machine Learning, Volume 59:1 Pages: 5–30.

NETWORK FLOWS IN OPTIMISATION PROBLEMS AND THEIR EXTENSIONS

Miloš Šeda
Institute of Automation and Computer Science
Faculty of Mechanical Engineering, Brno University of Technology
Technická 2, Brno 616 69, Czech Republic
E-mail: seda@fme.vutbr.cz
EPI Kunovice, Osvobození 699, Kunovice 686 04, Czech Republic

KEYWORDS
Maximum flow, Steiner tree, flows with priorities, multicommodity flow, NP-hard problem, heuristic method

ABSTRACT

Network flow problems are among the most important ones in graph theory. Since there are many well-known polynomial algorithms for solving the classical Maximum Flow Problem, we, instead of summarising them, focus on special formulations and their transformation into the basic one, and because other graph theory problems may be formulated with the help of network flow tools, we show how to formulate the Minimum Steiner Tree Problem using the maximum network flow terminology and derive its mathematical model.
Finally, we discuss the Integer Maximal Multi-commodity Flow Problem. Since this network flow version belongs to the class of NP-hard combinatorial problems, for large scale instances, it must be solved by approximation or heuristic techniques. We present a stochastic heuristic approach based on a simulated annealing algorithm.

INTRODUCTION

The flow problems have many specific formulations depending on the constraints imposed resulting in various applications in transportation, distribution, telecommunications, etc. Examples of flow networks can be found, e.g., in electrical circuits, pipe networks, and city plans. In the literature, many other important applications are described, such as machine scheduling, assignment of computer modules to computer processors, tanker scheduling, project management, facility layout and location, production and inventory planning, selection problem solving, open-pit mining problem, forest clearing problem, producing memory chips (VLSI layout), maximum closure problem, vertex cover on bipartite graphs, and mesh generation.

For some of the flow problems such as maximum flow problem, minimum cost flow problem, polynomial time algorithms are known (Ahuja et al 1993), others such as integer maximum multicommodity flow problem (Ouorou et al, 2000), (Hochbaum 1996), (Plesník 1983) are NP-hard combinatorial problems and, for large-scale instances, must be solved by approximation or using heuristic techniques.

Now, we will summarise some network-flow-related graph-theoretic concepts and, without giving proofs, mention the key theorems that result in the classical algorithms to compute a maximal flow.

Definition 1.
A (*single source – single sink*) *network* is a quadruple $N=(G,s,t,c)$ where $G=(V,E)$ is a simple directed graph without loops, $s \in V$ is the *source* of network N, $t \in V$, $t \neq s$ is the *sink* of network N and c: $E(G) \rightarrow \mathbf{R}^+$ is a nonnegative edge-weight function called the *capacity* of network N. For each edge $e \in E$, we call the corresponding value $c(e)$ the capacity of edge e. □

Definition 2.
A *flow* in a network $N=(G,s,t,c)$ is an edge weight function f: $E(G) \rightarrow \mathbf{R}$ satisfying the following conditions (the second one is also known as Kirchhoff's law):

(a) $(u,v) \in E(G)$: $0 \leq f(u,v) \leq c(u,v)$ [*capacity constraints*]

(b) $\forall v \in V - \{s,t\}$: $\sum_{(u,v)\in E(G)} f(u,v) - \sum_{(v,w)\in E(G)} f(v,w) = 0$

[*conservation-of-flow constraints*]

□

Definition 3

- The *value* $|f|$ *of flow* f is the total network flow leaving the source (= network flow entering the sink).

$$|f| = \sum_{(s,v)\in E(G)} f(s,v) = \sum_{(v,t)\in E(G)} f(v,t)$$

- If $f(u,v) = c(u,v)$, then we say that (u,v) is a *saturation arc*.
- A flow f is called *saturated* if, for each directed path $s = v_0, \ldots, v_k = t$ from s to t, there is i such that $f(v_{i-1}, v_i) = c(v_{i-1}, v_i)$.
- A *maximum flow* f^* in a capacitated network N is a flow in N having the maximum value, i.e. $|f| \leq |f^*|$ for every flow f in G. □

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

Maximum flow problem is formulated as the following linear programming problem:

$$\text{Maximise } |f| = \sum_{(s,v)\in E(G)} f(s,v) \quad (1)$$

subject to

$$(u,v)\in E(G): 0 \le f(u,v) \le c(u,v) \quad (2)$$

$$\forall v \in V - \{s,t\}: \sum_{(u,v)\in E(G)} f(u,v) - \sum_{(v,w)\in E(G)} f(v,w) = 0 \quad (3)$$

SPECIAL FORMULATIONS AND THEIR TRANSFORMATION

In real situations, many other constraints may be added: the capacities of vertices may be constrained in addition to edge constraints (the number of cars, for example, passing a crossroads during a certain time interval).

Networks with constrained vertex capacities

If the maximum flow that can pass through a vertex is constrained, such as by $\sum_{u\in V} f(u,v) \le w_v$, then

1. we split vertex v into 2 new vertices v' and v" connected by an edge having the capacity wv,
2. all incoming edges of v will enter into v', and all outgoing edges of v will be moved to start in v".

Networks with more sources and sinks

Let us assume that we need to determine the maximal flow in a network with m sources and n sinks and each source can supply each sink.

This problem can be transformed into the basic problem with a single source and sink by adding a supersource and a supersink. The supersource will be connected by an outgoing edge with each original source and the supersink by incoming edges with each original sink. The weights of these new edges are initialised with unrestricted capacities.

Maximal flow with priorities of sources and sinks

In practice, sources and sinks are frequently ordered by their priorities (Plesník 1983), i.e. first the demands are met from a source with the highest priority, then from the source with the second highest priority, etc. Similarly, priorities can be defined for sinks, i.e. first the demands of a sink with the highest priority are met, etc.

Definition 4
Let f be a flow from sources $s_1, s_2, \dots, s_m$ to sinks $t_1, t_2, \dots, t_n$ and $|f^+(s)| = \Sigma_{(s,v)\in E(G)} f(s,v)$ denote the outflow from source s and $|f^-(t)| = \Sigma_{(v,t)\in E(G)} f(v,s)$ inflow into sink t.
Then the flow f is called *lexicographically maximal with respect to sources* if the vector $(|f^+(s_1)|,|f^+(s_2)|,\dots,|f^+(s_m)|)$ is lexicographically maximal. Similarly, the flow f is called *lexicographically maximal with respect to sinks* if the vector $(|f^-(t_1)|,|f^-(t_2)|,\dots,|f^-(t_n)|)$ is lexicographically maximal.
□

Note that the *lexicographical order* $\ge$ is defined as follows: Let $\mathbf{a} = (a_1, a_2, \dots, a_n)$ and $\mathbf{b} = (b_1, b_2, \dots, b_n)$, then $\mathbf{a} \ge \mathbf{b} \Leftrightarrow_{df}$ $\mathbf{a} = \mathbf{b}$ or $\exists\, k\in\{1, \dots, n\}: a_k > b_k$ and $a_i = b_i\ \forall i\ (1 \le i < k)$.

The lexicographically maximal flow with respect to sources can be found as shown in Figure 2. All sinks are connected to a supersink t with edges having infinite capacities, a supersource s is connected to source s_1 by an edge of infinite capacity and a maximal flow f_1' from s to t is computed. Then the edge $s - s_2$ is added and, starting from f_1', the flow f_{12}' is determined and, from f_{12}' after addition of the edge $s - s_3$, the flow f_{123}' is determined, etc. The lexicographically maximal flow with respect to sinks can be obtained by a symmetrical approach.

Theorem 3. *The lexicographically maximal flow with respect to sources* (*or sinks*) *is the maximal flow.*
Proof. This is straightforward since, otherwise, an augmenting path from s_i to t_j should exist and, by increasing the flow along the augmenting path, we would find a flow that was lexicographically bigger. ■

A lexicographically maximal flow with respect to sources $f_{12\dots m}$' and a lexicographically maximal flow with respect to sinks $f_{12\dots n}$'' can be combined. Since both these flows are maximal, there is a cut $(A, A\sim)$ separating the source and the sink where edges $(u,v)\in(A\times A\sim)\cap E$ are saturated by both of them and, along each edge $(v,u)\in(A\sim\times A)\cap E$, the flow is zero. The resulting flow can be created by using the flow $f_{12\dots m}'(u,v)$ for edges $(u,v)\in(A\times A)\cap E$ and the flow $f_{12\dots n}''(u,v)$ for the other edges .

NETWORK STEINER TREE PROBLEM AND ITS NETWOK FLOW FORMULATION

The *Network Steiner tree problem* (NSTP) (or *Steiner tree problem in graphs*) (Du et al. 2000), (Hwang et al. 1992) is concerned with connecting a subset of vertices at a minimal cost. More precisely, given an undirected connected graph $G=(V,E)$ with vertex set V, edge set E, nonnegative weights associated with the edges, and a subset B of V (called *terminals* or *customer vertices*), the problem is to find a subgraph T that connects the vertices in B so that the sum of the weights of the edges in T is minimised. It is obvious that the solution is always a tree and it is called a *Steiner minimum tree* for B in G.

Applications of the NSTP are frequently found in the layout of connection structures in networks and circuit. Their common feature is that of connecting together a set of terminals (communications sites or circuits components) by a network of the minimal total length.

If $|B|=2$ then the problem reduces to the *shortest path problem* and can be solved by Dijkstra's algorithm. In

the case of $B=V$ the NSTP reduces to the *minimum spanning tree* (MST) *problem* and can be solved by Jarník's (Prim's), Borůvka's or Kruskal's algorithm. All these algorithms are polynomial. However, in the general case the NSTP is NP-complete (Hwang et al. 1992), (Plesník 1983) and therefore it cannot be solved exactly for larger instances, i.e. heuristic or approximation methods must be used. Normally a Steiner minimum tree is not a minimum spanning tree only, it can also span some nonterminals, called *Steiner vertices*.

Let $V=\{1,2,\ldots,n\}$ and S be a set of Steiner vertices. For every edge (i,j), c_{ij}, $c_{ij} \geq 0$ is a weight of the edge. The aim is to find a connected graph $G'=(B \cup S, E')$ (Steiner tree), $E' \subset E$, for the sum of weights to be minimal.

In other words, the Steiner minimum tree problem can be described as a problem of finding a set of edges that connects terminals.Therefore we can define a bivalent variable x_{ij} for each edge $(i,j) \in E$ indicating whether the edge (i,j) is included into the Steiner tree ($x_{ij} = 1$) or not ($x_{ij} = 0$) and similarly a bivalent variable f_i indicating whether vertex i is included in the Steiner tree ($f_i = 1$) or not ($f_i = 0$). For terminals, $i \in B$, it is satisfied $f_i = 1$, and $f_i = 0$ for the other vertices, $i \in (V-B)$.

Using these denotations, we can derive the model based on a network flow formulation of the NSTP as follows. The variable y_{ij} represents a flow through the edge $(i,j) \in E'$.

$$\text{Minimise} \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} x_{ij} \tag{4}$$

subject to

$$r := \min \{k \mid k \in B\} \tag{5}$$

$$\forall j \in (V - \{r\}): x_{rj} = 0 \tag{6}$$

$$\forall i \in (V - \{r\}): \sum_{j=1}^{n} x_{ij} = f_i \tag{7}$$

$$\forall i \in (V - \{r\}): \sum_{j=1}^{n} y_{ij} - \sum_{j=1, j\neq r}^{n} y_{ij} = f_i \tag{8}$$

$$\forall i,j \in V: \; y_{ij} \leq (n-1) x_{ij} \tag{9}$$

$$\forall i,j \in V: \; y_{ij} \in \mathbf{Z}_+ \tag{10}$$

$$\forall i,j \in V, c_{ij} = 0: x_{ij} := 0 \tag{11}$$

$$\forall i,j \in V: \; x_{ij} \in \{0,1\} \tag{12}$$

$$\forall i \in B: \; f_i = 1 \tag{13}$$

$$\forall i \in (V-B): f_i \in \{0,1\} \tag{14}$$

where $\mathbf{Z}_+$ denotes the set of nonnegative integers.

MULTICOMMODITY FLOWS

Multicommodity flow problems are a special class of operations research problems with many applications such as transportation, distribution and telecommunications. If the commodities do not share common facilities, we would solve each single-commodity problem separately using traditional polynomial algorithms. Unfortunately, some resources (e.g. capital, labour, equipment, space) can be shared by several commodities.

This problem has a variety of formulations depending on the constraints defined. The paper (Ouorou *et al*., 2000) provides a survey of algorithms for convex multicommodity flow problems. A special case of the problem when the capacity constraints are modelled by random variables because some of the network nodes may fail, is studied in (Lin 2000). The paper (Ouorou and Mahey 2000) presents a nonlinear version of the problem where a nonlinear cost function associated with each edge is considered. Special cases of multicommodity flow problems are studied in (Lozovanu and Fonoberova 2006), (Wang and Kleinberg 2009) and (Srivastav and Stangier 2000).

Consider a network represented by a directed graph $G=(V,E)$ with $n=|V|$ vertices and $m=|E|$ edges. Let K be a set of commodities, let, for $k \in K$, s^k and t^k represent the source and sink vertices for the commodity k, x_{ij}^k and c_{ij}^k be the flow of commodity k along the edge (i,j) and its capacity constraint, let further c_{ij} represent the amount of some common resource available for all commodities combined, w_{ij}^k be the amount of this resource needed for processing one unit of commodity k along the edge (i,j), and let, finally, f^k be the total flow of commodity k leaving the source vertex s^k and reaching the sink vertex t^k. Using this notation and requiring the flows to be restricted to integer values, the integer maximal multicommodity flow problem can be formulated as the following integer linear programming problem:

$$\text{Maximise} \quad \sum_{k \in K} f^k \tag{15}$$

subject to

$$\sum_{j:(i,j)\in E} x_{ij}^k - \sum_{j:(j,i)\in E} x_{ji}^k = \begin{cases} f^k & , \; i = s^k, k \in K \\ 0 & , \; i = V - \{s^k, t^k\}, k \in K \\ -f^k & , \; i = t^k, k \in K \end{cases} \tag{16}$$

$$x_{ij}^k \leq c_{ij}^k , \quad (i,j) \in E, k \in K \tag{17}$$

$$\sum_{k \in K} w_{ij}^k x_{ij}^k \leq c_{ij} , \quad (i,j) \in E \tag{18}$$

$$x_{ij}^k \in Z_+ , \quad (i,j) \in E, k \in K \tag{19}$$

$$f^k \in Z_+ , \quad k \in K \tag{20}$$

Equation (20) can be omitted because this integrality constraint results from (19). However, this implication does not hold in the opposite direction. In spite of an optimal integer-valued solution f^k, the flows along the

edges may be multiples of ½. Furthermore the optimal flows are not guaranteed to be integer when the capacity of each of the edges is integer as it is in the case of a single commodity flow problem (Ahuja et al 1993).

The integer maximal multicommodity flow problem will be solved using the following allocation strategy. Initially, assign the maximum possible integer capacity to each edge $(i,j)\in E$ for commodity 1 and find the maximum flow. Reduce the capacity of each edge (i,j) by the flow for commodity 1, assign the maximum possible integer capacity for each edge (i,j) for commodity 2, and find the maximum flow. Continue the procedure until the maximum flows are found for all commodities.

Consider r_{ij}^{k} as a remaining capacity available to be assigned to commodity k on edge (i,j). Denote $\lceil x \rceil$ the largest integer $\leq x$ and IIF(a,b,c) the "if-then-else" function which returns b or c depending on whether a condition a is satisfied or not. Then we can formulate the following algorithm

ALLOCATION-PROCEDURE

$k := 1$;

$\forall (i,j)\in E$ **do** $r_{ij}^{k} := c_{ij}$;

while $k \leq |K|$ **do**

begin SINGLE-COMMODITY-PROBLEM(

k,

$$x_{ij}^{k} \in \left[0, IIF\left(w_{ij}^{k} = 0, c_{ij}^{k}, \min\left\{ c_{ij}^{k}, \left\lceil \frac{r_{ij}^{k}}{w_{ij}^{k}} \right\rceil \right\} \right) \right], \quad (i,j) \in E,$$

$f^{k'}$ ... maximal flow,

$x_{ij}^{k'}$ ... optimal integer flow along the edge (i,j)**)**;

$k := k+1$;

if $k \leq |K|$

then $\forall (i,j)\in E$ **do** $r_{ij}^{k} := r_{ij}^{k-1} - w_{ij}^{k} x_{ij}^{k'}$

end ;

The values $f^{k'}$, $x_{ij}^{k'}$ for all $(i,j)\in E$ and k = 1,2,...,$|K|$ give a feasible integer solution with a value $f' = \sum_{k\in K} f^{k'}$. The allocation of the capacity of edge (i,j) for commodity k is given by $w_{ij}^{k} x_{ij}^{k'}$ for k = 1,2,...,$|K|$–1 and $c_{ij} - \sum_{k=1}^{|K|-1} w_{ij}^{k} x_{ij}^{k'}$ for commodity $|K|$.

As to the single commodity flow problem, there are many algorithms for solving it, e.g., Goldberg & Tarjan's, Dinic's or Karzanov's algorithm (Ahuja et al 1993), their implementations are described in (Palubják 2003).

It is straightforward that the allocation of the edge capacities among the commodities and the corresponding combined maximal flow depend on the order in which the commodities are selected. The optimal ordering of the commodities in the allocation procedure is a difficult combinatorial problem. For $|K|$ commodities the number of their possible orderings is $|K|!$ because the number of all permutations of 1,2, ... ,$|K|$ is $|K|!$. Therefore, a search for the optimum in the space of permutations is only feasible for a not very high number of commodities.

This means that the time complexity of the problem in question is equal to that of the travelling salesman problem (Gutin and Punnen 2002). For higher numbers of commodities, we must use a heuristic or approximation method. Stochastic heuristics, mainly genetic algorithms, tabu-search and simulated annealing (Michalewicz 1996), (Michalewicz and Fogel 2000), (Reeves 1993), are among them the most popular methods.

The proposed algorithm was implemented using the simulated annealing (SA) technique. Its idea is based on simulating the cooling of a material in a heat bath - a process known as annealing. More details can be found in (Reeves 1993). For short, only note, that its performance is sensitive to the choice of a cooling schedule, i.e. it depends on the initial temperature, final temperature and temperature-reduction function.

Parameters of the simulated annealing were set as follows: Initial temperature: 1800, final temperature: 12 to 290, temperature reduction function $\alpha(t) = t/(1+a\,t)$ where $a = 0.000008$, neighbourhood – shift operation (two randomly selected positions are used in a different way, it removes a value at one position and puts it at another position. An example of this procedure is shown in Figure 1.). Each test was executed 30 times.

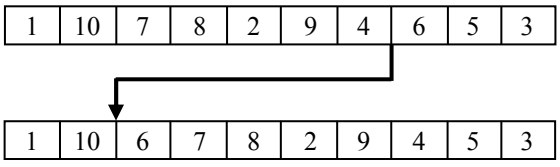


Figure 1. Shift operation for determining a permutation neighbour.

The proposed modules for solving the maximal multicommodity integer flow problem were implemented and tested for three types of networks N_1, N_2, N_3 with the following numbers of vertices and edges: (1) $|V|$=100, $|E|$=300, (2) $|V|$=100, $|E|$=600 and (3) $|V|$=100, $|E|$=900. Each network was generated for 5, 6, … , 11 commodities.

The results are summarized in Table 1.

Table 1: Computational results for 3 types of networks and 5 to11 commodities.

number of products (number of iterations)	network	opt.	the best solution	average solution
5 (120)	N_1	89868	89868	89852
	N_2	15767	15767	15767
	N_3	35143	35143	35143
6 (300)	N_1	95510	95510	95495
	N_2	17192	17192	17181
	N_3	44319	44319	44301
7 (2000)	N_1	97644	97644	97644
	N_2	17578	17578	17578
	N_3	47368	47368	47368
8 (4000)	N_1	96998	96998	96965
	N_2	19769	19769	19757
	N_3	53098	53098	53083
9 (5800)	N_1	87700	87700	87700
	N_2	18909	18909	18818
	N_3	54555	54555	54555
10 (8000)	N_1	90636	90636	90625
	N_2	15897	15897	15875
	N_3	40861	40861	40847
11 (10200)	N_1	105267	105267	104923
	N_2	17333	17333	17324
	N_3	55519	55519	55430

CONCLUSIONS

In this paper we studied network flows and their role for the modelling of other problems and presented for the Minimum Steiner Tree Problem. Besides it, we presented possible extensions of the basic maximum flow problem, e.g., for finding maximum flow with priorities of sources and sinks.

Finally, we studied the integer maximal multi-commodity flow problem, which belongs to the class of NP-hard combinatorial problems. In contrast to obvious approaches, mainly based on deterministic decomposition algorithms, we propose a stochastic heuristic approach using the simulated annealing algorithm.

Computational results show that, for suitable parameter settings presented in the paper, this approach is able to find the optimal solution almost in all cases or at least a solution very close to optimum when the test is executed several times. Furthermore, the proposed algorithm is stable because even the average results gained from all the executions are close to optimum.

In the future, we foresee further tests with other stochastic heuristic methods – genetic algorithms and tabu-search and their parameter settings because their temporary results are worse than the simulated annealing results.

Acknowledgments This research has been supported by the Czech Science Foundation GA ČR in the frame of GA ČR 102/09/1668 project Control Algorithm Design by Means of Evolutionary Approach and the Czech Ministry of Education in the frame of research plan MSM 0021630518 Simulation Modelling of Mechatronic Systems.

REFERENCES

Ahuja, R.K.; T.L. Magnanti; and J.B. Orlin. 1993. *Network Flows. Theory, Algorithms and Applications*. Prentice Hall, Englewood Cliffs, New Jersey.

Du, D.-Z.; J.M. Smith; and J.H. Rubinstein. 2000. *Advances in Steiner Trees*. Kluwer Academic Publishers, Dordrecht.

Gutin, G. and A.P. Punnen (eds.). 2002. *The Traveling Salesman Problem and Its Variations*. Kluwer Academic Publishers, Dordrecht.

Hochbaum, D. (ed.). 1996. *Approximation Algorithms for NP-Hard Problems*. PWS Publishing Company, Boston.

Hwang, F.K.; D.S. Richards, and P. Winter. 1992. *The Steiner Tree Problem*. North-Holland, Amsterdam.

Lin, Y.K. 2002. "Study on the System Capacity for a Multicommodity Stochastic-Flow Network with Node Failure". *Reliability Engineering and System Safety*, Vol. 78, 2002, 57-62.

Lozovanu, D. and M. Fonoberova. 2006. "Optimal Dynamic Multicommodity Flows in Networks". *Electronic Notes in Discrete Mathematics,* Vol. 25, 93-100.

Michalewicz, Z. and D.B. Fogel. 2000. *How to Solve It: Modern Heuristics*. Springer-Verlag, Berlin.

Michalewicz, Z. 1996. *Genetic Algorithms + Data Structures = Evolution Programs*. Springer Verlag, Berlin, Germany.

Mrad, M. and M. Haouari. 2008. "Optimal solution of the discrete cost multicommodity network design problem". *Applied Mathematics and Computation*, Vol. 204, No. 2, 745-753.

Ouorou, A. and P. Mahey. 2000. "A Minimum Mean Cycle Cancelling Method for Nonlinear Multicommodity Flow Problems". *European Journal of Operational Research*, Vol. 121, 532-548.

Ouorou, A.; P. Mahey; and J.-P. Vial. 2000. "A Survey of Algorithms for Convex Multicommodity Flow Problems". *Management Science*, Vol. 46, No. 1, 126-147.

Palubják, P. 2003. Multicommodity Network Flows (in Czech). PhD thesis, Brno University of Technology, Faculty of Mechanical Engineering, Brno, Czech Republic, 142 pp.

Plesník, J.. 1983. *Graph Algorithms* (in Slovak). Veda, Bratislava, Slovakia.

Reeves, C.R. 1993. *Modern Heuristic Techniques for Combinatorial Problems*. Blackwell Scientific Publications, Oxford.

Srivastav, A. and P. Stangier. 2000. "On Complexity, Representation and Approximation of Integral Multicommodity Flows". *Discrete Applied Mathematics*, Vol. 99, 183-208.

Wang, D and R. Kleinberg. 2009. "Analyzing Quadratic Unconstrained Binary Optimization Problems via Multicommodity Flows". *Discrete Applied Mathematics*, Vol. 157, No. 18, 3746-3753.

AUTHOR BIOGRAPHY

MILOŠ ŠEDA was born in Uherské Hradiště, Czech Republic. He graduated from Technical University of Brno in 1976 (majoring in Technical Cybernetics) and Masaryk University of Brno in 1985 (majoring in Computer Science). He received his Ph.D. degree in Technical Cybernetics in 1998, Assoc. Prof. degree in Applied Mathematics in 2001 and in 2009 he was appointed professor of Design and Process Engineering at Brno University of Technology. His work is concerned with graph theory, computational geometry, robotics, combinatorial optimisation, scheduling manufacturing processes, project management, fuzzy sets applications and database systems.

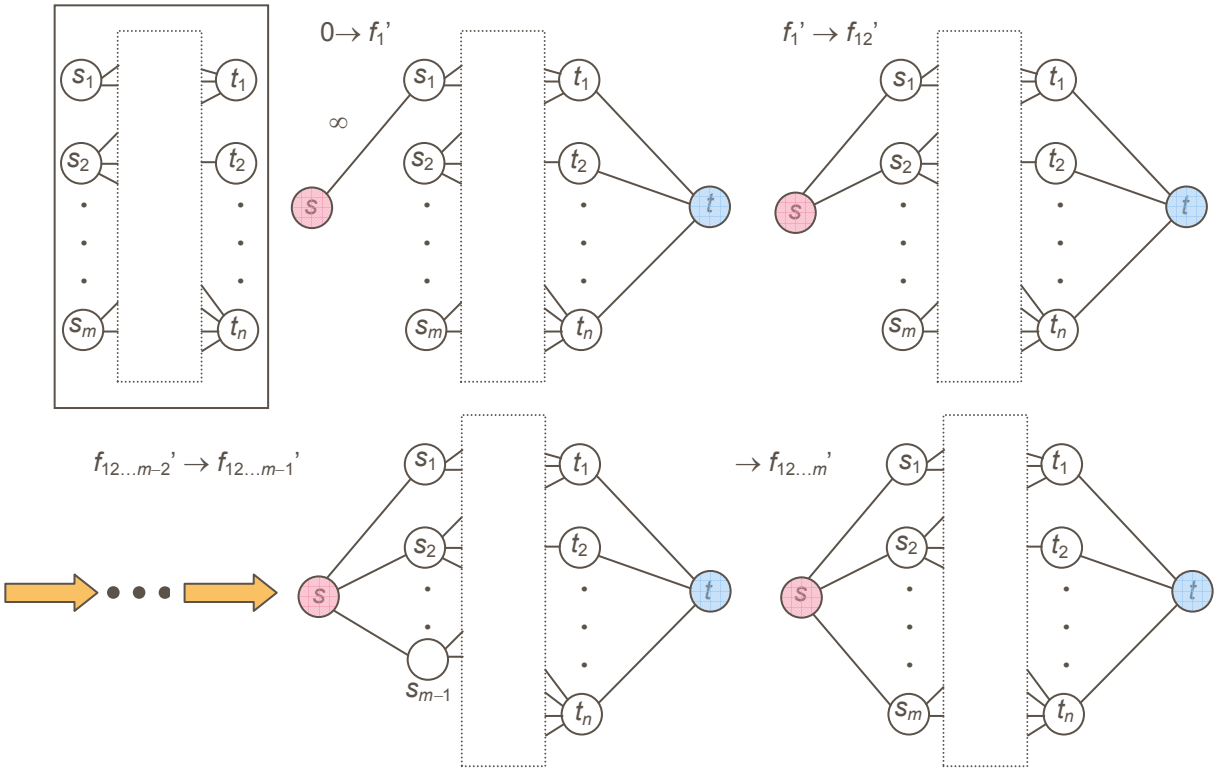


Figure 2. Maximal flow with priorities of sources and sinks

SYNTHESIS OF CONTROL LAW FOR CHAOTIC HENON SYSTEM PRELIMINARY STUDY

Zuzana Oplatkova, Roman Senkerik, Ivan Zelinka, Jiří Hološka
Faculty of Applied Informatics
Tomas Bata University in Zlin
Nad Stranemi 4511, 760 05 Zlin
Czech Republic
{Oplatkova, Senkerik, Zelinka}@fai.utb.cz

KEYWORDS
Control law, Henon system, synthesis, evolutionary computation, analytic programming.

ABSTRACT

The paper deals with a synthesis of control law for a discrete chaotic Henon system by means of analytic programming. This is a preliminary study in which the aim is to show that tool for symbolic regression – analytic programming - is possible to use for such kind of problems. The paper consists of description of analytic programming as well as chaotic Henon system. This article contents only 21 successful simulations in the result section and will be extended within future tests in this field. SOMA (Self-Organizing Migrating Algorithm) with analytic programming was used for experiments in this case.

INTRODUCTION

The interest about the control of chaotic systems is spread day by day. First steps were done in (Zelinka et al., 2006), (Zelinka et al., 2007), (Senkerik et al., 2006) where the control law was based on Pyragas method: Extended delay feedback control – ETDAS (Pyragas, 1995). That papers were concerned to tune several parameters inside the control technique for chaotic system. Compared to that, a presented paper shows a possibility how generate the whole control law (not only to optimize several parameters) for the purpose of stabilization of a chaotic system. The synthesis of control is inspired by the Pyragas's delayed feedback control technique (Just, 1999), (Pyragas, 1992). Unlike the original OGY control method (Ott, 1990) it can be simply considered as a targeting and stabilizing algorithm together in one package (Kwon, 1999). Another big advantage of Pyragas method is the amount of accessible control parameters.

Instead of evolutionary algorithms (EA) utilization, analytic programming (AP) is used here. AP is a superstructure of EAs and is used for synthesis of analytic solution according the required behavior. A control law from the proposed system can be viewed as a symbolic structure, which can be created according the requirements for the stabilization of a chaotic system. The advantage is that it is not necessary to have some "preliminary" control law and only to estimate its parameters. This system will generate the structure of the law also with suitable parameter values.

Firstly, a problem design is proposed. The next paragraph is focused on AP description. Results and conclusion follow afterwards.

PROBLEM DESIGN

The chosen example of chaotic system was the two dimensional Henon map in form (1).

$$\begin{aligned} x_{n+1} &= a - x_n^2 + b y_n \\ y_{n+1} &= x_n \end{aligned} \tag{1}$$

This is a model invented with a mathematical motivation to investigate chaos. The Henon map is a discrete-time dynamical system, which was introduced as a simplified model of the Poincaré map for the Lorenz system. It is one of the most studied examples of dynamical systems that exhibit chaotic behavior and in fact it is also a two-dimensional extension of the one-dimensional quadratic map. The map depends on two parameters, a and b, which for the canonical Henon map have values of $a = 1.4$ and $b = 0.3$. For the canonical values the Henon map is chaotic (Hilborn, 2000).

The example of this chaotic behavior can be clearly seen from bifurcation diagram – Figure 1.

This figure shows the bifurcation diagram for the Henon map created by plotting of a variable x as a function of the one control parameter for the fixed second parameter.

This work is focused on explanation of AP application for synthesis of a whole control law instead of demanding tuning of EDTAS method control law to stabilize desired Unstable Periodic Orbits (UPO). As a study case a p-1 (a fixed point) desired UPO is used only in this preliminary study. Until today, 21 successful simulations out of 21 have been carried out and the others are running.

EDTAS method was obviously an inspiration for preparation of sets of basic functions and operators for AP.

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

The original control method – ETDAS in the discrete form suitable for two-dimensional Henon map has the form (2).

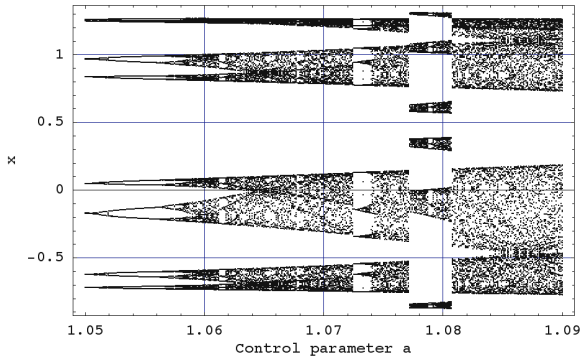


Figure 1: Bifurcation diagram of Henon Map

$$\begin{aligned} x_{n+1} &= a - x_n^2 + by_n + F_n \\ F_n &= K\left[(1-R)S_{n-m} - x_n\right] \\ S_n &= x_n + RS_{n-m} \end{aligned} \tag{2}$$

Where K and R are adjustable constants, F is the perturbation, S is given by a delay equation utilizing previous states of the system and m is the period of m-periodic orbit to be stabilized. The perturbation F_n in equations (2) may have arbitrarily large value, which can cause diverging of the system outside the interval {-1.5, 1.5}. Therefore, F_n should have a value between $-F_{\max}$, $F_{\max}$. In this preliminary study a suitable $F_{\max}$ value was taken from the previous research. To find the optimal value also for this parameter is in future plans.

COST FUNCTION FOR STABILIZATION TESTING

Proposal for the cost function comes from the simplest Cost Function (CF) presented in (Senkerik et al., 2008). The core of CF could be used only for the stabilization of p-1 orbit. The idea was to minimize the area created by the difference between the required state and the real system output on the whole simulation interval – τ_i.

But another cost function had to be used for stabilizing of higher periodic orbit. It was synthesized from the simple CF and other terms were added. In this case, it is not possible to use the simple rule of minimizing the area created by the difference between the required and actual state on the whole simulation interval – τ_i, due to the many serious reasons, for example: degrading of the possible best solution by phase shift of periodic orbit.

This CF is in general based on searching for desired stabilized periodic orbit and thereafter calculation of the difference between desired and found actual periodic orbit on the short time interval - τ_s (approx. 20 - 50 iterations) from the point, where the first min. value of difference between desired and actual system output is found. Such a design of CF should secure the successful stabilization of higher periodic orbit anywise phase shifted.

This CF can be also used for p-1 orbit. The CF_{Basic} has the form (3).

$$CF_{Basic} = penalization_1 + \sum_{t=\tau 1}^{\tau 2} \left| TS_t - AS_t \right| \tag{3}$$

where: TS - target state, AS - actual state

τ_1 - the first min. value of difference between TS and AS

τ_2 – the end of optimization interval ($\tau_1 + \tau_s$)

$penalization_1 = 0$ if $\tau_i - \tau_2 \geq \tau_s$;

$penalization_1 = 10*(\tau_i - \tau_2)$ if $\tau_i - \tau_2 < \tau_s$ (i.e. late stabilization)

ANALYTIC PROGRAMMING

Basic principles of the AP were developed in 2001 (Zelinka et al., 2005), (Zelinka et al., 2008), (Oplatkova et al., 2009). Until that time only genetic programming (GP) and grammatical evolution (GE) had existed. GP uses genetic algorithms while AP can be used with any evolutionary algorithm, independently on individual representation. To avoid any confusion, based on use of names according to the used algorithm, the name - Analytic Programming was chosen, since AP represents synthesis of analytical solution by means of evolutionary algorithms.

The core of AP is based on a special set of mathematical objects and operations. The set of mathematical objects is set of functions, operators and so-called terminals (as well as in GP), which are usually constants or independent variables. This set of variables is usually mixed together and consists of functions with different number of arguments. Because of a variability of the content of this set, it is called here "general functional set" – GFS. The structure of GFS is created by subsets of functions according to the number of their arguments. For example GFS_{all} is a set of all functions, operators and terminals, GFS_{3arg} is a subset containing functions with only three arguments, GFS_{0arg} represents only terminals, etc. The subset structure presence in GFS is vitally important for AP. It is used to avoid synthesis of pathological programs, i.e. programs containing functions without arguments, etc. The content of GFS is dependent only on the user. Various functions and terminals can be mixed together (Zelinka et al., 2005), (Zelinka et al., 2008), (Oplatkova et al., 2009).

The second part of the AP core is a sequence of mathematical operations, which are used for the program synthesis. These operations are used to transform an individual of a population into a suitable program. Mathematically stated, it is a mapping from an individual domain into a program domain. This mapping consists of two main parts. The first part is

called discrete set handling (DSH) (Figure 2) (Zelinka et al., 2005), (Lampinen & Zelinka, 1999) and the second one stands for security procedures which do not allow synthesizing pathological programs. The method of DSH, when used, allows handling arbitrary objects including nonnumeric objects like linguistic terms {hot, cold, dark…}, logic terms (True, False) or other user defined functions. In the AP DSH is used to map an individual into GFS and together with security procedures creates the above mentioned mapping which transforms arbitrary individual into a program.

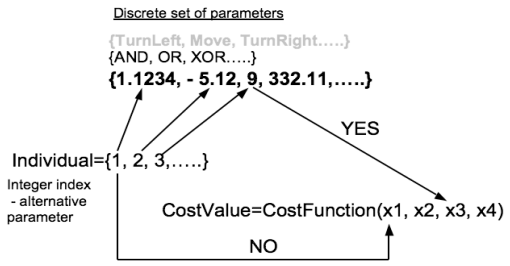


Figure 2: Discrete set handling

AP needs some evolutionary algorithm (Zelinka, 2004) that consists of population of individuals for its run. Individuals in the population consist of integer parameters, i.e. an individual is an integer index pointing into GFS. The creation of the program can be schematically observed in Figure 3. The individual contains numbers which are indices into GFS. The detailed description is represented in (Zelinka et al., 2005), (Zelinka et al., 2008), (Oplatkova et al., 2009).
AP exists in 3 versions – basic without constant estimation, AP_{nf} – estimation by means of nonlinear fitting package in Mathematica environment and AP_{meta} – constant estimation by means of another evolutionary algorithms; meta means metaevolution.

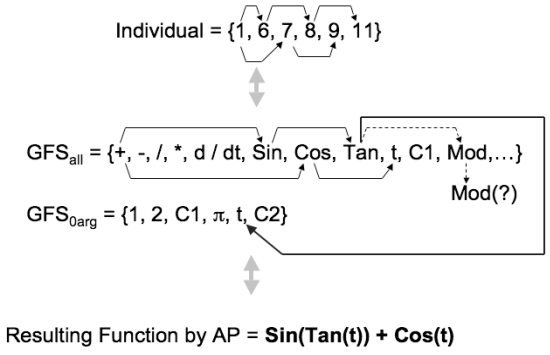


Figure 3: Main principles of AP

USED EVOLUTIONARY ALGORITHMS

Self Organizing Migrating Algorithm (SOMA) is a stochastic optimization algorithm that is modelled on the social behaviour of cooperating individuals (Zelinka, 2004). It was chosen because it has been proven that the algorithm has the ability to converge towards the global optimum (Zelinka, 2004). SOMA works on a population of candidate solutions in loops called *migration loops*. The population is initialized randomly distributed over the search space at the beginning of the search. In each loop, the population is evaluated and the solution with the highest fitness becomes the leader *L*. Apart from the leader, in one migration loop, all individuals will traverse the input space in the direction of the leader. Mutation, the random perturbation of individuals, is an important operation for evolutionary strategies (ES). It ensures the diversity amongst the individuals and it also provides the means to restore lost information in a population. Mutation is different in SOMA compared with other ES strategies. SOMA uses a parameter called PRT to achieve perturbation. This parameter has the same effect for SOMA as mutation has for genetic algorithms.
The novelty of this approach is that the PRT Vector is created before an individual starts its journey over the search space. The PRT Vector defines the final movement of an active individual in search space.
The randomly generated binary perturbation vector controls the allowed dimensions for an individual. If an element of the perturbation vector is set to zero, then the individual is not allowed to change its position in the corresponding dimension.
An individual will travel a certain distance (called the PathLength) towards the leader in *n* steps of defined length. If the PathLength is chosen to be greater than one, then the individual will overshoot the leader. This path is perturbed randomly.

RESULTS

As above mentioned, AP needs an evolutionary algorithm for its run. In this paper AP_{meta} version was used. It was easier to set all parameters than to use nonlinear fitting package, which was used with a big success in other cases.
SOMA algorithm (Zelinka, 2004) was used for both optimization tasks – to find a suitable solution of the control law and in metaevolution - to find suitable estimated values of constants in the obtained control law. Both settings were similar (Table 1 and Table 2).

Table 1: SOMA settings for AP

PathLength	3
Step	0.11
PRT	0.1
PopSize	50
Migrations	4
Max. CF Evaluations (CFE)	5345

Table 2: SOMA settings for meta evolution

PathLength	3
Step	0.11
PRT	0.1
PopSize	40
Migrations	5
Max. CF Evaluations (CFE)	5318

During all simulations, 21 successful results were obtained. A minimum number of cost function evaluations in the case of SOMA for AP was 100, maximum 3093. The average through all simulations was 1345. Simulations were stopped when CF was under 10^{-8}. These numbers were only for AP, if also second evolution for obtaining parameter values would be taken, each number has to be multiplied by 5318, i.e. total number of cost function evaluations was from 0.532 millions to 16.449 millions.
As was said the novelty of this approach represents the synthesis of feedback control law F_n (4) (perturbation) for the Henon system inspired by original ETDAS control method.

$$x_{n+1} = a - x_n^2 + by_n + F_n \qquad (4)$$

Following control laws are examples of obtained results in version without Ks estimated from AP and the notation with simplification after estimation by means of second SOMA. The first case was stored for further processing or better tuning.

a) without estimation

$$F_n = \frac{x_{n-1}x_n}{x_{n-1}(2x_n - x_{n-1}) - \frac{x_n + K_1}{x_{n-1} - x_n}}$$

with Ks estimation

$$F_n = \frac{x_{n-1}x_n}{x_{n-1}(2x_n - x_{n-1}) - \frac{x_n + 0.0349}{x_{n-1} - x_n}}$$

In this case, number 2 inside is not supposed as some K but simplification of original formula $x_{n-1} + x_{n-1}$
Stabilization was reached in 33th step.

b) without estimation

$$F_n = \frac{x_n(x_n - x_{n-1})}{K_1}$$

with Ks estimation

$$F_n = 0.9203 * x_n(x_n - x_{n-1})$$

The system was stabilized in 25th step.

c) without estimation

$$F_n = K_1(x_n - x_{n-1})$$

with Ks estimation

$$F_n = 0.76899 * (x_n - x_{n-1})$$

Stabilization was reached in minimal number of steps (from all simulations) – in 20th step. Simulation output of the stabilization is depicted in Figure 4.

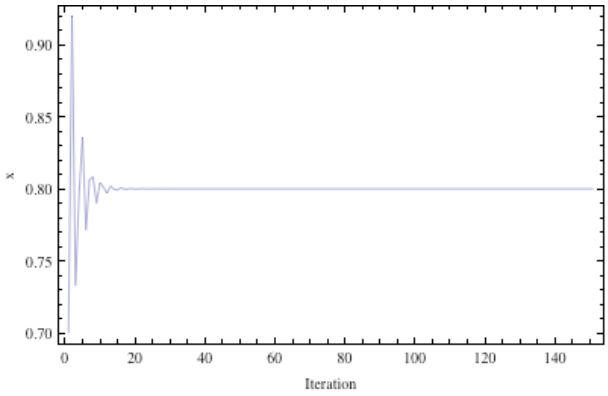


Figure 4: Example of result – stabilization of chaotic system by means of control law given in c)

d) without estimation

$$F_n = \frac{(x_{n-1} - x_n)K_1(-\frac{x_{n-1}}{K_2} - x_{n-1}x_n)}{x_n}$$

with Ks estimation

$$F_n = \frac{1.111(x_{n-1} - x_n)(0.011x_{n-1} - x_{n-1}x_n)}{x_n}$$

Stabilization was reached in maximal number of steps (from all simulations) - in 47th step. Simulation output of the stabilization is depicted in Figure 5.

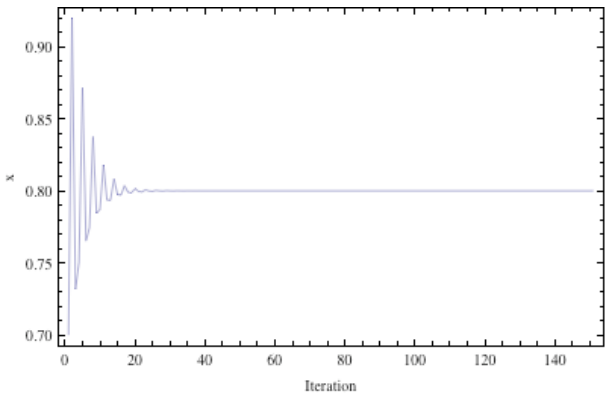


Figure 5: Example of result – stabilization of chaotic system by means of control law given in d)

e) without estimation

$$F_n = \frac{(x_{n-1} - x_n)}{x_{n-1}(x_n + K_1)}$$

with Ks estimation

$$F_n = \frac{(x_{n-1} - x_n)}{x_{n-1}(x_n - 2.2856)}$$

Stabilization was reached in maximal number of steps (from all simulations) - in 41th step. Simulation output of the stabilization is depicted in Figure 6.

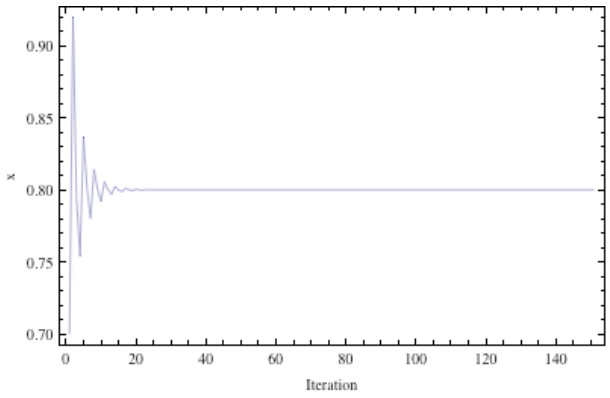


Figure 6: Example of result – stabilization of chaotic system by means of control law given in e)

Also an original ETDAS control law, which was tuned by means of evolutionary algorithms within (Senkerik et al., 2008) was found in the case of AP.
The quality of the solution was more or less the same. The cost function value is in order -15 or -16, i.e. almost zero. Thus the stabilization according to the cost value was very precise. The reached steps were counted as firstly observed required number rounded to 10^{-6}. According to this rule, the fastest stabilization was observed within 20 steps. The average was 34 steps.
The reached stabilization can be viewed also from the other side as a first minimal difference between reached and required solution. For the case with 20 steps it was reached 92 steps for successful stabilization according the second rule. The average amount of steps was 104.

CONCLUSION

This paper deals with a synthesis of a control law for stabilization of chaotic Henon system. In this case the analytic programming was used instead of tuning of parameters by means of evolutionary algorithms as in the previous research. AP is able to synthesize a symbolic notation for required behaviour of the system, EA tunes only parameters of expected law borrowed from ETDAS.

The presented results show that AP is able to solve problems of this kind and to produce the control law in a symbolic way. Within this preliminary study SOMA algorithm was used as an optimization algorithm for AP and also for estimating parameters in the second evolutionary process (meta-evolutionary approach).
Future plans are concerned to further tests to obtain more results for this chaotic system to reach a better statistics and also to use other chaotic systems. Further simulations will be considered also for higher orbit stabilization.

ACKNOWLEDGMENT

This work was supported by the grant NO. MSM 7088352101 of the Ministry of Education of the Czech Republic and by grants of Grant Agency of Czech Republic GACR 102/09/1680

REFERENCES

Hilborn R.C., 2000. Chaos and Nonlinear Dynamics: An Introduction for Scientists and Engineers, Oxford University Press, 2000, ISBN: 0-19-850723-2.

Just W., 1999, "Principles of Time Delayed Feedback Control", In: Schuster H.G., Handbook of Chaos Control, Wiley-Vch, ISBN 3-527-29436-8.

Kwon O. J., 1999. "Targeting and Stabilizing Chaotic Trajectories in the Standard Map", Physics Letters A. vol. 258, 1999, pp. 229-236.

Lampinen J., Zelinka I., 1999. *New Ideas in Optimization – Mechanical Engineering Design Optimization by Differential Devolution*, Volume 1. London: McGraw-hill, 1999, 20 p., ISBN 007-709506-5.

Oplatková, Z., Zelinka, I.: 2009. Investigation on Evolutionary Synthesis of Movement Commands, Modelling and Simulation in Engineering, Volume 2009 (2009), Article ID 845080, 12 pages, Hindawi Publishing Corporation, ISSN: 1687-559, e-ISSN: 1687-5605, doi:10.1155/2009/845080.

Ott E., C. Greboki, J.A. Yorke, 1990. "Controlling Chaos", Phys. Rev. Lett. vol. 64, 1990, pp. 1196-1199.

Pyragas K., 1992, "Continuous control of chaos by self-controlling feedback", Physics Letters A, 170, 421-428.

Pyragas K., 1995. "Control of chaos via extended delay feedback", Physics Letters A, vol. 206, 1995, pp. 323-330.

Senkerik R., Zelinka I., Navratil E., 2006. "Optimization of feedback control of chaos by evolutionary algorithms", in proc 1st IFAC Conference on analysis and control of chaotic systems, Reims, France, 2006.

Senkerik R., Zelinka I., Oplatkova Z., 2008. Evolutionary Techniques for Deterministic Chaos Control, CISSE'08, In Proc. IETA 2008, International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering, 5-13 December 2008, ISBN 978-90-481-3655-1.

Zelinka I., 2004. "SOMA – Self Organizing Migrating Algorithm", In: *New Optimization Techniques in Engineering*, (B.V. Babu, G. Onwubolu (eds)), chapter 7, 33, Springer-Verlag, 2004, ISBN 3-540-20167X.

Zelinka I.,Oplatkova Z, Nolle L., 2005. *Boolean Symmetry Function Synthesis by Means of Arbitrary Evolutionary Algorithms-Comparative Study*, International Journal of Simulation Systems, Science and Technology, Volume 6,

Number 9, August 2005, pages 44 - 56, ISSN: 1473-8031, online http://ducati.doc.ntu.ac.uk/uksim/journal/Vol-6/No.9/cover.htm, ISSN: 1473-804x.

Zelinka I., Senkerik R., Navratil E., 2006. "Investigation on Real Time Deterministic Chaos Control by Means of Evolutionary Algorithms", CHAOS'06, In Proc. 1st IFAC Conference on Analysis and Control of Chaotic Systems, Reims, France, 28-30 June 2006, pages 211-217

Zelinka I., Senkerik R., Navratil E., 2007. "Investigation on Evolutionary Optimitazion of Chaos Control", CHAOS, SOLITONS & FRACTALS (2007), doi:10.1016/j.chaos.2007.07.045.

Zelinka, I., Guanrong Ch., Celikovsky S., 2008. Chaos Synthesis by Means of Evolutionary algorithms, International Journal of Bifurcation and Chaos, Vol. 18, No. 4 (2008) 911–942.

AUTHORS BIOGRAPHIES

ZUZANA OPLATKOVA was born in Czech Republic, and went to the Tomas Bata University in Zlin, where she studied technical cybernetics and obtained her MSc. degree in 2003 and Ph.D. degree in 2008. She is a lecturer (Artificial Intelligence) at the same university. Her e-mail address is: oplatkova@fai.utb.cz

ROMAN SENKERIK was born in the Czech Republic, and went to the Tomas Bata University in Zlin, where he studied Technical Cybernetics and obtained his MSc degree in 2004 and Ph.D. degree in Technical Cybernetics in 2008. He is now a lecturer at the same university (Applied Informatics, Cryptology, Artificial Intelligence, Mathematical Informatics). Email address: senkerik@fai.utb.cz

IVAN ZELINKA was born in the Czech Republic, and went to the Technical University of Brno, where he studied Technical Cybernetics and obtained his degree in 1995. He obtained Ph.D. degree in Technical Cybernetics in 2001 at Tomas Bata University in Zlin. Now he is a professor (Artificial Intelligence, Theory of Information) and a head of the department. Email address: zelinka@fai.utb.cz .

MODELLING AND CONTROL OF HOT-AIR SYSTEM UNDER CONDITIONS OF UNCERTAINTY

Radek Matušů and Roman Prokop
Department of Automation and Control Engineering
Faculty of Applied Informatics
Tomas Bata University in Zlín
Nad Stráněmi 4511, 76005 Zlín, Czech Republic
E-mail: {rmatusu; prokop}@fai.utb.cz

KEYWORDS
Hot-Air System, Parametric Uncertainty, Algebraic Synthesis.

ABSTRACT

The contribution deals with modelling and control of temperature in laboratory model of hot-air system under conditions of parametric uncertainty. In the first instance, the first and second order parametrically uncertain mathematical models of the plant are constructed, and then they are utilized for design of various controllers with conventional structure. The control synthesis exploits general solutions of Diophantine equations in the ring of proper and stable rational functions. Robust stability of final closed control loops is tested using the value set concept and zero exclusion condition.

INTRODUCTION

The uncertainty represents serious problem in many real control applications. One of convenient approaches to uncertain modelling and description supposes no variations in the structure but only in parameters of the controlled system. In such case, one speaks about parametric uncertainty. In spite of the uncertain conditions, the often requirement consists in application of a cheap controller with simple PI or PID structure and fixed coefficients which would ensure stability and desired control behaviour for all expected values of the uncertain parameters.

A potential solution of this task consists in the usage of continuous-time controllers designed through general solutions of Diophantine equations in the ring of proper and stable rational functions (R_{PS}), Youla-Kučera parameterization and divisibility conditions. The principal idea of this approach is adopted from (Vidyasagar 1985; Kučera 1993) while the control design itself is proposed and analysed e.g. in (Prokop and Corriou 1997; Prokop et al. 2002; Matušů et al. 2008). This method brings a single tuning parameter $m > 0$ which can be used for influencing the control response. Later on, closed-loop robust stability can be verified for example with the assistance of the value set concept and zero exclusion condition (Barmish 1994; Bhattacharyya et al. 1995).

This paper aims to present a simple way of constructing a model with parametric uncertainty and also an algebraic approach to continous-time robust control design. The proposed techniques are applied during control of bulb temperature in laboratory model of hot-air tunnel. In a set of experiments, the controlled system is approximated by first or second order transfer functions with parametric uncertainty, the controllers are designed, the robust stability is verified, and the final control responses are tested and evaluated.

HOT-AIR PLANT DESCRIPTION

The controlled plant has been represented by laboratory model of hot-air tunnel constructed in VŠB – TU of Ostrava (Smutný et al. 2002). Generally, this object can be seen as multi-input multi-output (MIMO) system, however, the experiments have been done on a selected single-input single-output (SISO) loop. The model is composed of the bulb, primary and secondary ventilator and an array of sensors covered by tunnel. The bulb is powered by controllable source of voltage and serves as the source of light and heat energy while the purpose of ventilators is to ensure the flow of air inside the tunnel. All components are connected to the electronic circuits which adjust signals into the voltage levels suitable for CTRL 51 unit. Finally, this control unit is connected with the PC via serial link RS232. The diagram of the plant and the whole control system is shown in fig. 1.

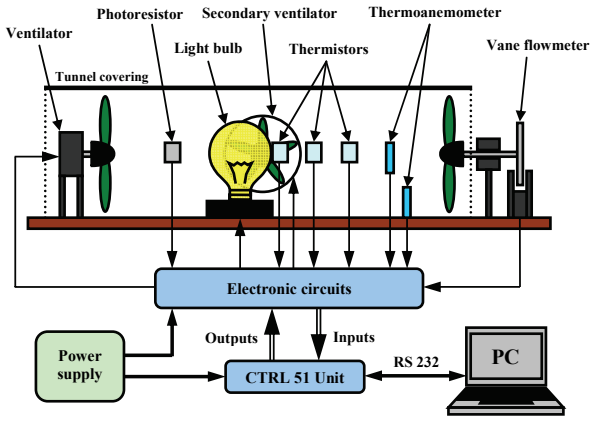


Figure 1: Scheme of Hot-Air Tunnel and Control System

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

The CTRL 51 unit has been produced in the Academy of Sciences of the Czech Republic (Klán et al. 2003). The tables 1 and 2 denote the meaning of input and output channels of this unit, respectively.

Table 1: Connection of Input Signals of CTRL 51 Unit

Input channel	Sensor
Input 1 ($y1$)	Light intensity of the bulb (photoresistor)
Input 2 ($y2$)	Temperature a few mm from the bulb (2nd thermistor)
Input 3 ($y3$)	Temperature of the bulb (1st thermistor)
Input 4 ($y4$)	Temperature at the end of the tunnel (3rd thermistor)
Input 6 ($y6$)	Airflow speed (thermoanemometer)
Input 7 ($y7$)	Airflow speed (vane flowmeter)

Table 2: Connection of Output Signals of CTRL 51 Unit

Output channel	Actuator
Output 1 ($u1$)	Bulb voltage (control of light intensity and bulb temperature)
Output 2 ($u2$)	Voltage of the primary ventilator (control of revolutions)
Output 3 ($u3$)	Voltage of the secondary ventilator (control of revolutions)

All presented identification and control experiments were performed in MATLAB 6.5.1 environment. The communication between MATLAB and CTRL 51 unit was arranged through four user functions (for initialization, reading and writing of data and for closing) and the synchronization of the program with real time was done via „semaphore“ principle (forthermore, the utilization of MATLAB functions „tic“ and „toc“ as an alternative were tested). To ensure the sufficient emulation of the continuous-time control algorithms, the sampling time 0.1 s was used. The detailed information about utilization of serial link under MATLAB including mentioned user routines, program synchronization mechanism and several tests can be found in (Dušek and Honc 2002). The discretization of control laws was carried out by left rectangle approximation method.

The considered loop covers bulb voltage $u1$ (control signal), which influences temperature of the bulb y3 (controlled variable). The other actuating signals were preseted to constant values – primary ventilator voltage $u2$ to 2 V and secondary one u3 to 0 V.

IDENTIFICATION OF THE SYSTEM

Naturally, the first task was to determine static and dynamic behaviour of the system. The trio of static characteristics measured during 3 different days is plotted in fig. 2.

Note, that the system properties markedly depend on current conditions and that it can be saturated in higher levels of $u1$. Therefore, the value 10 V was excluded from the subsequent process of identification. The fig. 3 shows the set of step responses with the starting point $u1 = 0\,\mathrm{V}$ while the final value of $u1$ is from 1 to 9 V and the fig. 4 depicts the similar responses from $u1 = 5\,\mathrm{V}$ to 6, 7, 8 and 9 V.

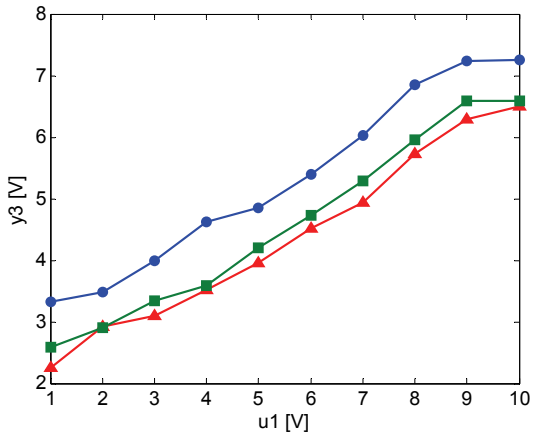


Figure 2: Static Characteristics of the System

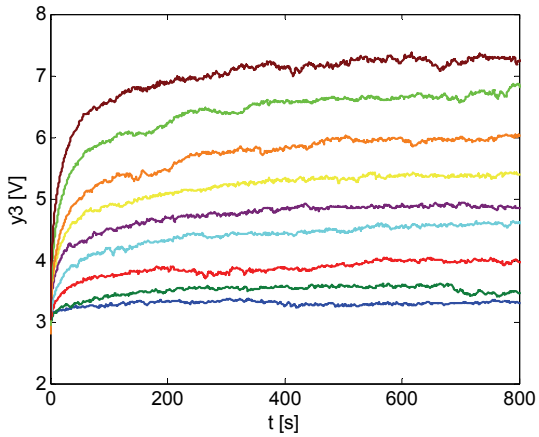


Figure 3: Step Responses Starting from $u1 = 0\,\mathrm{V}$

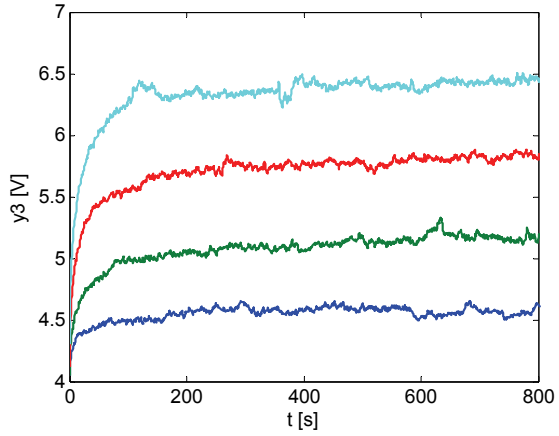


Figure 4: Step Responses Starting from $u1 = 5\,\mathrm{V}$

All measured responses were normalized and approximated by step response of system with selected structure. In the first instance, it has been approximated by first order system:

$$G(s) = \frac{K}{Ts+1} \tag{1}$$

However, with respect to the character of dynamics which is initially very fast and gradually starts to slow, the first order plant represents simplified solution. On that account, also the second order system:

$$G(s) = \frac{K(\tau s+1)}{(T_1 s+1)(T_2 s+1)} \tag{2}$$

has been assumed. The least squares method was used for identification of time constants. The example of approximation by first (1) and second order system (2) is given in fig. 5. It belongs to $u1$ step-change from 0 to 6 V. In this particular case, the transfer functions are:

$$G(s) = \frac{0.4107}{37.2289s+1} \tag{3}$$

$$G(s) = \frac{0.4107(119.7679s+1)}{(9.5334s+1)(186.3907s+1)} \tag{4}$$

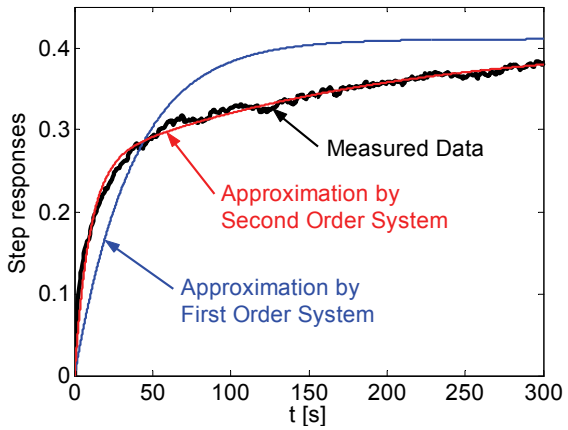


Figure 5: Example of Approximations

In an effort to stress more the initial part of responses with fast dynamics, only first 100 s have been included in optimization of T for first order. The complete results both for first and second order are shown in table 3.

Table 3: Identification Results

$u1$ [V]	K [-]	T [s]	τ [s]	T_1 [s]	T_2 [s]
0 - 1	0.2435	41.2259	27.1160	3.6437	72.1970
0 - 2	0.2833	47.2704	33.6512	3.2631	92.8624
0 - 3	0.3594	43.3580	115.9249	11.9881	186.0546
0 - 4	0.4274	48.2019	109.6962	10.8675	195.5283
0 - 5	0.3740	37.2064	76.1006	6.6347	130.6889
0 - 6	0.4107	37.2289	119.7679	9.5334	186.3907
0 - 7	0.4599	41.2868	119.2653	9.5686	194.5886
0 - 8	0.4889	37.9254	114.5048	9.5139	180.0520
0 - 9	0.4680	30.7657	93.7720	9.4876	137.8641
5 - 6	0.5656	35.1384	79.5117	4.9491	137.3248
5 - 7	0.5505	34.5624	90.2150	9.6484	139.7103
5 - 8	0.5676	28.0184	93.1630	8.1383	135.3549
5 - 9	0.5403	26.0924	94.5474	11.5862	123.6040

The set of data from these tables and advisement of substantive properties have led to the construction of models with parametric uncertainty. The lower bound of time constant T in model (5) has been moved down to 5 s because of fast initial dynamics which should also be taken into consideration. Although the intended working point corresponds to reference values of $y3$ at 4 and 5 V, the models are going to cover all measured temperature areas:

$$G(s,K,T) = \frac{[0.2;\,0.7]}{[5;\,50]s+1} \tag{5}$$

$$G(s,K,\tau,T_1,T_2) = \frac{[0.2;\,0.7]([25;\,130]s+1)}{([3;\,14]s+1)([70;\,210]s+1)} \tag{6}$$

ALGEBRAIC SYNTHESIS

The fractional approach developed by Vidyasagar (1985) and Kučera (1993) and discussed in (Prokop and Corriou 1997; Prokop et al. 2002) supposes that transfer functions of continuous-time linear causal systems in $\mathrm{R_{PS}}$ are expressed as:

$$G(s) = \frac{b(s)}{a(s)} = \frac{b(s)(s+m)^{-n}}{a(s)(s+m)^{-n}} = \frac{B(s)}{A(s)} \tag{7}$$

where $n = \max\{\deg(a), \deg(b)\}$ and $m > 0$.

Consider a two-degree-of-freedom (2DOF) control system from fig. 6. Take notice that the traditional one-degree-of-freedom (1DOF) system is obtained simply by $R = Q$.

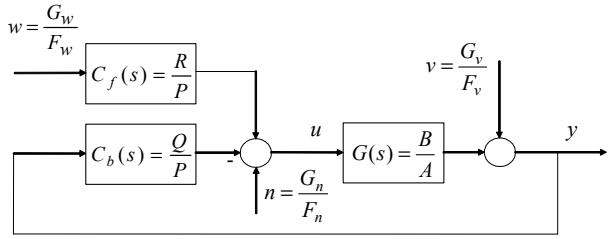


Figure 6: Two-Degree-of-Freedom Control System

External signals $w=\frac{G_w}{F_w}$, $n=\frac{G_n}{F_n}$ and $v=\frac{G_v}{F_v}$ represent the reference, load disturbance and disturbance signal, respectively. The most frequent case is a stepwise for reference and load disturbance signal and a harmonic signal for disturbance. Denominators of their transfer functions are then $F_w=F_n=\frac{s}{s+m}$ and $F_v=\frac{s^2+\omega^2}{(s+m)^2}$, respectively.

Basic relations following from fig. 6 are:

$$y=\frac{B}{A}u+v; \qquad u=\frac{R}{P}w-\frac{Q}{P}y+n \tag{8}$$

Further, the following equations hold:

$$y=\frac{BR}{AP+BQ}\frac{G_w}{F_w}+\frac{BP}{AP+BQ}\frac{G_n}{F_n}+\frac{AP}{AP+BQ}\frac{G_v}{F_v} \tag{9}$$

Provided that no disturbances affect the control system, i.e. $n=v=0$, the control error is given by:

$$e=w-y=\left(1-\frac{BR}{AP+BQ}\right)\frac{G_w}{F_w} \tag{10}$$

For the structure 1DOF ($R=Q$), the last relation takes the form:

$$e=\frac{AP}{AP+BQ}\frac{G_w}{F_w} \tag{11}$$

The basic task is to ensure stability of the system in fig. 6. All stabilizing feedback controllers are given by all solutions of the linear Diophantine equation:

$$AP+BQ=1 \tag{12}$$

with a general solution $P=P_0+BT$, $Q=Q_0-AT$, where T is free in R_{PS} and P_0, Q_0 is a pair of particular solutions (Youla – Kučera parameterization of all stabilizing controllers). Details and proofs can be found e.g. in (Prokop and Corriou 1997; Prokop et al. 2002). Then relations (10) and (11) take the form:

$$e=(1-BR)\frac{G_w}{F_w} \tag{13}$$

$$e=AP\frac{G_w}{F_w} \tag{14}$$

For asymptotic tracking then follows:

F_w must divide AP for 1DOF

F_w must divide $(1-BR)$ for 2DOF

The last condition gives the second Diophantine equation for 2DOF structure:

$$F_wS+BR=1 \tag{15}$$

The design process is demonstrated for first order system. A nominal transfer function is supposed as:

$$G(s)=\frac{b_0}{s+a_0} \tag{16}$$

Further, step-wise reference with $F_w=\frac{s}{s+m}$ and no disturbances are assumed. The Diophantine equation (12) takes the form:

$$\frac{s+a_0}{s+m}p_0+\frac{b_0}{s+m}q_0=1 \tag{17}$$

Multiplying by $(s+m)$ and comparing coefficients give the general stabilizing solution in the form:

$$P(s)=p_0+\frac{b_0}{s+m}T; \quad Q(s)=q_0-\frac{s+a_0}{s+m}T \tag{18}$$

where $q_0=\frac{m-a_0}{b_0}$; $p_0=1$ and T is free in R_{PS}. The asymptotic tracking for a stepwise reference w is given by divisibility of $F_w=\frac{s}{s+m}$ and AP (or only P in this case). It is achieved for $T=t_0=-\frac{m}{b_0}$ so that P has zero absolute coefficient in the numerator. Then inserting t_0 into (18) gives:

$$P(s)=\frac{s}{s+m}; \quad Q(s)=\frac{\tilde{q}_1 s+\tilde{q}_0}{s+m} \tag{19}$$

$$\tilde{q}_1=\frac{2m-a_0}{b_0}; \quad \tilde{q}_0=\frac{m^2}{b_0} \tag{20}$$

Finally, the 1DOF controller has the transfer function:

$$\frac{Q(s)}{P(s)}=\frac{\tilde{q}_1 s+\tilde{q}_0}{s} \tag{21}$$

Note that $\tilde{q}_1$, $\tilde{q}_0$ depend on single tuning parameter $m>0$. Hence, another topic of interest should be an appropriate choice of m. A potential way of nominal tuning can be found e.g. in (Matušů and Prokop 2008).

CONTROL EXPERIMENTS

First, the uncertain model (5) and nominal system (for controller design):

$$G_N(s)=\frac{0.5}{25s+1}=\frac{0.02}{s+0.04} \tag{22}$$

have been assumed. The tuning parameter $m = 0.0748$, which corresponds to 2 % of first overshoot from (Matušů and Prokop 2008), has been selected. The computed 1DOF PI controller (21), (20) is:

$$C_b(s) = \frac{\tilde{q}_1 s + \tilde{q}_0}{s} = \frac{5.48s + 0.2798}{s} \tag{23}$$

The characteristic polynomial of closed loop with plant (5) and controller (23) can be easily computed:

$$\begin{aligned} p(s,K,T) &= Ts^2 + (1 + K\tilde{q}_1)s + K\tilde{q}_0 = \\ &= [5;\,50]s^2 + [2.096;\,4.836]s + [0.05596;\,0.1959] \end{aligned} \tag{24}$$

This simple polynomial is obviously robustly stable, because it is of second order and all its coefficients are possitive, i.e. the whole system is robustly stable. The real closed-loop control behaviour can be seen in fig. 7. The control signal is depicted only in 25 % of its true size because of better perspicuity of controlled variable.

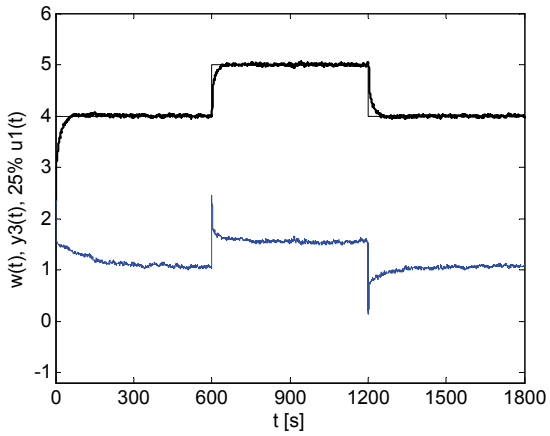


Figure 7: Control of Bulb Temperature by (23)

Next, it has been supposed the system with parametric uncertainty (6) and nominal plant:

$$G_N(s) = \frac{0.5(100s+1)}{(9s+1)(150s+1)} = \frac{0.0\overline{37}s + 0.00\overline{037}}{s^2 + 0.11\overline{7}s + 0.00\overline{074}} \tag{25}$$

Unfortunately, the single tuning parameter entails restraint for control design here and it is not easy to find suitable m with "quality" control response. Using synthesis technique from previous section, the chosen value $m = 0.025$ leads to 1DOF PID regulator:

$$\begin{aligned} C_b(s) &= \frac{\tilde{q}_2 s^2 + \tilde{q}_1 s + \tilde{q}_0}{s^2 + \tilde{p}_1 s} = \\ &= \frac{-1.1556s^2 + 0.01324s + 0.001055}{s^2 + 0.02502s} \end{aligned} \tag{26}$$

The plant (6) and controller (26) leads to closed-loop characteristic polynomial:

$$\begin{aligned} &p(s,K,\tau,T_1,T_2) = T_1T_2\left(s^4 + \tilde{p}_1 s^3\right) + \\ &+ (T_1 + T_2)\left(s^3 + \tilde{p}_1 s^2\right) + K\tau\left(\tilde{q}_2 s^3 + \tilde{q}_1 s^2 + \tilde{q}_0 s\right) + \\ &+ K\left(\tilde{q}_2 s^2 + \tilde{q}_1 s + \tilde{q}_0\right) + \left(s^2 + \tilde{p}_1 s\right) \end{aligned} \tag{27}$$

The value sets of this family with multilinear uncertainty structure, plotted via the Polynomial Toolbox for Matlab (Polyx; Šebek et al. 2000) for several non-negative frequencies, are depicted in fig. 8. Unluckily, the zero exclusion condition indicates that polynomial (27) and thus also the whole control system is not robustly stable for assumed range of uncertain parameters, because the value sets include the zero point. The boundaries in (6) are too broad (the requirements are too strong). Margins have to be narrowed to gain the closed loop robustly stable with the controller (26). For details about this very universal and effective technique for graphical testing of robust stability and related topics see e.g. (Barmish 1994; Bhattacharyya et al. 1995). However, the real system has been stable in used working point (with non-minimum phase behaviour), as can be seen in fig. 9.

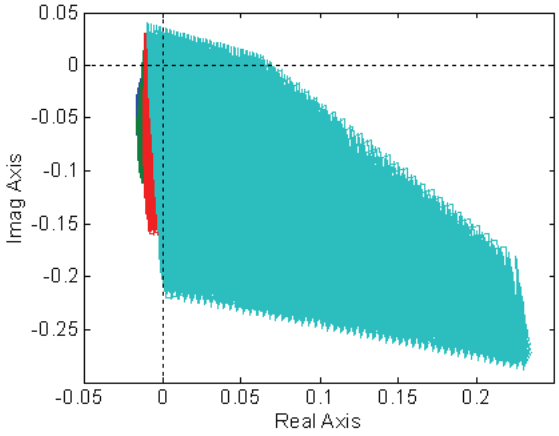


Figure 8: Value Sets for (27) with Parameters (26)

If 2DOF structure and $m = 0.02$ is used, the final controller arises in the form:

$$\begin{aligned} C_b(s) &= \frac{-1.3701s^2 + 0.01727s + 0.000432}{s^2 + 0.01297s} \\ C_f(s) &= \frac{1.08s^2 + 0.0432s + 0.000432}{s^2 + 0.01297s} \end{aligned} \tag{28}$$

The feedforward part does not influence robust stability, i.e. this controller would ensure it under similar conditions as in the previous case. The fig. 10 presents the control results.

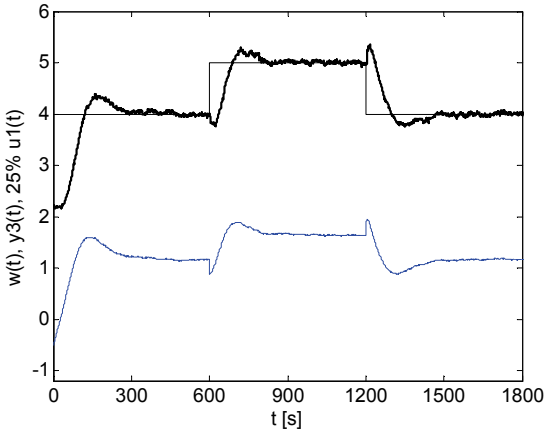


Figure 9: Control of Bulb Temperature by (26)

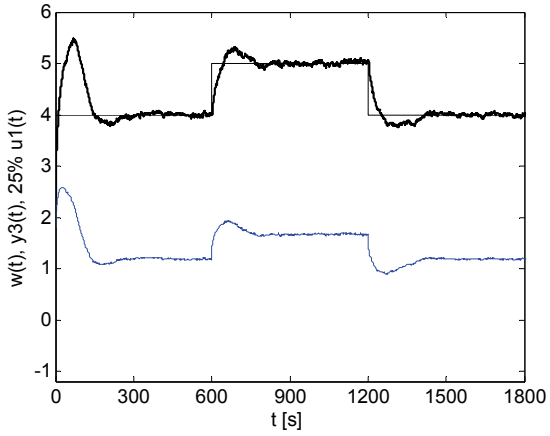


Figure 10: Control of Bulb Temperature by (28)

Another possibility of simplifying (instead of approximation by first order model) can be done via additional order reduction in identified second order system. The order reductions first only in numerator and afterward both in numerator and denominator lead to nominal transfer functions, respectively:

$$G_N(s) = \frac{0.00\overline{037}}{s^2 + 0.11\overline{7}s + 0.00\overline{074}} = \\ = \frac{0.5}{(9s+1)(150s+1)} \approx \frac{0.5(100s+1)}{(9s+1)(150s+1)} \tag{29}$$

$$G_N(s) = \frac{0.003145}{s+0.006289} = \frac{0.5}{159s+1} \approx \frac{0.5(100s+1)}{(9s+1)(150s+1)} \tag{30}$$

The former approximation (29) and $m = 0.04$ result in:

$$C_b(s) = \frac{10.4933s^2 + 0.6068s + 0.006912}{s^2 + 0.04222s} \tag{31}$$

while the latter one (30) and $m = 0.0204$ (6 % first overshoot for nominal system) lead to the controller:

$$C_b(s) = \frac{10.9733s + 0.1323}{s} \tag{32}$$

The plant (6) and the controller (31) give, again, the closed-loop characteristic polynomial with structure (27).

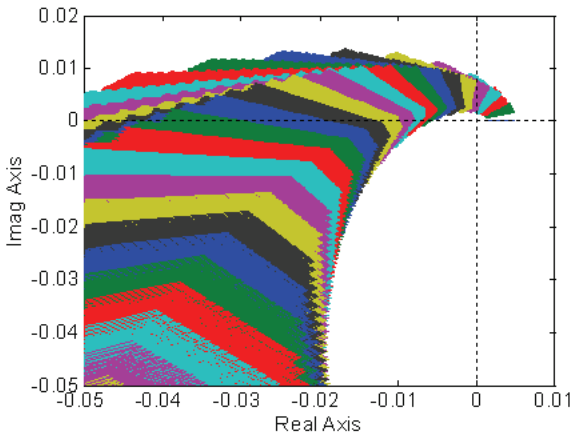


Figure 11: Value Sets for (27) with Parameters (31)

Nevertheless, in this instance, the closed-loop polynomial (27) is robustly stable which results from the fact that the origin of the complex plane is not included in the value sets (see Fig. 11) and (27) has a stable member (Barmish 1994). Thus the whole control system is robustly stable. Furthermore, the controlled system (6) and the regulator (32) yield the polynomial:

$$p(s,K,\tau,T_1,T_2) = T_1T_2s^3 + (T_1+T_2)s^2 + \\ + K\tau(\tilde{q}_1s^2 + \tilde{q}_0s) + K(\tilde{q}_1s + \tilde{q}_0) + s \tag{33}$$

which is also robustly stable as follows from Fig. 12. The figs. 13 and 14 show the control responses for both cases.

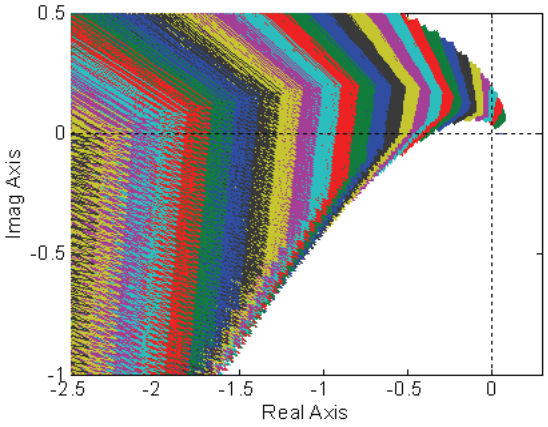


Figure 12: Value Sets for (33) with Parameters (32)

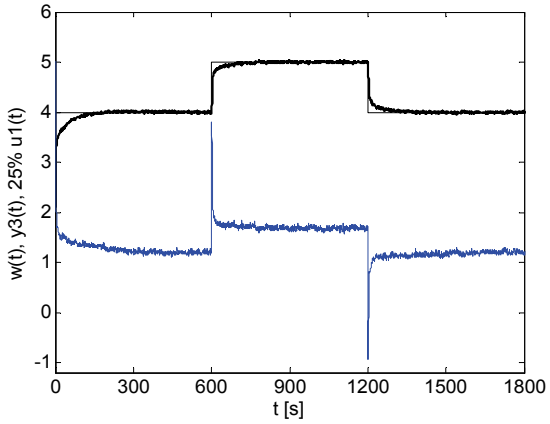


Figure 13: Control of Bulb Temperature by (31)

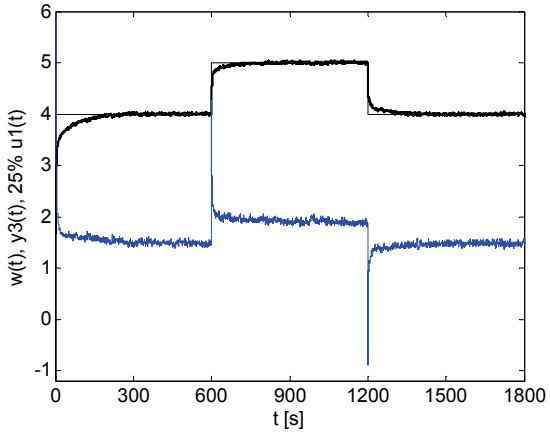


Figure 14: Control of Bulb Temperature by (32)

EVALUATION AND CONCLUSION

The objective evaluation of quality has been performed by meaning of Integrated Squared Error (ISE) criterion. The quantification is expressed in table 4.

Table 4: Outcomes of ISE Calculations

Controller	ISE
(23)	19.2579
(26)	370.7898
(28)	166.0868
(31)	18.8931
(32)	28.4738

The controllers (31) and (23) achieve the best ISE results. However, the regulator (23) generates "less aggressive" actuating signal after step-changes of reference. On the contrary, controller (26) is the worst, moreover with non-minimum phase behaviour. Not an application of 2DOF structure (28) brings about considerable improvement. Problems in control, which paradoxically emerge during use of identified second order model (6) and nominal system (25) are caused by synthesis limitation using single tuning parameter. It arises here as the cost for tuning simplicity.

ACKNOWLEDGEMENT

The work was supported by the Ministry of Education, Youth and Sports of the Czech Republic under Research Plan No. MSM 7088352102. This assistance is very gratefully acknowledged.

REFERENCES

Barmish, B.R. 1994. *New Tools for Robustness of Linear Systems*. Macmillan, New York, USA.

Bhattacharyya, S.P.; H. Chapellat and L.H. Keel. 1995. *Robust Control: The Parametric Approach*. Prentice Hall, Englewood Cliffs, New Jersey, USA.

Dušek, F. and D. Honc. 2002. "Utilization of serial link under MATLAB 6 (Využití sériové linky pod MATLABem verze 6)". In *Proceedings of conference MATLAB 2002* (Prague, Czech Republic). (In Czech).

Klán, P.; D. Honc and J. Jindřich. 2003. "New measuring unit CTRL V3 (Nová měřicí jednotka CTRL V3)". In *Proceedings of conference MATLAB 2003* (Prague, Czech Republic). (In Czech).

Kučera, V. 1993. "Diophantine equations in control - A survey". *Automatica*, Vol. 29, No. 6, 1361-1375.

Matušů, R. and R. Prokop. 2008. "Single-Parameter Tuning of PI Controllers: From Theory to Practice". In *Proceedings of the 17th IFAC World Congress* (Seoul, Korea).

Matušů, R.; R. Prokop and M. Dlapa. 2008. "Robust Control of Temperature in Hot-Air Tunnel". In *Proceedings of the 16th Mediterranean Conference on Control and Automation* (Ajaccio, France).

Polyx. "The Polynomial Toolbox". [online]. [cit. 01-02-2010]. Available from URL: <http://www.polyx.com/>.

Prokop, R. and J.P. Corriou. 1997. "Design and analysis of simple robust controllers". *International Journal of Control*, Vol. 66, No. 6, 905-921.

Prokop, R.; P. Husták and Z. Prokopová. 2002. "Simple robust controllers: Design, tuning and analysis". In *Proceedings of the 15th IFAC World Congress* (Barcelona, Spain).

Šebek, M.; M. Hromčík and J. Ježek. 2000. "Polynomial toolbox 2.5 and systems with parametric uncertainties". In *Proceedings of the 3rd IFAC Symposium on Robust Control Design* (Prague, Czech Republic).

Smutný, L.; J. Škuta and J. Farník. 2002. "Model of hot-air circuit (Model teplovzdušného obvodu)". Technical report to HS 311107 "Technická pomoc při návrhu a zhotovení modelu teplovzdušného obvodu", VŠB-TU Ostrava, Czech Republic. (In Czech).

Vidyasagar, M. 1985. *Control System Synthesis: A Factorization Approach*. MIT Press, Cambridge, MA, USA.

AUTHOR BIOGRAPHIES

RADEK MATUŠŮ was born in Zlín, Czech Republic in 1978. He is a Researcher at Faculty of Applied Informatics of Tomas Bata University in Zlín. He graduated from Faculty of Technology of the same university with an MSc in Automation and Control Engineering in 2002 and he received a PhD in Technical Cybernetics from Faculty of Applied Informatics in 2007. He worked as a Lecturer from 2004 to 2006. The main fields of his professional interest include robust systems and application of algebraic methods to control design. His e-mail address is: `rmatusu@fai.utb.cz` and his web-page can be found at: `http://zamestnanci.fai.utb.cz/~matusu/`.

ROMAN PROKOP was born in Hodonín, Czech Republic in 1952. He graduated in Cybernetics from the Czech Technical University in Prague in 1976. He received post graduate diploma in 1983 from the Slovak Technical University. Since 1995 he has been at Tomas Bata University in Zlín, where he presently holds the position of full professor of the Department of Automation and Control Engineering and a vice-dean of the Faculty of Applied Informatics. His research activities include algebraic methods in control theory, robust and adaptive control, autotuning and optimization techniques. His e-mail address is: `prokop@fai.utb.cz`.

RELAY FEEDBACK AUTOTUNING – A POLYNOMIAL DESIGN APPROACH

Roman Prokop, Jiří Korbel and Zdenka Prokopová
Faculty of applied informatics
Tomas Bata University in Zlín
Nad Stráněmi 4511, 760 05 Zlín, Czech Republic
E-mail: prokop@fai.utb.cz

KEYWORDS
Auto-tuning, Diophantine equation, RPS synthesis, PID controller, relay feedback.

ABSTRACT

A method of autotuning using an asymmetric relay with hysteresis feedback test is proposed and developed. Then, three parameters for aperiodic first or second order transfer functions can be obtained. After the identification relay experiment, controller parameters are computed through linear diophantine equation in the ring of proper and stable rational functions. This algebraic approach for a traditional 1-DOF feedback structure generates a class of PI or PID controllers. The pole placement principle in the methodology brings a scalar positive "tuning knob" for additional controller tuning. A Matlab-Simulink program implementation was developed for simulation and verification of the studied approach. Two illustrative examples support simplicity and efficiency of the proposed methodology.

INTRODUCTION

Despite of expressive evolution of control hardware, PID controllers remain the main tool in industrial feedback loops and they survived many changes in technology. The practical advantages of PID controllers can be seen in a simple structure, in an understandable principle and in control capabilities. It is widely known that PID controllers are quite resistant to changes in the controlled process without meaningful deterioration of the loop behavior.

However, real industrial processes suffer from many unpleasant drawbacks as non-linearity, complexity and time variance. These features induce difficulties to control their loops properly. Adequate and sufficient tuning of controllers needs to know relevant process parameters. One way how to overcome the mentioned problems consists in automatic tuning of controllers. The development of various autotuning principles was started by a simple symmetrical relay feedback experiment (Aström and Hägglund 1984) for a PID structure. Ultimate gain and ultimate frequency are then used for adjusting of parameters by common known Ziegler-Nichols rules. During the period of more than two decades, many studies have been reported to extend and improve autotuners principles; see e.g. (Aström and Hägglund 1995; Ingimundarson and Hägglund 2000; Majhi and Atherton 1998; Morilla at al. 2000). The extension in an experimental phase was performed in (Yu 1999; Pecharromán and Pagola 2000; Kaya and Atherton 2001) by an asymmetry and hysteresis of a relay, see (Thyagarajan and Yu 2002), and experiments with asymmetrical and dead-zone relay feedback are reported in (Vítečková and Víteček 2004; Vyhlídal 2000). Also, various control design principles and rules can be investigated in mentioned references. Nowadays, almost all commercial industrial PID controllers provide the feature of autotuning. This paper is focused on a novel combination for autotunig method of PI and PID controllers. The method combines an asymmetrical relay identification experiment and a control design which is based on a pole-placement principle. The pole placement problem is formulated through a Diophantine equation and it is tuned by an equalization setting proposed in (Gorez and Klán 2000).

PROCESS PARAMETERS IDENTIFICATION

System identification of the process parameters is a crucial point for many auto-tuning principles. The identification rules with relay in feedback loops can utilize various types of relays. The relay feedback proposed by Aström in 1984 used a symmetrical relay without hysteresis. The identification procedure with a relay in the feedback loop can utilize various types of relays. The relay feedback proposed by Aström in 1984 utilizes symmetrical relay without hysteresis. This procedure gives the ultimate parameters of the process and control design may follow. Unfortunately, the process gain must be known in advance because the symmetrical relay test cannot identify it. On the other hand, the process gain can be obtained during the relay feedback test when an asymmetrical relay is utilized. A typical data response from the relay experiment can be seen in Figure 1.
In this paper, an asymmetric relay with hysteresis is used. It enables to identify the parameters of the transfer function, such as proportional gain, time constant, as well as time delay term. Time delay is approximated by Pade approximation before the algebraic controller design.

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

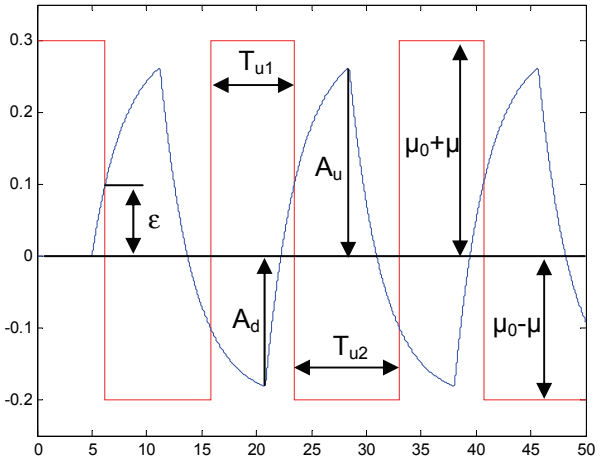


Figure 1: Relay feedback test of stable process

First order system

The most popular and simplest approximation of aperiodic industrial processes can be characterized by the first order transfer function with time delay (FOPDT) in the form:

$$G(s) = \frac{K}{Ts+1} \cdot e^{-Ls} \tag{1}$$

When the asymmetric relay is used for the relay feedback test, it is shown in Figure 1, the output y converges to the stationary oscillation in one period. These oscillations are characterized by equations (Hang et al. 2001):

$$A_u = (\mu_0 + \mu) \cdot K \cdot \left(1 - e^{-\frac{L}{T}}\right) + \varepsilon \cdot e^{-\frac{L}{T}} \tag{2}$$

$$A_d = (\mu_0 - \mu) \cdot K \cdot \left(1 - e^{-\frac{L}{T}}\right) - \varepsilon \cdot e^{-\frac{L}{T}} \tag{3}$$

$$T_{u1} = T \cdot \ln \frac{2 \cdot \mu \cdot K \cdot e^{\frac{L}{T}} + \mu_0 \cdot K - \mu \cdot K + \varepsilon}{\mu \cdot K + \mu_0 \cdot K - \varepsilon} \tag{4}$$

$$T_{u2} = T \cdot \ln \frac{2 \cdot \mu \cdot K \cdot e^{\frac{L}{T}} - \mu_0 \cdot K - \mu \cdot K + \varepsilon}{\mu \cdot K - \mu_0 \cdot K - \varepsilon} \tag{5}$$

The proportional gain can be alternatively computed from the equation (Vyhlídal 2000):

$$K = \frac{\int_0^{T_y} y(t)dt}{\int_0^{T_y} u(t)dt}, \quad T_y = T_{u1} + T_{u2} \tag{6}$$

The normalized dead time of the process (L/T) is obtained from (2) or (3) in the form (Hang et al. 2001):

$$\Theta = \ln \frac{(\mu_0 + \mu) \cdot K - \varepsilon}{(\mu_0 + \mu) \cdot K - A_u} \tag{7}$$

or

$$\Theta = \ln \frac{(\mu - \mu_0) \cdot K - \varepsilon}{(\mu - \mu_0) \cdot K + A_d} \tag{8}$$

Next, the time constant can be computer from (4) or (5) by solving these formulas:

$$T = T_{u1} \cdot \left(\ln \frac{2 \cdot \mu \cdot K \cdot e^{\Theta} + \mu_0 \cdot K - \mu \cdot K + \varepsilon}{\mu \cdot K + \mu_0 \cdot K - \varepsilon} \right)^{-1} \tag{9}$$

or

$$T = T_{u2} \cdot \left(\ln \frac{2 \cdot \mu \cdot K \cdot e^{\Theta} - \mu_0 \cdot K - \mu \cdot K + \varepsilon}{\mu \cdot K - \mu_0 \cdot K - \varepsilon} \right)^{-1} \tag{10}$$

The dead time is $L = T \cdot \Theta$.

Second order system

The process is described by the second order transfer function with time delay (SOPDT):

$$G(s) = \frac{K}{(Ts+1)^2} \cdot e^{-Ls} \tag{11}$$

The process gain can be computed by the same equation (6) as for the first order system.

The time constant and time delay term can be estimated by relations (Vítečková and Víteček 2004)

$$T = \frac{T_y}{2\pi} \cdot \sqrt{\frac{4 \cdot K \cdot u_0}{\pi \cdot a_y} - 1} \tag{12}$$

$$L = \frac{T_y}{2\pi} \cdot \left[\pi - 2 \arctan \frac{2\pi T}{T_y} - \arctan \frac{\varepsilon}{\sqrt{a_y^2 - \varepsilon^2}} \right] \tag{13}$$

where $T_y = T_{u1} + T_{u2}$, $u_0 = \mu$, $a_y = \frac{A_u + A_d}{2}$.

CONTROLLER DESIGN

The control design is based on the fractional approach (Vidyasagar 1987; Kučera 1993; Prokop and Corriou 1997; Prokop et al. 2002). Any transfer function G(s) of a (continuous-time) linear system is expressed as a ratio of two elements of R_{PS}. The set R_{PS} means the ring of (Hurwitz) stable and proper rational functions. Traditional transfer functions as a ratio of two polynomials can be easily transformed into the fractional form simply by dividing, both the polynomial denominator and numerator by the same stable polynomial of the order of the original denominator.

Then all transfer functions can be expressed by the ratio:

$$G(s) = \frac{b(s)}{a(s)} = \frac{\frac{b(s)}{(s+m)^n}}{\frac{a(s)}{(s+m)^n}} = \frac{B(s)}{A(s)} \tag{14}$$

$$n = \max(\deg(a), \deg(b)), \quad m > 0 \tag{15}$$

All feedback stabilizing controllers according to Figure 2 are given by a general solution of the Diophantine equation:

$$AP + BQ = 1 \tag{16}$$

which can be expressed with Z free in R_{PS}:

$$\frac{Q}{P} = \frac{Q_0 - AZ}{P_0 + BZ} \tag{17}$$

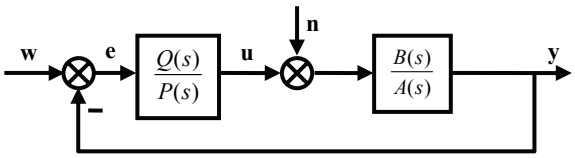


Figure 2: Feedback (1DOF) control loop

Asymptotic tracking is then ensured by the divisibility of the denominator P in (17) by the denominator of the reference $w = \frac{G}{F}$ which is supposed in the form:

$$F_w = \frac{s}{s+m}; \quad m > 0 \tag{18}$$

The set of reference signals with denominator (18) represents all stepwise signals which are most frequent references. The divisibility is achieved by a suitable choice of rational function Z in (17), see (Prokop et al. 2002).

The control design for first order systems (1) without time delay gives the Diophantine equation (16) in the form:

$$(Ts+1)p_0 + Kq_0 = s + m \tag{19}$$

and the general solution is given by:

$$\begin{aligned} P &= \frac{1}{T} + \frac{K}{s+m} \cdot Z \\ Q &= \frac{Tm-1}{TK} - \frac{Ts+1}{s+m} \cdot Z \end{aligned} \tag{20}$$

where Z is free in the ring R_{PS}. Asymptotic tracking is achieved by the choice:

$$Z = -\frac{m}{TK} \tag{21}$$

and the resulting PI controller is in the form:

$$C(s) = \frac{Q}{P} = \frac{q_1 s + q_0}{s} \tag{22}$$

where parameters q_1 a q_0 are given by:

$$q_1 = \frac{2Tm-1}{K} \qquad q_0 = \frac{Tm^2}{K} \tag{23}$$

Second order systems give the design equation in the form:

$$(Ts+1)^2 \cdot s \cdot p_0 + K \cdot (q_2 s^2 + q_1 s + q_0) = (s+m)^3 \tag{24}$$

and after similar manipulations the resulting PID controller gives the transfer function:

$$C(s) = \frac{Q}{P} = \frac{q_2 s^2 + q_1 s + q_0}{s(s+p_0)} \tag{25}$$

with parameters:

$$\begin{aligned} p_0 &= \frac{1}{T^2}; \qquad q_2 = \frac{3Tm-2}{KT} \\ q_1 &= \frac{3T^2m^2-1}{KT^2}; \qquad q_0 = \frac{m^3}{K} \end{aligned} \tag{26}$$

The scalar parameter *m>0* seems to be a suitable „tuning knob" influencing control behavior as well as robustness properties of the closed loop system. Some principles and approaches exist for a "good" or "optimal" choice of this parameter. One of them is based on the equalization principle, proposed in Gorez and Klán 2000.

Polynomial control synthesis is based on neglecting of time delay terms. However, ignoring of this term generally can not be acceptable, especially for its higher values. Then a Pade approximation can be used in the form

$$e^{-Ls} \approx \frac{1 - \frac{Ls}{2}}{1 + \frac{Ls}{2}} \tag{27}$$

and naturally, control equation (16) leads to higher orders of control transfer functions. For automatic control design, a program system in Matlab and Simulink with the support of the Polynomial Toolbox

was developed. Controller formulas then are more complex than (23) and (26) and resulting controllers lose PI or PID structures. Figure 3 illustrates the main menu of the program system.

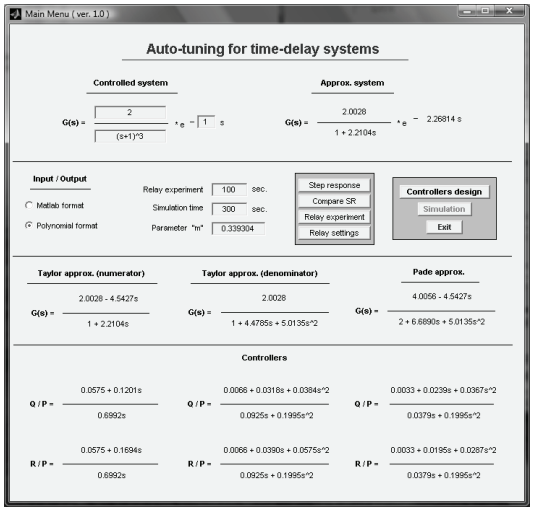


Figure 3: Main menu of the program

SIMULATION EXAMPLES

Example 1: A third order controlled system with the transfer function:

$$G(s) = \frac{2}{(2s+1.1)^3} \cdot e^{-2s} \tag{28}$$

was identified in the relay feedback loop as a first order model:

$$\widetilde{G}(s) = \frac{1.5}{4.5s+1} \cdot e^{-3.8s} \tag{29}$$

Step responses of the controlled system and its identified model are compared in Figure 4.

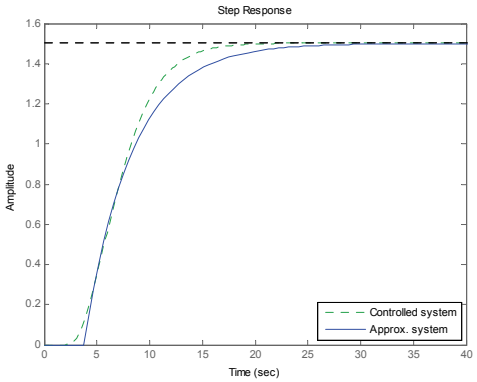


Figure 4: Step responses of systems

The following controller design was performed for two different tuning parameters *m*. Control responses can be seen in Figure 5 and Figure 6.

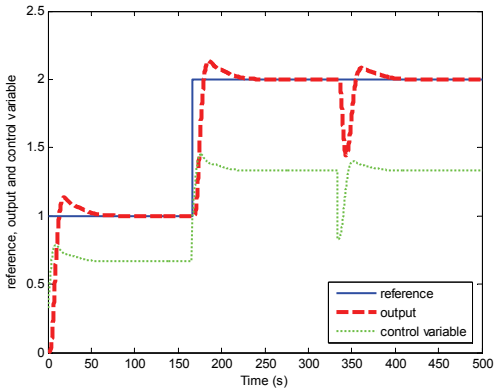


Figure 5: Control response for m=0.17

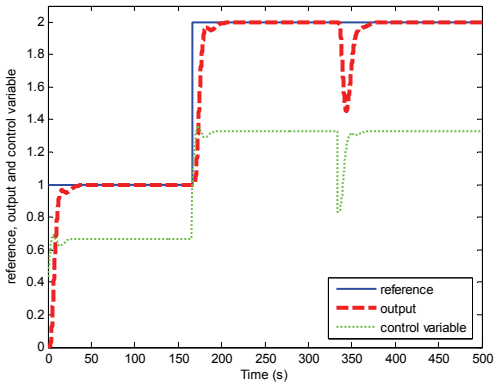


Figure 6: Control response for m=0.3

Example 2: A fourth order controlled system governed by the transfer function:

$$G(s) = \frac{1.5}{(1.5s+1)^4} \cdot e^{-3s} \tag{30}$$

was again approximated in similar way as a first order model:

$$\widetilde{G}(s) = \frac{1.5}{3.9s+1} \cdot e^{-5.7s} \tag{31}$$

The comparison of step responses is depicted in Figure 7. In (31), two different approaches to the time delay term were considered. The first one represents a simple neglecting of the time delay. The control response is depicted in Figure 8. The second approach utilizing Pade approximation (27) leads to responses shown in Figure 9.

As can be seen in the control responses, the value of the tuning parameter m>0 strongly influences the control response, e.g. the increasing values of m>0 accelerate the control response but escalate overshoots.

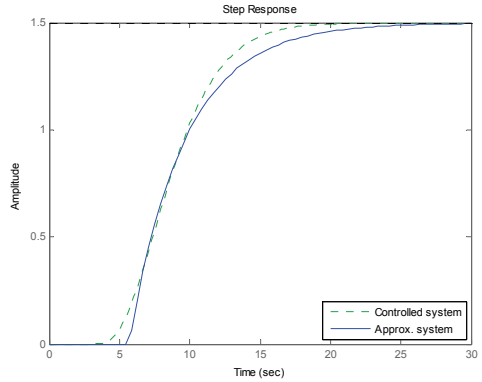


Figure 7: Comparison of step responses (30), (31)

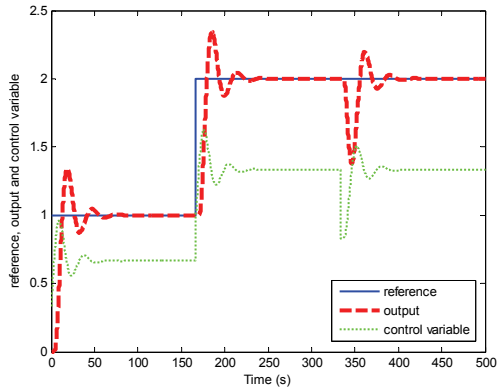


Figure 8: Neglected time delay, m=0.19

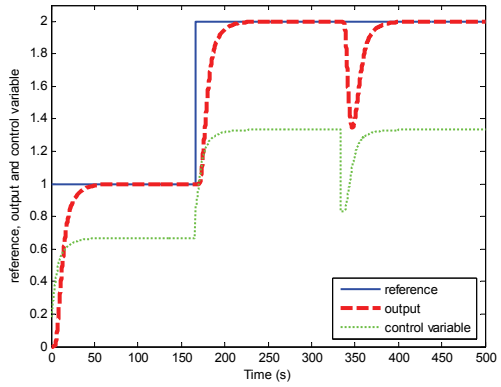


Figure 9: Approximation of time delay, m=0.19

Example 3: A higher order system was simultaneously identified as a first and second order model with transport delay. The controlled system was governed by transfer function:

$$G(s) = \frac{2}{(2s+1)^8} \cdot e^{-3s} \qquad (32)$$

After relay feedback experiments, the identification procedure for (1) and (11) gives two transfer functions:

$$\tilde{G}(s) = \frac{2}{7.7s+1} \cdot e^{-12.9s} \qquad (33)$$

$$\tilde{\tilde{G}}(s) = \frac{2}{(4.4s+1)^2} \cdot e^{-10.7s} \qquad (34)$$

In both cases, time delay terms were approximated by Pade simple formula (27) and then the polynomial control design procedure was performed.

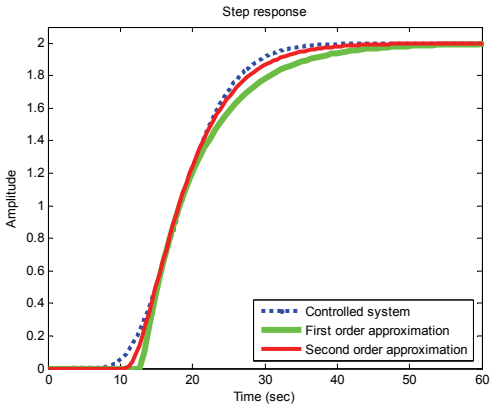


Figure 10: Comparison of step responses (32)-(34)

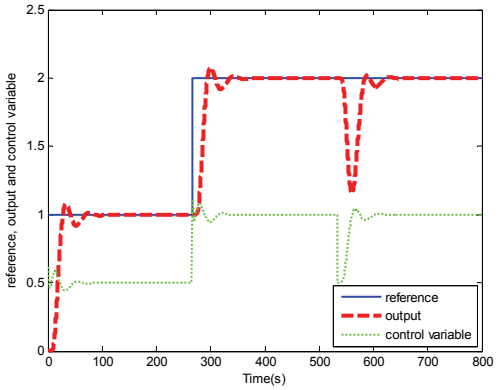


Figure 11: FOPDT model (33) - control response

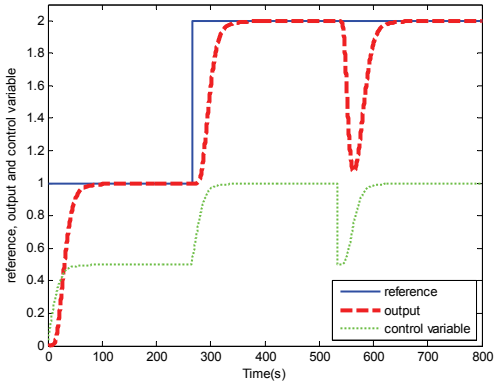


Figure 12: SOPDT model (34) – control response

For both models, controller parameters were generated with the same parameter $m=0.15$. Naturally, for control simulations the original system (32) was utilized. The importance and relevance of the order approximation is obvious and apparent. A higher order in Figure 12 exhibit very smooth and acceptable control behavior.

CONCLUSION

An autotuning method using an asymmetric relay feedback test is studied. The identification procedure yields three process parameters for aperiodic first or second order transfer functions. After the identification relay experiment, controller parameters are designed using linear diophantine equation in the ring of proper and stable rational functions. This algebraic approach for a traditional 1-DOF feedback structure gives a class of PI or PID controllers. The pole placement principle in the methodology brings a scalar positive "tuning knob" for additional controller tuning. A Matlab, Simuling program implementation was developed for simulation and verification of the studied approach. Two illustrative examples support simplicity and efficiency of the proposed methodology.

ACKNOWLEDGMENT

This work was supported by the grant of Ministry of Education, Youth and Sports of the Czech Republic, MSM 708 835 2102.

REFERENCES

Åström, K.J. and T. Hägglund 1984. "Automatic tuning of simple regulators with specification on phase and amplitude margins," *Automatica,* vol. 20, pp. 645-651.

Åström, K.J. and T. Hägglund 1995. *PID Controllers: Theory, Design and Tuning.* Research Triangle Park, NC: Instrumental Society of America.

Gorez, R. and P. Klán 2000. "Nonmodel-based explicit design relations for PID controllers," in: *Preprints of IFAC Workshop PID'00*, pp. 141-146.

Hang, C.C., K.J. Aström and Q.G. Wang 2001. "Relay feedback auto-tuning of process controllers – a tutorial review" In *Journal of Process Control 12*, pp. 143-162.

Ingimundarson, A. and T. Hägglund 2000. "Robust automatic tuning of an industrial PI controller for dead-time systems" In: *Preprints of IFAC Workshop PID'00,* 149-154.

Kaya, I. and D.P. Atherton 2001. "Parameter estimation from relay autotuning with asymmetric limit cycle data," in: *Journal of Process Control*, pp. 429-439.

Kučera, V. 1993. "Diophantine equations in control - A survey," *Automatica*, Vol. 29, pp. 1361-75.

Majhi, S. and D.P. Atherton 1998. "Autotuning and controller design for unstable time delay processes" In: *Preprints of UKACC Int. Conf. on Control,* 769-774.

Morilla, F., A. Gonzáles and N. Duro 2000. "Auto-tuning PID controllers in terms of relative damping" In: *Preprints of IFAC Workshop PID'00,* 161-166.

Pecharromán, R.R. and F.L. Pagola 2000. "Control design for PID controllers auto-tuning based on improved identification," in: *Preprints of IFAC Workshop PID'00*, pp. 89-94.

Prokop, R. and J.P. Corriou 1997. "Design and analysis of simple robust controllers," *Int. J. Control*, Vol. 66, pp. 905-921.

Prokop, R., P. Husták, and Z. Prokopová 2002. "Simple robust controllers: Design, tuning and analysis," in: *Preprints of 15th IFAC World Congress*.

Thyagarajan, T. and Ch.Ch. Yu 2002. "Improved autotuning using shape factor from relay feedback," in: *Preprints of IFAC World Congres.*

Vidyasagar, M. 1987. *Control system synthesis: a factorization approach.* MIT Press, Cambridge, M.A.

Vítečková, M. and A. Víteček 2004, "Experimentální identifikace metodou relé," in: *Automatizácia a informatizácia.*

Vyhlídal, T. 2000. "Anisochronic first order model and its application to internal model control," in: *ASR '2000 Seminar*.

Yu, Ch.Ch. 1999. *Autotuning of PID Controllers.* Springer, London.

AUTHOR BIOGRAPHIES

ROMAN PROKOP was born in Hodonin, Czech Republic in 1952. He graduated in Cybernetics from the Czech Technical University in Prague in 1976. He received post graduate diploma in 1983 from the Slovak Technical University. Since 1995 he has been at Tomas Bata University in Zlín, where he presently holds the position of full professor of the Faculty af Applied Informatics and a vice-dean of the faculty. His research activities include algebraic methods in control theory, robust and adaptive control, autotuning and optimization techniques. His e-mail address is: prokop@fai.utb.cz.

JIŘÍ KORBEL was born in Zlín, Czech Republic. He studied automatic control and informatics at the Tomas Bata University and graduated in 2004, now he is an post-graduate student and assistant at the Faculty of Applied Informatics in Zlín. His research activities include autotuning principles, algebraic and polynomial syntheses and modeling and simulations. His e-mail address is: korbel@fai.utb.cz.

ZDENKA PROKOPOVÁ was born in Rimavská Sobota, Slovak Republic. She graduated from Slovak Technical University in 1988 (Automatic Control) and the PhD degree she received in Technical Cybernetics in 1993 from the same university. Since 1995 she been working in Tomas Bata University in Zlin, Faculty of Applied Informatics. She works there as an associate professor. Her main research activities include mathematical modeling, simulation, control of technological systems, programming and application of database systems. Her e-mail address is: prokopova@fai.utb.cz.

TRANSPLANT EVOLUTION FOR OPTIMIZATION OF GENERAL CONTROLLERS

[1]Roman Weisser
[2]Pavel Ošmera
[3]Miloš Šeda
Institute of Automation and Computer Science
University of Technology
Technicka 2896/2, 616 69 Brno, Czech Republic
[1]E-mail: roman.weisser@gmail.com
[2]E-mail: osmera@fme.vutbr.cz
[3]E-mail: seda@fme.vutbr.cz

[4]Oldřich Kratochvíl
European Polytechnical Institute Kunovice
Osvobození 699, 686 04 Kunovice
Czech Republic
[4]E-mail: kratochvil@edukomplex.cz

KEYWORDS
Transplant evolution, grammatical-differential evolution, object trees, hierarchical structures, algebraic reducing of trees, crossover by linking.

ABSTRACT

This paper describes a new method of evolution that is named Transplant Evolution (TE). None of the individuals of the transplant evolution contains genotype as in Grammatical Evolution (GE). Each individual of the transplant evolution contains the phenotype in the tree structure. Reproduction methods as crossover and mutation work with parts of phenotypes (sub-trees). The hierarchical structure of grammar-differential evolution that is used for finding optimal structures and parameters of general controllers is described.

INTRODUCTION

This paper describes a new method of evolution that is named Transplant Evolution (TE). The transplant evolution is the two-level Grammar-Differential Evolution of Object Sub-trees (GDEOS), which is created by linking the grammatical and differential evolution together. Every individual is represented by the form of an object tree structure without the genotype. The meaning of name TE and GDEOS is equivalent. The GDEOS can be understood as a combination of a Grammatical Evolution (GE) (O'Neill and Ryan 2003), based on a grammar, and a Genetic Programming (GP) (Koza 1992). GDEOS unlike GE is based on tree structures and the individual of GDEOS does not store the genotype. The Grammar-Differential Evolution (GDE) is two-level evolutionary optimization, which combines modified grammatical evolution and Differential Evolution (DE). The modified GE (TE) creates optimal general structure and the DE sets suitable numeric value of the parameters. The main advantage of the Grammatical Evolution of Object Sub-trees (GEOS) or GDEOS compared to GE is the representation of the individual. The individual in GEOS (GDEOS) only generate phenotype without saving the genotype. The object tree structure of GEOS's individuals is created by randomly generated items of genotype. This approach allows dynamically change the production rules of GEOS, without losing already generated phenotype of the individual, further use an appropriate method of reducing the tree structure of the object, e.g. using algebraic operations. The TE method (GDEOS) enable to create a new type of crossover, called the crossover by linking of sub-trees and method that is called Algebraic Reducing of object Trees (ART). The ART can be used for algebraic minimization of tree structures. The GEOS or GDEOS also allows use of a different probability of selection of each rule in the grammar.

THE PRESENTATION OF THE OBJECT TREE STRUCTURES

The phenotype representation of the individual is stored in the object tree structure. Each of nodes in the tree structure, including the sub-nodes, is an object that is specified by a terminal symbol and the type of terminal symbols (Weisser 2010). All nodes are independent and correctly defined mathematical functions that can be calculated, e.g. the function *x-3*, shown on Fig. 1, is a tree structure containing a functional block (sub-tree).

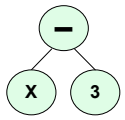


Fig. 1. Function block

Creating the object tree is a key part of GEOS, which this method differs from other evolutionary algorithms. When the object tree is generated, similar methods to a traditional grammatical evolution are used. But the GEOS does not store the genotype, because the production rules are selected by randomly generated genes that are not saved in chromosomes of individuals. The final GEOS's individual contains only phenotype expressed in an object tree structure.

The algorithm of GEOS uses a generative grammar (O'Neill and Ryan 2003) whose translation process starts from the initial symbol *S* and continues randomly with using the rules of defined grammar (Weisser 2010). The basic procedure of the translation algorithm

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

is shown on Fig. 2. The genotype is a random sequence of integers. From this genotype the tree structure is generated by defined production rules as in GE. We store only result tree structure, not the genotype.

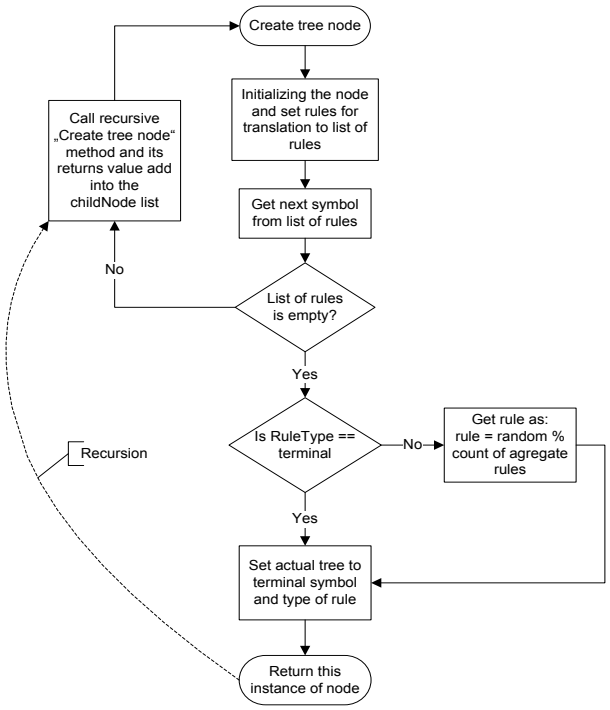


Fig. 2 Creation of object tree

CROSSOVER

The crossover is a distinctive tool for genetic algorithms and is one of the methods in evolutionary algorithms that are able to acquire a new population of individuals. For crossover of object trees can be used following methods, similar as in Genetic Programming (GP) (Koza 1992):

Crossover the parts of object trees (sub-trees)

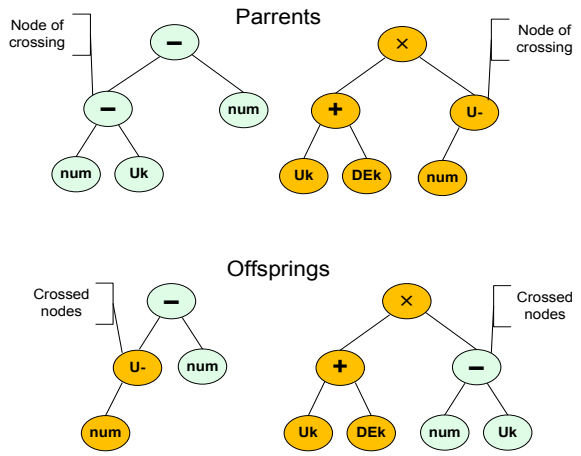


Fig. 3. Classical Crossover (CC)

The method of crossover object trees is based on the selection of two parents from the population and changing each other part of their sub-trees. For each of the parents cross points are randomly selected and their nodes and sub-trees are exchanged. This is the principle of creating new individuals into subsequent population as is shown on Fig. 3.

Crossover by linking trees or sub-trees

This method, as well as the previous one, is based on the crossover of two parents who are selected from the previous population. But the difference is in the way how the object trees are crossed. This method, unlike the previous one, does not exchange two randomly selected parts of the parents but parts of individuals are linked together with new and randomly generated node. This node will represent a new root of the tree structure of the individual. This principle is shown on Fig. 4.

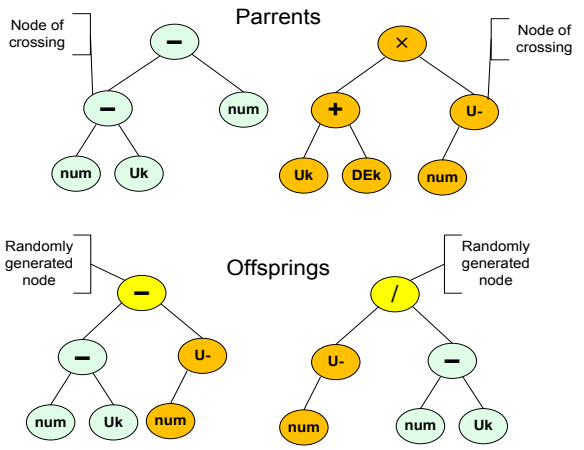


Fig. 4. Crossover by linking method (LC - Linking Crossing)

MUTATION

Mutation is the second of the operators to obtain new individuals. This operator can add new structures, which are not included in the population so far. Mutation is performed on individuals from the old population. The nodes in the individuals for mutation are selected randomly. The mutation operator can be subdivided into two types:

- Non-structural Mutation (NM)
- Structural Mutation (SM)

Non-structural Mutation (NM)

Non-structural mutations do not affect the structure of already generated individual. In the individual who is selected for mutation, chosen nodes of object sub-tree are further subjected to mutation. The mutation will randomly change chosen nodes, whereas used grammar is respected. For example it means that mutated node, which is a function of two variables (i.e. + - × ÷) cannot be changed by node representing function of one variable or only a variable, etc. see Fig. 5.

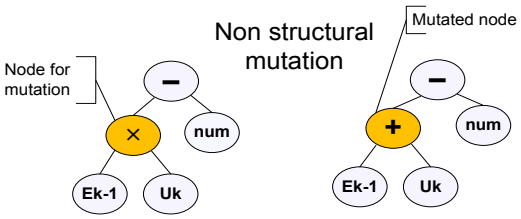


Fig. 5. Nonstructural mutation

Structural Mutation (SM)

Structural mutations, unlike non-structural mutations, affect the tree structure of individuals. Changes of the sub-tree by extending or shortening its parts depend on the method of structural mutations. Structural mutation can be divided into two types: Structural mutation which is extending an object tree structure (ESM) and structural mutation which is shortening a tree structure (SSM). This type of mutation operator can be subdivided into two types:

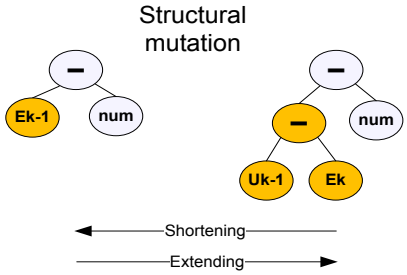


Fig. 6. Structural mutation

Extending Structural Mutation (ESM)
In the case of the extending mutation, a randomly selected node is replaced by a part of the newly created sub-tree that respects the rules of defined grammar (see fig. 3). This method obviously does not always lead to the extension of the sub-tree but generally this form of the mutation leads to extension of sub-tree. (see Fig. 6).

Shortening Structural Mutation (SSM)
Conversely the shortening mutation replaces a randomly selected node of the tree, including its child nodes, by node which is described by terminal symbol (i.e. a variable or a number). This type of mutation can be regarded as a method of indirectly reducing the complexity of the object tree (see Fig. 6).

The complexity of the tree structure can be defined as the total number of objects in the tree of individual.

DIRECT TREE REDUCTION

The minimal length of an object tree is often one of the factors required in the optimal problem solution. This requirement can be achieved in several ways:

- By penalizing the part of the individual fitness which contains a complex object tree,
- Method of targeted structural mutation of individual (see SSM),
- The direct shortening of the tree using algebraic adjustments - algebraic reducing tree (ART).

The last-mentioned method can be realised by the GEOS, where all of individuals does not contain the genotype, and then a change in the phenotype is not affected by treatment with genotype. The realisation of above mentioned problem with individual, which use genotype would be in this case very difficult. This new method is based on the algebraic arrangement of the tree features that are intended to reduce the number of functional blocks in the body of individuals (such as repeating blocks "unary minus", etc.). The method described above is shown on Fig. 7 and Fig. 8.

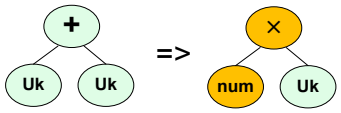


Fig. 7. ART – substitution of nodes

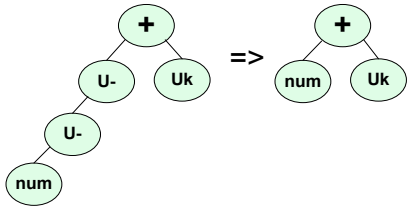


Fig. 8. ART – reduction multiple unary minus

In view of the object tree complexity of the individual and also for subsequent crossover is preferable to have a function in the form $x = 3a$ than $x = a + a + a$, or more generally $x = n \times A$. Another example is the shortening of the function $x = -(-a)$, where is preferable to have the form $x = a$ (it is removing redundant marks in the object tree individual). The introduction of algebraic modifications of individual phenotype leads to the shorter result of the optimal solution and consequently to the shorter presentation of the individual, shortening the time of calculation of the function that is represented in object tree and also to find optimal solutions faster because of higher probability of crossover in the suitable points with higher probability to produce meaningful solutions. The essential difference stems from the use of direct contraction of trees, which leads to significantly shorter resulting structure than without using this method.

HIERARCHICAL STRUCTURE OF TE (GDEOS) FOR OPTIMISATION OF THE CONTROLLER

The hierarchical structure of the transplant evolution can be used for optimisation of the structure and parameters of a general controller. This structure contains three layers. First two layers (GE + DE) are contained in TE. Those two layers are used for

optimisation of the structure and parameters of general controller. The third layer which is named layer of controller is used for computation of fitness in TE.

At the beginning of GDEOS an initial population is created (see Fig. 2) and then fitness of individuals is calculated. In the case of finding the optimal solution in the first generation, the algorithm is terminated, otherwise creates a new population of individuals by crossover and mutation operators, with the direct use of already created parent's object tree structures (it is analogy as transplantation of already created organs, without necessary know-ledge of DNA – "Transplant Evolution (TE)"). If the result of GDEOS needs some numerical parameters (for example *num* in (Weisser 2010)), the second level with Differential Evolution (DE) is used for optimization their parameter setting. The DE gives better results in finding optimal values of unknown numerical parameters that are expressed in the form of real numbers, then in the GE. Due to the use of GDEOS for optimization of controllers in the next stage of calculation of fitness is model of controller used (see Fig. 9). This model is represented by the equation in incremental form (recurrent algorithm). Quality of controller is determined depending on the type of criterial function (see equation 3). For fitness calculation are various criterial functions used. Basic criterion is linear control area, quadratic control area, linear or quadratic control area extended with overshoot, oscillation of action value of the controller.

The flowchart of TE (GDEOS) for a controller is shown on Fig. 9

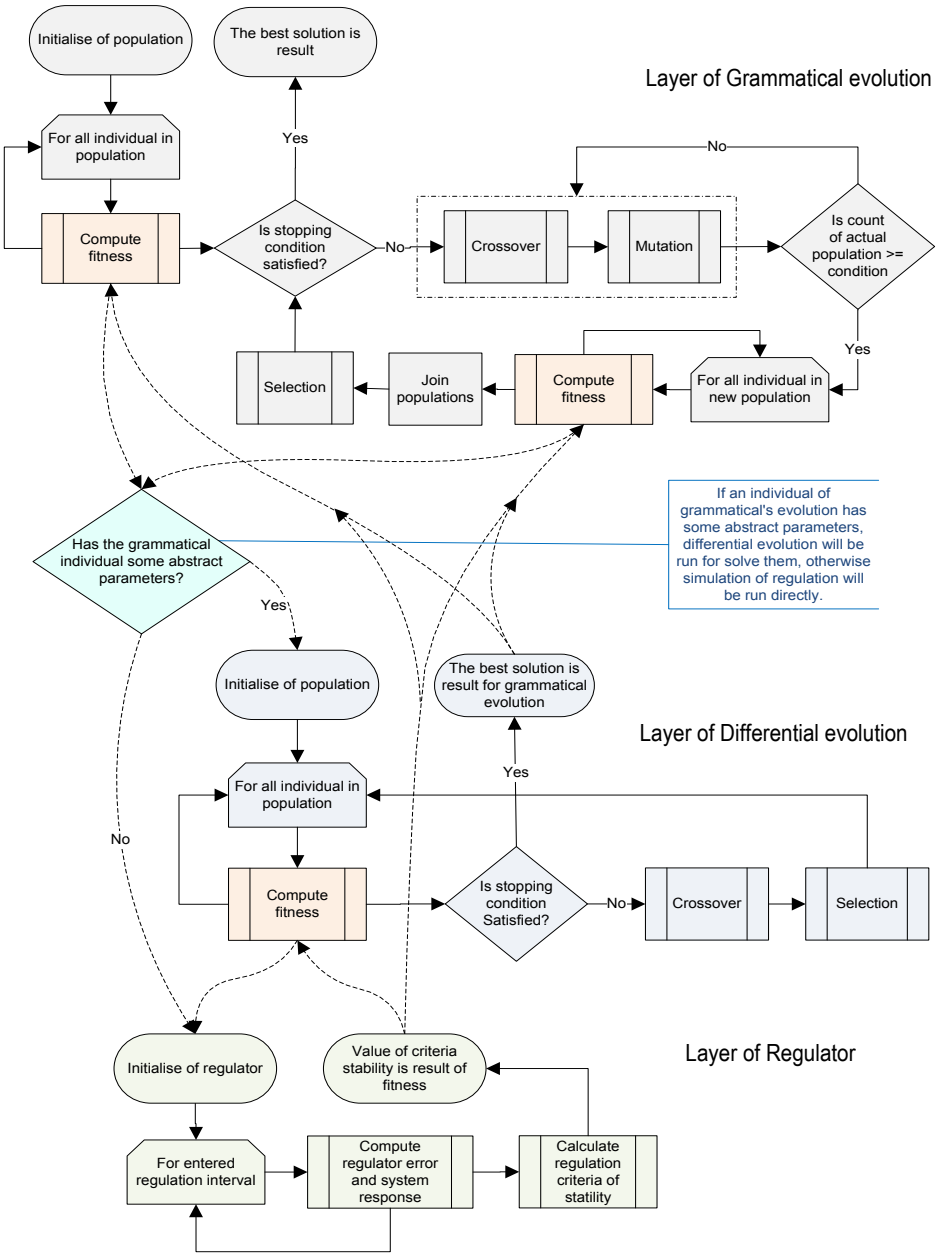


Fig. 9. Flowchart of TE (GDEOS) for controller

RESULTS

The TE and TE + ART methods for optimization of equation for general controller were compared.

The resulting form of the recurrent equation of general controller without using the direct method shortening of the tree (ART) is following (equation 1):

$$u_k=((((((E_k-(-((E_k+E_k))))\times3)\times2)+E_{k-3})-(-(2)))-(-(((((((E_{k-3}+((((((E_k\times3)+(E_k+(((dE_{k-1}+E_k)\times2)\times1.63)))\times2)\times2)+(-(((E_{k-4}+(E_k-(-(3))))+E_{k-2}))))\times3))+(-(((E_{k-4}+(E_k-E_k))+E_k-2))))\times2)+((dE_{k-1}+(4.47-((((E_k-(-((((((((((E_k\times3)+(E_k+(((dE_{k-1}+E_k)\times2)\times2)))\times2)\times2)\times2)+(-(-3.61)))+E_k)+E_k)+(-((3+E_{k-2})))))))+((E_k\times2)+E_k))+E_k)-(-(((E_k+2)\times(3\times((E_{k-4}+2)+E_{k-2})))))))-((((E_k+((((E_k+2)+E_k)+E_k)+(E_k/(-((6.88-((dE_{k-1}+(1.79-E_{k-3}))-2)))))))\times2)\times3)+(-(((E_{k-4}+((E_k+E_k)\times2))+E_k))))))+E_k)-(-(E_{k-2})))))$$

(1)

The resulting form of the recurrent optimization algorithm in the case with using the direct method of contraction tree is following (equation 2):

$$u_k = (E_{k-3} - E_k \times 1.93) \times 33.97 + E_{k-1} + E_{k-2} \quad (2)$$

As you can see, the resulting lengths of recurrent equation of the general controller, is shorter in case of using TE + ATR then TE without ART.

Bellow is shown result of optimisation parameters of PSD controllers and optimisation of the structure and parameters of general controllers. The parameters of PSD controllers were optimised with using DE and structure and parameters of general controller were optimised with using TE + ART method.

The basic criterion of minimal integral control area was used as criterial function for optimisation of PSD or general controllers, (see equation 3)

$$\boldsymbol{J = \int_0^{\infty} |e(t)|\, dt} \quad (3)$$

On the Fig. 10 is show regulatory process of PSD controller and on the Fig. 11 is shown regulatory process of general controller. The results on both pictures are for system with 5 second time delay. The parameters of PSD controller was optimised with using DE. Structure and parameters of general controller was optimised with using TE.

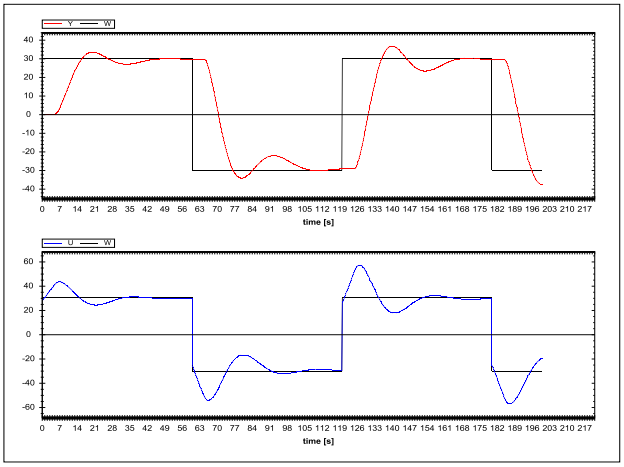


Fig. 10. Regulatory process of PSD controller for second order system with 5s time delay

(Top figure shows the system response. On the bottom figure the action output of the controller is shown.)

On the Fig. 11 is shown regulatory process of general controller. The structure and parameters of this controller was optimised with using TE (GDEOS) + ART method. The equation of general controller is following (see equation 4).

$$u_k = E_k \times 23.01 + 10.91 \times E_{k-3} + E_{k-1} \times (-33.91) + U_{k-1} \quad (4)$$

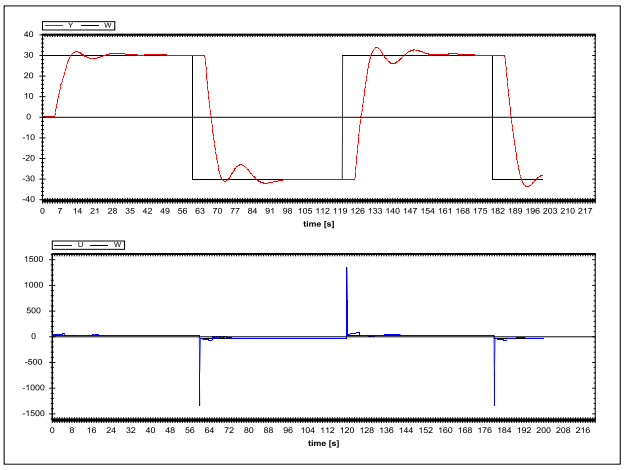


Fig. 11 Regulatory process of general controller for second order system with 5s time delay

(Top figure shows the system response. On the bottom figure the action output of the controller is shown.)

On the Fig. 12 and Fig. 13 are results of optimisation of PSD controller and general controller to control the identical system with 2 second time delay.

On the Fig. 12 is shown regulatory process of PSD controller. The parameters of PSD controller were optimised with using DE.

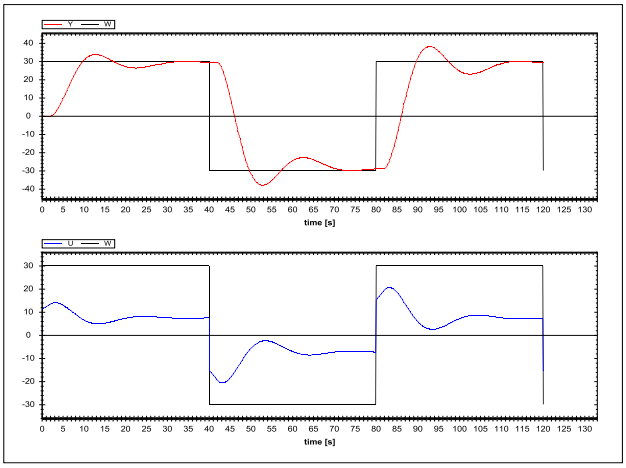


Fig. 12 Regulatory process of PSD controller for second order system with 2s time delay

(Top figure shows the system response. On the bottom figure the action output of the controller is shown.)

On the Fig. 13 is shown regulatory process of general controller. The structure and parameters of this controller was optimised with using TE (GDEOS) + ART method. The equation of general controller is following (see equation 5).

$$u_k = (32.55 \times E_{k-5}) + (58.79 \times E_k) + (-89.97 \times E_{k-2}) \quad (5)$$

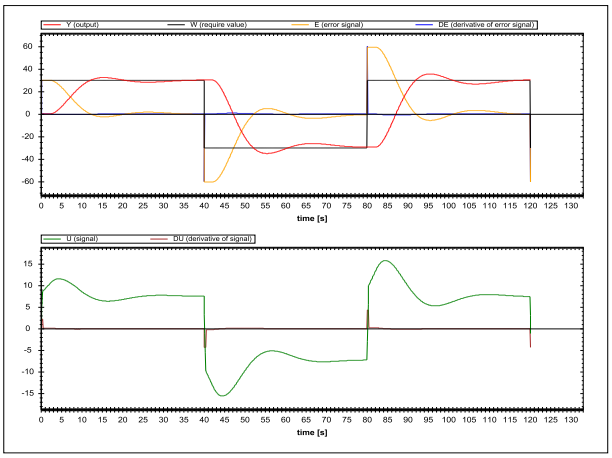


Fig. 13. Regulatory process of general controller for second order system with 2s time delay

(Top figure shows the system response. On the bottom figure the action output of the controller is shown.)

CONCLUSION

We have described the new method of evolution that was named GDEOS or even TE. This evolution method does not store a genotype of individual. The TE was proved for optimization of the controller. The proposed method can be very useful in case if the model of control system is in some of non-standard form. From the experimental session it can be concluded that modified versions of TE can create better results than classical versions of PSD controllers.

The TE can be used for the automatic generation of control formulae. We are far from supposing that all difficulties are removed but first results with TE are very promising.

Although we are at early stages of experiments, but it seems that it is possible to use parallel grammatical evolution with backward processing to generate combinatorial logic circuits. The grammatical algorithm can be outperformed with algorithms, which are designed specifically for this purpose.

REFERENCES

Koza J.R. 1992: Genetic Programming: On the Programming of Computers by Means of Natural Selection, The MIT Press

Kratochvíl O. and Ošmera P. and Popelka O. 2009: Parallel grammatical evolution for circuit optimization, in Proc. WCECS, World Congress on Engineering and Computer Science, San Francisco, 1032-1040.

Li Z. and Halang W. A. and Chen G. 2006: Integration of Fuzzy Logic and Chaos Theory; paragraph: Osmera P.: Evolution of Complexity, Springer, 527 – 578.

O'Neill M. and Ryan C. 2003: Grammatical Evolution: Evolutionary Automatic Programming in an Arbitrary Language Kluwer Academic Publishers.

O'Neill M. and Brabazon A. and Adley C. 2004: The Automatic Generation of Programs for Classification Problems with Grammatical Swarm, Proceedings of CEC, Portland, Oregon, 104 – 110.

Ošmera P. and Šimoník I. and Roupec J. 1995: Multilevel distributed genetic algorithms. In Proceedings of the International Conference IEE/IEEE on Genetic Algorithms, Sheffield, 505–510.

Ošmera P. and Roupec J. 2000: Limited Lifetime Genetic Algorithms in Comparison with Sexual Reproduction Based GAs, Proceedings of MENDEL, Brno, Czech Republic, 118 – 126.

Piaseczny W. and Suzuki H. and Sawai H. 2004: Chemical Genetic Programming – Evolution of Amino Acid Rewriting Rules Used for Genotype-Phenotype Translation, Proceedings of CEC, Portland, Oregon, 1639 - 1646.

Price K. 1996. Differential evolution: a fast and simple numerical optimizer, Biennial Conference of the North American Fuzzy Information Processing Society, NAFIPS, IEEE Press, New York, NY, 524-527.

Weisser R., 2010: Optimization of structure and parameters of controller, conference ICSC, Hodonin, Czech Republic (Jan.)

ROMAN WEISSER was born in Karvina, Czech republic and went to the University of Technology in Brno, where he studied mechanical engineering and automation and computer science. He obtained master degree in 2003. Since 2003 works as doctoral student at institute of automation and computer science. He worked for a couple of years for IT company that develop Manufacturing Execution Systems (MES). His e-mail address is roman.weisser@gmail.com

OLDŘICH KRATOCHVÍL was born in Czech republic. Since 1999 he is rector of EPI Kunovice. He obtained Dr.h.c degree at REA in Moscow. His e-mail address is kratochvil@edukomplex.cz

PAVEL OŠMERA was born in Trebic, Czech republic. He obtained master degree at Technical University of Brno in 1969. Since 2009 he works as professor at institute of automation and computer science. His e-mail address is osmera@fme.vutbr.cz

MILOŠ ŠEDA was born in Uherské Hradiště, Czech republic. He obtained master degree at Technical University of Brno in 1976 and RNDr. degree at Masaryk University of Brno in 1984. He was appointed professor of Design and Process Engineering at Brno University of Technology. His e-mail address is seda@fme.vutbr.cz

A Data Management Framework Providing Online-Connectivity in Symbiotic Simulation

Sebastian Bohlmann, Matthias Becker, Helena Szczerbicka
Department of Simulation and Modelling
Leibniz University Hannover
Welfengarten 1
30167 Hannover, Germany
{bohlmann, xmb, hsz}@sim.uni-hannover.de

Volkhard Klinger
Department of Embedded Systems
FHDW Hannover
Freundallee 15
30173 Hannover, Germany
volkhard.klinger@fhdw.de

KEYWORDS

Hybrid Process Simulation, Process Control, Communication Protocols, Symbiotic Simulation

ABSTRACT

Symbiotic simulation in industrial applications requires efficient connectivity between industrial processes and embedded legacy simulators. One main challenge is to handle heterogeneous system parameters like bandwidth, latency, redundancy, security and data representation. Moreover, data management needs to be improved in terms of unified access providing an interface to online, historical and corresponding simulation data.

This paper proposes a framework for symbiotic simulation addressing the problem of connectivity. We introduce the Process Data Streaming Protocol (PDSP) managing distributed process data flows. Additionally we present PDSP based modules covering different modes of operation for data processing. The Framework interacts with distributed control systems via object linking and embedding for process control as well as to embedded systems using different hierarchic operation modes. Historical data can be provided transparently through an integrated stream-database. The framework is primarily optimized to be used in JAVA-based simulation environments, but is not limited to these. Finally we demonstrate the usability of the system while interacting with a simulation environment for a hybrid process and present some experimental results.

INTRODUCTION

Nowadays it is possible by the help of today's microprocessor technology as well as the simulation software to simulate complex technical and industrial systems in a speed corresponding to the real process. Furthermore simulators use data provided by the underlying process (for example for reasons of the performance) to get better simulation results in a narrowly limited time period. As a consequence it is necessary to bind process and simulation with each other closely. This synchronization with the original process is seen as problematic because of the heterogeneous descriptions and interfaces. Since a consistent connection is, however, a key feature to maintain a symbiotic behaviour between a simulation and the real process, we will introduce a protocol to handle this. Furthermore experiences with a reference implementation of this component are presented. These are suitably for simulators targeting hybrid (continuous and discrete) processes. They can for example be used in plant simulation in the pulp and paper industry. For a prototypical integration to achieve experimental results the Hybrid Process Net Simulator (HPNS) framework (Bohlmann et al., 2009) will be used here. This currently under development hybrid simulator is based on models similar to petri nets as description form. The underlying communication protocol and the areas of application of the current implementation are in the foreground, though.

Technical systems are specified by the combination of different components which allow only a part-specific and heterogeneous description on account of the mechanical, electronic and information-processing subsystems. Hence, the modeling of the entire system in a closed simulation framework is not possible as a rule. The hardware/software-co-simulation represents a good example for this situation: Based on an interface and a synchronization process the simulators of the hardware and for the software subsystems are coupled together to simulate and to evaluate without a standardized model view. In particular in the automation technology there are almost no closed model-based simulation methods for both, the process and the necessary process control. In this paper we want to show a strategy to online simulate complex systems consisting of physical/chemical processes, mechanical systems and Hardware/Software-subsystems with the help of PDSP based components. Physical, chemical and mechanical subsystems are characterized primary by a continuous behaviour. Hardware subsystems can be characterized by continuous and discrete behaviour, while software is characterized by discrete behaviour.

PDSP ARCHITECTURE

In this section we describe the architecture of the Process Data Streaming Protocol, characterised by different modes of operation and requirements in simulation technology in general. PDSP is designed to be used in mixed continuous and discrete environments (Zeigler et al., 2000) referred to as hybrid. Focusing on symbiotic simulation (Fujimoto et al., 2002) PDPS is primary

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

designed to satisfy four modes of operation.

- Analytic mode
  The analytic mode is used to analyse a process based on captured data. Therefore historical data is used to create and tune models. In this mode no real-time or redundancy requirements exist. Data throughput and repeatability are the most important parameters.
- Transparent mode
  In this mode an online connection to the process is available. However the system is not influenced. Offline and online data are available and will be switched depending on the time index. This is transparent to the connected simulator application. Transparent mode is designed to be used for online and offline verification and process supervision e.g. soft sensor applications. In this way an online verification procedure can be re-executed on historical data multiple times with deterministic results if the simulator is deterministic.
- Online Mode
  Online mode is used to simulate a process and transmit the results back to the process. The data is directly transmitted between the process control system and the simulator. Therefore latency is minimized although proxy servers may be necessary for large scale simulations.
- Prediction mode
  This mode is the most advanced and primary used for proactive and symbiotic simulation. Advanced synchronisation methods are provided and multiple prediction results over a time horizon can be transferred back to the process. The transparent data source switching feature is also included and extended to be able to reuse previously predicted data. In this mode historical, present and future data are available to the process control system and the simulator application. Redundancy in this operation mode is an important feature to be integrated in industrial process control. Multiple simulation runs can connect in parallel to the data source with different wall-clock time. This is especially useful to simulate models with dead times.

PDSP should be classified as an application protocol (layer 7) in the Open System Interconnection reference model. The protocol encapsulates three inner layers. Figure 1 visualizes this structure. The bottom layer is used to multiplex multiple connections over a single connection. In Ethernet based TCP/IP networks it would also be possible to use multiple connections on lower OSI-layers but the resulting data flows would differ. This happens because of a PDSP assurance. All data flows being multiplexed in one connection can be linked to provide cross connection time-ordered data transmission. This means that no earlier sample or event than the current one can occur. Further no new authentication is processed resulting in lower response times. The layer 1 frame consists

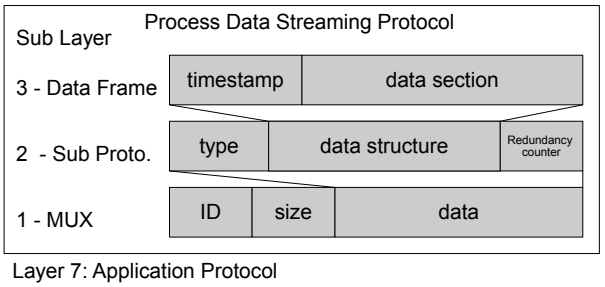


Figure 1: PDSP frame format

of a 32-bit stream id, a data size descriptor and the data part for the next layer. Layer 2 is used to implement different sub protocols e.g. for data transmission or flow management. The complete list will follow later on. The top layer 3 is not used by all layer 2 protocols. It is a data representation frame consisting of a 64-bit time stamp and a well defined data section. The format of the data section is constant for a single transmission stream and is communicated when the stream is initialized. It should be mentioned that the 3 layer structure is not used due to security reasons until authentication procedure is positive.

The PDSP protocol can be extended by protocols in inner layer 2. Although some sub protocols are predefined and necessary during startup operation. When a simulation environment connects to a process using PDSP it must define a minimum connectivity level. This number is used to let PDSP scale from simple embedded systems to Distributed Control Systems. Not every small device is able to handle all sub protocols. For example a dynamic online code loading and execution mechanism is defined in PDSP Java to Java machine communication. Useful to distribute precompiled simulation kernel over a multiple simulation nodes. A simple embedded system without a Java VM would be unable to provide a handler for this. Based on this a set of PDSP schema level is defined. When a connection is established both sites transmit their own maximum schema level and compare the result with a for this connection required level. The highest possible level is used for further communication. Note that it is possible that no connection can be established if the requested features are not provided. PDSP protocol is defined as a symmetric protocol. This means that there is no server and both sides provide and use exactly the same interface. PDSP defines multiple sub protocols:

- Data Stream
  This is the basic protocol for stream based data transmission. Layer 3 frames are transmitted.
- Flow management
  Used to establish, close, rewind connections, adjust connection properties like buffer sizes and transmit control commands e.g. end of stream if no online source is connected.
- Multiplex management

Manages new layer 1 connections. At the beginning a single default layer 1 connection is established to be able to use this protocol.

- Redundancy management
  Distributes a layer one connection on multiple PDSP connections to provide redundancy. If more than 2 connections are used a voting procedure is used.
- Synchronisation
  Provides a fast synchronisation path to read remote system timer and delays. Protocol frames of this sub protocol are transmitted with the highest priority.
- Remote Methode Invocation (RMI)
  Each application using PDSP can provide own RMI based interfaces. These are compiled to a binary command protocol simple enough to be used by 8-bit embedded systems. For example the proxy and database server of the later introduced framework uses this mechanism. Three level of operation are supported. Higher levels support more complex data structures for parameters and results.
- Online code loading
  If a PDSP Java to PDSP Java connection is established code which is not present on one side can be obtained form the corresponding partner during RMI usage.

The authentication is defined separately and always enforced. It is based on a certificate store or passwords depending on the PDSP schema level. Table 1 explains the used sub protocols corresponding to a schema level.

FRAMEWORK

The PDSP-Framework implementation consists of a central server, multiple connectors, a set of data processing modules, a management console and a report generator. In this section we present implementation details of these components. The central design pattern for data access is a stream based one. Its structure is presented in Figure 2. Although streaming interfaces usually only use one direction in data flow, this interface includes the necessary control commands in the opposite direction. In Figure 2 data flows from left to right. Status indicators e.g. end of stream are transmitted in parallel. Control commands use the opposite direction. Each stream has a beginning and an end; cycles are not allowed. To be able to use a data stream a stream-controller has to be attached. It is automatically attached at the correct position which is usually at the beginning. Some stream components e.g. buffers attach their own controller on their input side and capture events from the originally attached controller. For example some buffers read bigger chunks of data in order to decrease latency when historical data is accessed. It should be mentioned that this interface is identical to the combination of the data stream and flow control subprotocols of the PDSP definition. So each connection can be mapped over a network via PDSP or executed directly inside one machine.

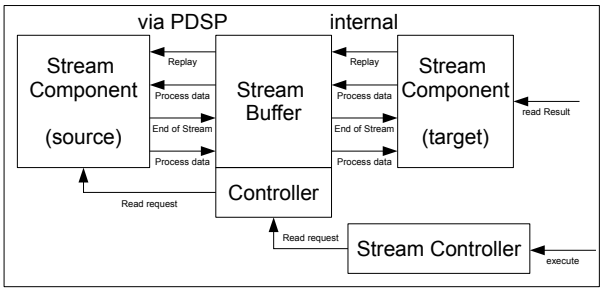


Figure 2: PDSP Stream Interface

Next we take a look into the central server. This server is combining features useful for standard employment scenarios. It is written in pure JAVA version 1.6 and can be extended by external jar-files. The structure is visualized in Figure 3. The central component is the Java implementation of the PDSP. There are several other modules accessible with the help of this interface:

- Database
  This component stores historical or predicted data streams. The back-end for this implementation is the relational database Derby. The Structured Query Language (SQL) interface of Derby is completely hidden. A stream based interface is used to hide all database operations. In symbiotic simulation environments the addressed data frequently is out of a defined sliding time window. This statement is supported by the database by the usage of a memory buffer. It provides fast access to recent data. Furthermore the database can be replicated to a remote database via PDSP without interruption.
- Proxy
  The server proxy can be used to split data streams over all connected PDSP links. In combination with the database component it is possible to access data streams at any time stamp where data is present. In symbiotic simulation this is can be used to spread the calculation across multiple compute nodes or implement security filter. Security filters are used to eliminate illegal output produced by a simulator.
- Catalog
  All PDSP based nodes connected to a server can announce services or data streams at a central place. The catalog can supervise data stream endpoints or RMI-interfaces.
- RMI Server
  The Remote Method Invocation server component is used to load Java based extensions provided by the user in a jar-file. The component must define an interface to be accessible via PDSP and can dynamically provide data stream endpoints to the catalog. Java based data connectors can use this to connect external systems.

To connect the framework to data sources usually a data connector has to be implemented. For faster eval-

Table 1: PDSP Schema Level and Features

Level	Authentication	Data Streaming	Multiplexing	Redundancy	Synchronisation	RMI	Online Code Loader
9	Yes	Yes	Yes	multi-path	Yes	full	Yes
8	Yes	Yes	Yes	partial	Yes	full	Yes
7	Yes	Yes	Yes	partial	Yes	full	No
6	Yes	Yes	Yes	partial	Yes	normal	No
5	Yes	partial	Yes	partial	basic	normal	No
4	Yes	partial	Yes	no	partial	basic	No
3	password	partial	Yes	no	partial	basic	No
2	password	partial	Yes	no	no	basic	No
1	password	partial	no	no	no	basic	No

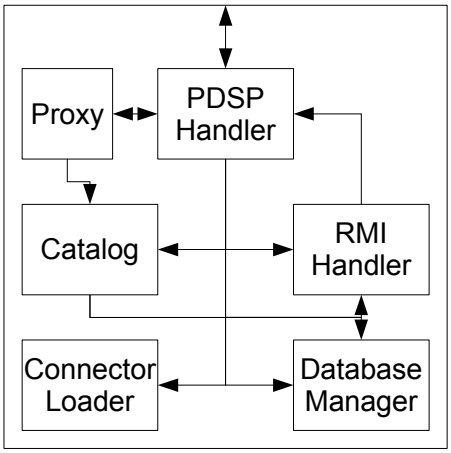


Figure 3: PDSP Server Structure

uation and easy usage some default connectors are implemented. The first one is a wrapper implemented in Java and uses the RMI server interface. It connects one or more serial devices or files. The in-/output format is equivalent to a comma separated text. The second more advanced connector is primary used to connect industrial equipment via OLE (Object Linking and Embedding) for Process Control (OPC-Task-Force, 1998). Because OPC uses the DCOM (Distributed Component Object Model) interface for communication this component is not written in Java. As PDSP is not limited to Java this connector uses C# and the .NET framework for OPC access via PDSP over TCP/IP. OPC is a popular interface in industrial automation equipment. Some of the functions are similar to the PDSP. Data can be read and written to items. The current PDSP-framework extends this data interface by the possibility to use other operation systems and programming languages, a transparent historical data access, reusable data processing components and so on. For industrial symbiotic simulation the performance of OPC is usually not sufficient. The third connector is still under development and enables a MSP430 microcontroller for PDSP access.

A core feature of the framework are the rich data processing components. All of these follow the stream pattern in Figure 2. In total there are more than 50 blocks implemented. Due to space constraints only four examples are presented. A continuous data endpoint transfers data samples with time stamps.

- Sample rate conversion filter
  For some simulations and analytic methods it is necessary to provide equidistant samples. Continuous signals have a bandwidth defined by the largest gap between two signals. The sample rate conversion filter uses an approximation algorithm to output higher rate equidistant data streams.

- Alignment filter
  This filter handles multiple continuous data streams in parallel. If a Distributed Control System (DCS) is this data source for the simulation usually not all sources provide a sample at the same time. For simulation it is sometimes useful to read the current system state at a well defined time. The alignment filter corrects this behaviour. It write the interpolated state to each output at fixed time stamp (with regard to the signal bandwidth).
- Forward filter
  A simple but sometimes useful filter. Some connected systems output data samples in incorrect temporal order. This filter has an internal buffer to provide the right ordered samples within a range and drop other samples.
- Event to RMI
  This stream component connects events arriving from a PDSP discrete stream to a PDSP remote method invocation interface. This is useful to control embedded simulations by the usage of the already existing DCS.

As mentioned before the PDSP framework includes a management console. Especially during Java development it can be used to test all RMI interfaces without extra implementation overhead. The structure is automatically generated from the Java class files. Special annotations can be used to provide extra information e.g. for the Linux style 'man' command. The console can be remote accessed only via PDSP to enforce authentication. It is enabled with a command line option in the framework server. Scripted command sequences are supported.

Last but not least the framework provides a interface for reporting. It is based on the eclipse Business Intelligence and Reporting Tools (BIRT). The Java PDSP implementation can be used to connect to BIRT using an Open Data Access connector (Weathersby et al., 2006). By the help of this tool chain high quality reports can be generated online during simulation and for example exported to a web server. Because reporting is often application specific to the simulation model, it is good to use the graphical BIRT editor for this. Almost no programming skills are necessary if the used simulator is supporting the framework database server. For symbiotic simulation the online reporting feature is essential to observe the performance of the hole system. With Java annotations all RMI interfaces can also be marked for ODA/BIRT usage. The corresponding stubs are automatically generated.

PDSP IN SYMBIOTIC SIMULATION

The rapid changing environment of manufacturing processes such as paper manufacturing requires a framework to identify process parameters and to optimize the process itself. State-of-the-art simulation systems (Low et al., 2007) provide no closed-coupled integration into the entire process. Even it concerns an on-line-simulator it does not emphasise this close relationship between the simulation system and the physical system. The paradigm of symbiotic simulation systems, defined at the Dagstuhl seminar (Fujimoto et al., 2002), reflects this interlocking of hard- and soft-information assembled inside the framework. Therefore the symbiotic simulation approach allows an interaction between the physical system, the automation and the simulation in an efficient and beneficial manner. To point out the characteristics of the PDSP-framework and to present its symbiotic characteristics we describe our work in developing this symbiotic simulation framework for a manufacturing process application.

In automation each technical process consists of three different parts: The process itself, the control equipment like control loops, PLCs (Programmable Logic Controller), process computer and the interface between process and automation, depicted in Figure 4. The whole automation environment is connected to an archive database. To achieve the project objective it is necessary to integrate the Hybrid Process Net Simulator framework into the entire automation environment. Therefore using the HPNS framework with a manufacturing or in more general with an arbitrary technical or functional process it is essential to link the HPNS framework with the process itself. There are different scenarios possible to realize this linkage and to fit the framework to the automation system. All scenarios provide another system perspective, for example the offline- or the online analysis. In the following sections the different modes of operation, introduced in section 2 are described without specifying the details of the connection.

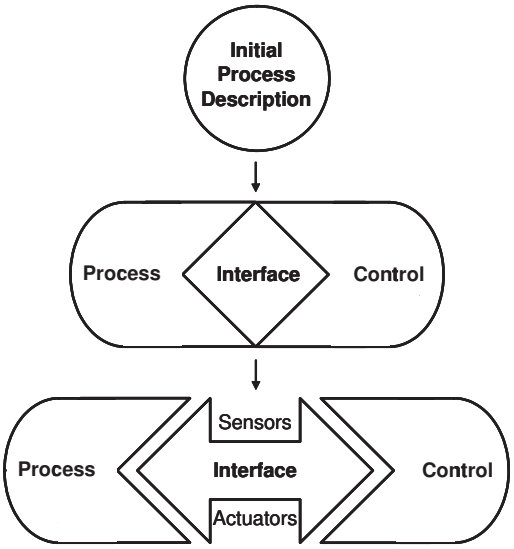


Figure 4: Process Decomposition

By using the analytic mode the HPNS framework can verify offline given scenarios based on available process data saved in the archive process database. This scenario provides the chance to evaluate the process behaviour and the process stability in special operating points. It avoids critical process states and supports the process operators to evaluate and to analyze the process characteristics and the process progression in more detail. To realize this it is necessary to connect the HPNS framework to the archive database. In addition a second database for triggering special states by using a list of event, called base sequences, is necessary. Based on this trigger data, special process working points can be defined for evaluation and analysis.

The transparent PDSP mode embeds the HPNS framework into the entire process without interaction. Here the HPNS is executed concurrent to the running manufacturing process. Due to the flexible architecture of the HPNS framework different options of embedding are possible. Every option shows other characteristics and provide a special point of view to fit different demands of interest. The two most interesting terms of embedding are introduced here:

- HPNS and automation (Figure 5, 1))
  HPNS is modelling the process on a required abstraction level. Therefore it is possible to connect the HPNS framework to the automation system and to replace the manufacturing process by the executed model. This is the reference scenario for the verification and testing of the automation system.
- HPNS and process (Figure 5, 2))
  As shown in Figure 5, 2) the process model consists of three different parts: The process, the control equipment and the interface between them. The HPNS framework is able to replace both entire parts or different subsystems to provide a type of development framework. Consequently, based on the

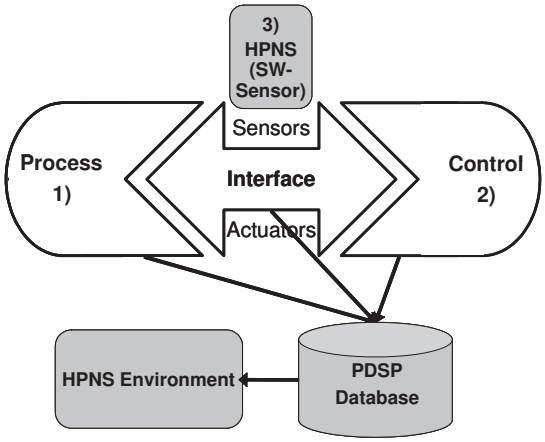


Figure 5: PDSP Process Integration

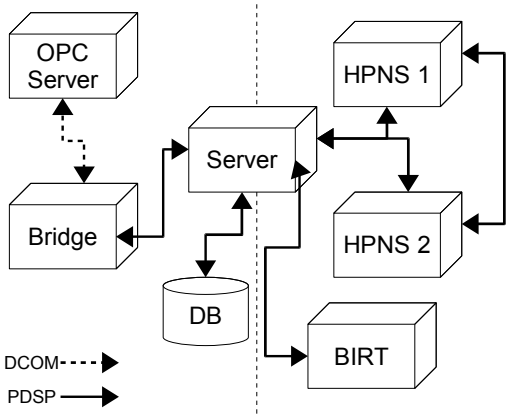


Figure 6: Data Flow

model the HPNS framework simulates the automation equipment.

The online mode is used to interact with the process. This is used e.g. for the SW-sensor capability (Figure 5, 3). The considered manufacturing process is realized based on state-of-the-art technology. Therefore not all process data can be achieved online, due to installation space, process dead-times, sensor failure et cetera. Soft sensors are inferential estimators, drawing conclusions from process observations when hardware sensors are unavailable, unsuitable or out of order. Based on a process model soft-sensors for monitoring and control of manufacturing processes requires an online model execution which can be realized using the HPNS framework. Therefore it is possible to deduce special process data from the process model even if the physical sensor is broken.

The prediction mode is the most complex interaction between simulator and process. Here the PDSP framework in combination with the HPNS is used to execute symbiotic dynamic models. In this mode the simulation environment can take advantage of low PDSP overhead to clone simulation instances for prediction (Bradley and Levent, 2008). In symbiotic simulation a signal feedback is provided using different voting algorithms. The application of modern SIMD Stream-Processors seems under real time restrictions promisingly and PDSP access had been implemented.

INTEGRATION EXAMPLES/RESULTS

To be able to get some first results and test the flexibility of the PDSP-framework it is integrated to the Hybrid Process Net Simulation environment. The HPNS-framework consists of different modules including one for model fitting and a simulation kernel for symbiotic simulation. Here we present two different experimental arrangements. The first one uses the framework database server connected via the PDSP-OPC-Bridge to a DCS in a paper mill. Over a few month about 1 TB data had been collected from 6,000 different data points. The data types were scalar and vector floating points. Then we connected a compute node via Gbit Ethernet to the server and applied a chain of data filters to the streams. These configuration is shown in Figure 6 on the left side. The single core database server could output a 190 Mbit constant data stream in analytic mode. This is equivalent to nearly 10,000,000 64-bit samples per second in random access to the different streams. But the compute node was unable to process this. The bottleneck as expected is the used alignment filter processing about 8,000 samples per second on a modern quad core machine. This indicates the usability of the framework in data analysis. The data throughput of two matching systems in common operation tasks should not be limited by the PDSP data transfer rates. For very simple computations the operation could also be invoked directly on the database server. This would reduce the transmission overhead.

The second scenario uses a USB-PLC connected by a specific implemented PDSP-bridge to an interface node. The PLC is interfacing a two tank system. A database server stores historical data a second transparent server connects the bridge in-line with an instance of the HPNS-Kernel. This second connection chain is using the predictive PDSP mode. Continuous sensor data e.g. water level or temperature and discrete events (for example limit switches) are transmitted to the simulator. Inside the HPNS are two models implemented. One for the two tank system and a second for the control logic (see Figure 5). Both are simulated in parallel. This structure is visualized in Figure 6. The control commands generated by the HPNS-Framework controlling the heater and the pumps are written back to the PDPS connection. In this experiment HPNS reverts time of the input data about 100 times a second to predict 100 possible behaviours. During operation no data transmission interruption could be measured. As expected the chronological access pattern enforced by the PDSP-server in prediction mode fits the data input behaviour of the HPNS-kernel. For test purposes this enforcement was disabled resulting in data transmission gaps of about 2ms. In combination with 100 occurrence this results in a 20% idling simulation

node. Beside the extra knowledge gained by chronological transmission order this indicates that for common simulation kernel a better performance could be generated. The round trip time for this control loop is 19ms in average. It is dominated by the two 8ms average delays in USB transmission. The four PDSP-links remain below 1ms each. This behaviour also proves that the sliding memory buffer of the PDSP server works because typical I/O-subsystems have higher access times.

CONCLUSIONS

The paradigm of symbiotic simulation environments can generate benefits in manufacturing processes. On one side symbiotic simulation environments have a distinct need of state of the art connectivity concerning performance and efficiency. On the other side the interfaces are getting more heterogeneous while simulation tools are used in more and more application areas. In this paper we introduce the PDSP framework as a possible solution. It couples tightly the data management and system connectivity. Furthermore online reporting for symbiotic simulation is part of the PDSP framework. While the framework is still under development we are able to obtain first performance parameters. In near future especially the consistent connectivity for embedded systems and radio based ad hoc networks will be enhanced. This will allow symbiotic simulation in new application areas using data from small wireless nodes.

REFERENCES

Bohlmann, S., Klinger, V., and Szczerbicka, H. (2009). Hpns : A hybrid process net simulation environment executing online dynamic models of industrial manufacturing systems. In Rosetti, M. D., Hill, R., and Johannsen, B., editors, *To be published in: Proceedings of the 2009 Winter Simulation Conference*. WSC Foundation.

Bradley, M. and Levent, Y. (2008). Symbiotic adaptive multisimulation: An autonomic simulation framework for real-time decision support under uncertainty. *ACM Transactions on Modeling and Computer Simulation*, 19.

Fujimoto, R., Lunceford, D., Page, E., and Uhrmacher, A. (2002). Technical report of the dagstuhl-seminar grand challenges for modelling and simulation. Technical Report No. 02351, Schloss Dagstuhl, Leibniz-Zentrum fr Informatik, Wadern, Germany.

Low, M. Y. H., Turner, S. J., Chan, L. P., Lendermann, P., Buckley, S., Ling, D., and Peng, H. L. (2007). Symbiotic simulation for business process re-engineering in high-tech manufacturing and service networks. In Henderson, S. G., Biller, B., Hsieh, M.-H., Shortle, J., Tew, J. D., and Barton, R. R., editors, *Proceedings of the 2007 Winter Simulation Conference*, pages 576–586. WSC Foundation.

OPC-Task-Force (1998). Opc overview 1.00. Technical report, www.opcfoundation.org.

Weathersby, J., French, D., Bondur, T., Tatchell, J., and Chatalbasheva, I. (2006). *Integrating and Extending BIRT (The Eclipse Series)*. Addison-Wesley Professional.

Zeigler, B. P., Praehofer, H., and Kim, T. G. (2000). *Theory of Modeling and Simulation: Integrating Discrete Event and Continuous Complex Dynamic Systems*. Academic Press, San Diego, USA, 2 edition.

AUTHOR BIOGRAPHIES

SEBASTIAN BOHLMANN is a Ph.D. candidate at Department of Simulation and Modelling - Institute of Systems Engineering at the Gottfried Wilhelm Leibniz Universität Hannover. He received a Dipl.-Ing. (FH) degree in mechatronics engineering from FHDW university of applied science. His research interests are machine learning and heuristic optimization algorithms, complex dynamic systems, control system synthesis and grid computing. His email address is <bohlmann@sim.uni-hannover.de>.

VOLKHARD KLINGER is a professor for embedded systems and computer science at the university of applied science FHDW in Hannover and Celle since 2002. After his academic studies at the RWTH Aachen he received his Ph.D. in Electrical Engineering from Technische Universität Hamburg-Harburg. During his 8-year research activity at the Technische Universität Hamburg-Harburg and research centres he focused on parallel and reconfigurable systems. He teaches courses in computer science, embedded systems, electrical engineering and ASIC/system design. His email address is <Volkhard.Klinger@fhdw.de>.

MATTHIAS BECKER is researcher and lecturer at the Department of Simulation and Modelling - Institute of Systems Engineering at the Gottfried Wilhelm Leibniz Universitt Hannover. He received his degree in Computer Science in Wrzburg 1996 and his Ph.D. from University Bremen in 2000. His research interests are simulation, modeling and optimization of discrete event systems. His email address is <xmb@sim.uni-hannover.de>.

HELENA SZCZERBICKA is head of the Department of Simulation and Modelling - Institute of Systems Engineering at the Gottfried Wilhelm Leibniz Universität Hannover. She received her Ph.D. in Engineering and her M.S in Applied Mathematics from the Warsaw University of Technology, Poland. Dr. Szczerbicka was formerly Professor for Computer Science at the University of Bremen, Germany. She teaches courses in discrete-event simulation, modelling methodology, queuing theory, stochastic Petri Nets, distributed simulation, computer organization and computer architecture. Her email address is <hsz@sim.uni-hannover.de>.

HYBRID ADAPTIVE CONTROL OF CSTR USING POLYNOMIAL SYNTHESIS AND POLE-PLACEMENT METHOD

Jiri Vojtesek, Petr Dostal and Roman Prokop
Faculty of Applied Informatics
Tomas Bata University in Zlin
Nad Stranemi 4511, 760 05 Zlin, Czech Republic
E-mail: {vojtesek,dostalp}@fai.utb.cz

KEYWORDS
Adaptive control, Polynomial approach, Pole-placement method, Recursive identification, CSTR

ABSTRACT

This paper deals with simulation studies of the adaptive control on the continuous stirred tank reactor (CSTR) as a typical chemical equipment with nonlinear behaviour and continuously distributed parameters. Mathematical model of this reactor is described by the set of two nonlinear ordinary differential equations (ODE). The simulation of the steady-state and dynamics results in optimal working point and external linear model (ELM) which is used in adaptive control. Adaptive approach here is based on the recursive identification of ELM and parameters of the controller are recomputed in each step too. The polynomial approach together with the pole-placement method and spectral factorization satisfies basic control requirements although the system has negative control properties.

INTRODUCTION

The most of the proceses in technical praxis has nonlinear behaviour and conventional control methods could lead to unoptimal or unwanted output responses. However, adaptive control (Åström 1989) is one way how to deal with the negative control properties such as non-minimum phase behaviour, nonlinearity, time delays etc.

There are several adaptive approaches. The one used in this work is based on the choice of an External Linear Model (ELM) parameters of which are recomputed recursively during the control (Bobal *et al.* 2005). It means that the steady-state and dynamic analyses need to be done for contructing of optimal ELM.

The ELM could be chosen from the range of the continuous-time or discrete-time models. Disadvantage of the continuous-time ELM is difficult on-line recursive identification. On the other hand, external delta models (Middleton and Goodwin 2004) used here for parameter estimation belong to the range of discrete models but parameters of these models approach up to some assumptions to their continuous-time counterparts (Stericker and Sinha 1993). Well known and widely used Ordinary recursive least squares method (Fikar and Mikles 1999) was used for parameter estimation during the control.

The polynomial method (Kucera 1993) used for controller synthesis together with the pole-placement method ensures basic control requirements such as stability, reference signal tracking and disturbance attenuation. Two control system configurations (Grimble 1994) were used – the first with one degree-of-freedom (1DOF) which has regulator only in feedback part and the second with two degrees-of-freedom (2DOF) with feedback and feedforward parts.

The resulting controllers are hybrid because polynomial synthesis is made for continuous-time but recursive identification runs on the delta-model, which belongs to the class of discrete-time models.

The Continuous Stired Tank Reactor (CSTR) which is an equipment widely used in chemical industry for its good control properties. Mathematical model of this plant is described by the set of two nonlinear ordinary differential equations which are solved numerically (Ingham *et al.* 2000). The system has two suitable input variables – volumetric flow rates of reactant and cooling and two outputs – product's temperature and concentration.

The article provides several control simulation studies on the CSTR, all of them were done on mathematical simulation software Matlab, version 6.5.1.

MODEL OF THE PLANT

The controlled process here is represented by the continuous stirred tank reactor (CSTR) and its scheme can be found in Figure 1.

Due to the simplification we expect that reactant is perfectly mixed and reacts to the final product with the concentration $c_A(t)$. The heat produced by the reaction is represented by the temperature of the reactant $T(t)$. Furthermore we also expect that volume, heat capacities and densities are constant during the control.

A mathematical model of this system is derived from the material and heat balances of the reactant and cooling. The resulted model is then a set of two Ordinary Differential Equations (ODEs) (Gao et al. 2002):

$$\frac{dT}{dt} = a_1 \cdot (T_0 - T) + a_2 \cdot k_1 \cdot c_A + a_3 \cdot q_c \cdot \left(1 - e^{\frac{a_4}{q_c}}\right) \cdot (T_0 - T) \quad (1)$$

$$\frac{dc_A}{dt} = a_1 \cdot (c_{A0} - c_A) - k_1 \cdot c_A$$

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

where $a_{1\text{-}4}$ are constants computed as

$$a_1 = \frac{q}{V}; a_2 = \frac{-\Delta H}{\rho \cdot c_p}; a_3 = \frac{\rho_c \cdot c_{pc}}{\rho \cdot c_p \cdot V}; a_4 = \frac{-h_a}{\rho_c \cdot c_{pc}} \quad (2)$$

variable t in previous equations denotes time, T is used for temperature of the reactant, V is volume of the reactor, c_A represents concentration of the product, q and q_c are volumetric flow rates of the reactant and cooling respectively. Indexes $(\cdot)_0$ denote inlet values of the variables and $(\cdot)_c$ is used for variables related to the cooling. The fixed values of the system are shown in Table 1 (Gao *et al*. 2002).

The nonlinearity of the model can be found in relation for the reaction rate, k_1, which is computed from Arrhenius law:

$$k_1 = k_0 \cdot e^{\frac{-E}{R \cdot T}} \quad (3)$$

where k_0 is reaction rate constant, E denotes an activation energy and R is a gas constant.

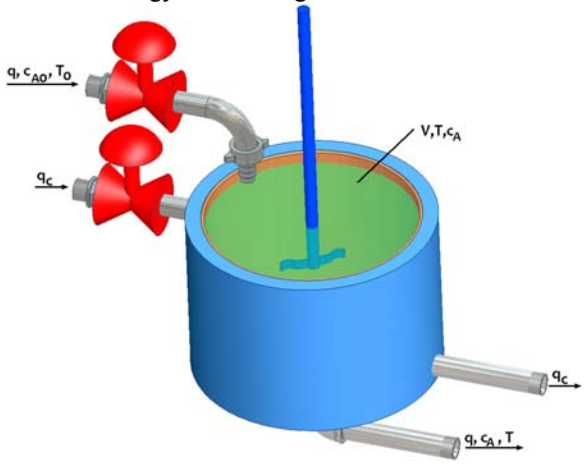


Figure 1: Continuous Stirred Tank Reactor

Table 1: Fixed parameters of the reactor

Reactant's flow rate	$q = 100\ l.min^{-1}$
Reactor's volume	$V = 100\ l$
Reaction rate constant	$k_0 = 7.2 \cdot 10^{10}\ min^{-1}$
Activation energy to R	$E/R = 1 \cdot 10^4\ K$
Reactant's feed temperature	$T_0 = 350\ K$
Reaction heat	$\Delta H = -2 \cdot 10^5\ cal.mol^{-1}$
Specific heat of the reactant	$c_p = 1\ cal.g^{-1}.K^{-1}$
Specific heat of the cooling	$c_{pc} = 1\ cal.g^{-1}.K^{-1}$
Density of the reactant	$\rho = 1 \cdot 10^3\ g.l^{-1}$
Density of the cooling	$\rho_c = 1 \cdot 10^3\ g.l^{-1}$
Feed concentration	$c_{A0} = 1\ mol.l^{-1}$
Heat transfer coefficient	$h_a = 7 \cdot 10^5\ cal.min^{-1}.K^{-1}$

STEADY-STATE AND DYNAMIC ANALYSES

The steady-state and dynamic analyses presented in (Vojtesek and Dostal 2008) have shown that the system has nonlinear behaviour and we cannot choose the exact optimal working point. The resulting working point is then combination of the possibly lowest values of the volumetric flow rates q and q_c:

$$q_c = 80\ l.min^{-1} \quad q = 100\ l.min^{-1} \quad (4)$$

The steady-state values of state variables T and are c_A for this working point

$$T^s = 354.26\ K \quad c_A^s = 0.9620\ mol.l^{-1} \quad (5)$$

The dynamic analysis for both outputs presented in the same paper (Vojtesek and Dostal 2008) shows that the output temperature could be described by the first or the second order transfer function and the second order transfer function could be used as a description of the output concentration c_A

ADAPTIVE CONTROL

The controlled variable was product's temperature T related to its steady-state value and the following simulation studies were done for both possible input volumetric flow rates of the reactant q and cooling q_c, i.e.

$$u_1(t) = \frac{q_c(t) - q_c^s}{q_c^s} \cdot 100\,[\%] \quad , \quad y(t) = T(t) - T^s\,[K] \quad (6)$$
$$u_2(t) = \frac{q(t) - q^s}{q^s} \cdot 100\,[\%]$$

External Linear Model(ELM)

Although the original system has nonlinear behaviour, the External Linear Model (ELM) is used as a reprentation of the controlled system.

As it is written above, the controlled output could be described by the second order transfer function with relative order one:

$$G(s) = \frac{Y(s)}{U(s)} = \frac{b(s)}{a(s)} = \frac{b_1 s + b_0}{s^2 + a_1 s + a_0} \quad (7)$$

This transfer function fulfils the condition of properness $\deg b \le \deg a$.

ELM described as (7) belongs to the class of continuous-time (CT) models. The identification of such processes is not very easy. The solution to this problem can be found in the use of so called δ–model which belongs to the class of discrete models but its parameters are close to the continuous ones for very small sampling period (Stericker and Sinha 1993).

The δ–model introduces a new complex variable γ computed as (Mukhopadhyay et al. 1992)

$$\gamma = \frac{z - 1}{\beta \cdot T_v \cdot z + (1 - \beta) \cdot T_v} \quad (8)$$

And it is obvious, that we can obtain infinitely many models for optional parameter β from the interval $0 \le \beta \le 1$ and a sampling period T_v, however a *forward δ-model* was used in this work which has γ operator computed via

$$\beta = 0 \Rightarrow \gamma = \frac{z - 1}{T_v} \quad (9)$$

The differential equation for ELM in the form of (7) is

$$y_\delta(k) = -a_1 y_\delta(k-1) - a_0 y_\delta(k-2) + \\ + b_1 u_\delta(k-1) + b_0 u_\delta(k-2) \tag{10}$$

where y_δ is the recomputed output to the δ-model:

$$y_\delta(k) = \frac{y(k) - 2y(k-1) + y(k-2)}{T_v^2}$$
$$y_\delta(k-1) = \frac{y(k-1) - y(k-2)}{T_v}$$
$$y_\delta(k-2) = y(k-2) \tag{11}$$
$$u_\delta(k-1) = \frac{u(k-1) - u(k-2)}{T_v}$$
$$u_\delta(k-2) = u(k-2)$$

and T_v is a sampling period, the data vector is then

$$\boldsymbol{\phi}_\delta^T(k-1) = \left[-y_\delta(k-1), -y_\delta(k-2), \ldots \right. \\ \left. \ldots, u_\delta(k-1), u_\delta(k-2)\right] \tag{12}$$

The vector of estimated parameters

$$\hat{\boldsymbol{\theta}}_\delta^T(k) = \left[a_1^\delta, a_0^\delta, b_1^\delta, b_0^\delta\right] \tag{13}$$

can be computed from the ARX (Auto-Regressive eXtrogenous) model

$$y_\delta(k) = \boldsymbol{\theta}_\delta^T(k) \cdot \boldsymbol{\varphi}_\delta(k-1) \tag{14}$$

by some of the recursive least squares methods.

The Recursive Least-Squares (RLS) method used for parameter estimation of the vector $\hat{\boldsymbol{\theta}}_\delta^T$ is well-known and widely used identification method (Fikar and Mikles, 1999). The ordinary RLS method is described by the set of equations:

$$\varepsilon(k) = y(k) - \boldsymbol{\varphi}^T(k) \cdot \hat{\boldsymbol{\theta}}(k-1)$$
$$\gamma(k) = \left[1 + \boldsymbol{\varphi}^T(k) \cdot \mathbf{P}(k-1) \cdot \boldsymbol{\varphi}(k)\right]^{-1}$$
$$\boldsymbol{L}(k) = \gamma(k) \cdot \mathbf{P}(k-1) \cdot \boldsymbol{\varphi}(k) \tag{15}$$
$$\mathbf{P}(k) = \mathbf{P}(k-1) - \gamma(k) \cdot \mathbf{P}(k-1) \cdot \boldsymbol{\varphi}(k) \cdot \\ \cdot \boldsymbol{\varphi}^T(k) \cdot \mathbf{P}(k-1)$$
$$\hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\theta}}(k-1) + \boldsymbol{L}(k)\varepsilon(k)$$

Where ε denotes a prediction error and $\mathbf{P}$ is a covariance matrix. This ordinary RLS method could be modified with some kind of forgetting, exponential or directional but simulation experiments have shown, that there is no need to use these modifications in this case .

Configuration of the Controller

Two control system configurations with one degree-of-freedom (1DOF) and two degrees-of-freedom (2DOF) were used here. The first, 1DOF, configuration has controller in the feedback part – see Figure 2.

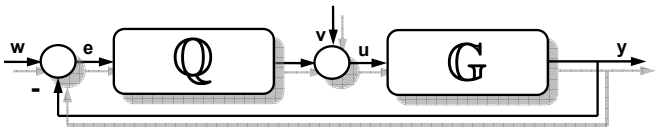


Figure 2: 1DOF control configuration

The configuration with two degrees-of-freedom (2DOF) displayed in Figure 3 has controller divided into feedback part (Q) and feedforward part (R).

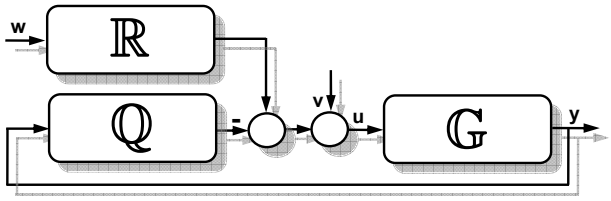


Figure 3: 2DOF control configuration

G in both configurations denotes transfer function (7) of controlled plant, w is the reference signal (wanted value), v is disturbance, e is used for control error, u is control variable and y is a controlled output.

The feedback and feedforward parts of the controller are designed with the use of polynomial synthesis:

$$Q(s) = \frac{q(s)}{s \cdot p(s)};\ R(s) = \frac{r(s)}{s \cdot p(s)} \tag{16}$$

where parameters of the polynomials $p(s)$, $q(s)$ and $r(s)$ are computed by the Method of uncertain coefficients which compares coefficients of individual s-powers from Diophantine equations (Kucera, 1993):

$$a(s) \cdot s \cdot p(s) + b(s) \cdot q(s) = d(s) \\ t(s) \cdot s + b(s) \cdot r(s) = d(s) \tag{17}$$

The resulted, so called "hybrid", controller works in the continuous time but parameters of the polynomials $a(s)$ and $b(s)$ are identified recursively in the sampling period T_v. It is clear, that the 1DOF control configuration counts only with the first equation in (17) while 2DOF configuration needs both Diophantine equations. Polynomial $t(s)$ in the second Diophantine equation is an additive stable polynomial with random coefficients, because these coefficients are not used for computing of coefficients of the polynomial $r(s)$ in 2DOF configuration. All these equations are valid for step changes of the reference and disturbance signals.

The feedback controller $Q(s)$ ensures stability, load disturbance attenuation for both configurations and asymptotic tracking for 1DOF configuration. On the other hand, feedforward part $R(s)$ ensures asymptotic tracking in 2DOF configuration.

The polynomial $d(s)$ on the right side of (17) is an optional stable polynomial. Roots of this polynomial are called poles of the closed-loop and their position affects quality of the control.

This polynomial could be designed for example with the use of Pole-placement method. The degree of the polynomial $d(s)$ is in this case

$$\deg d(s) = \deg a(s) + \deg \tilde{p}(s) + 1 \tag{18}$$

A choice of the roots needs some a priory information about the system's behaviour. It is good to connect poles with the parameters of the system via spectral factorization. The polynomial $d(s)$ then for our ELM (7) can be rewritten for aperiodical processes to the form

$$d(s) = n(s)\cdot(s+\alpha)^{\deg d - \deg n} \tag{19}$$

where $\alpha > 0$ is an optional coefficient reflecting closed-loop poles and stable polynomial $n(s)$ is obtained from the spectral factorization of the polynomial $a(s)$

$$n^*(s)\cdot n(s) = a^*(s)\cdot a(s) \tag{20}$$

The degrees of the polynomials $p(s)$, $q(s)$ and $r(s)$ are for this second order ELM deg $q(s)$ = 2, deg $q(s)$ = 1 and deg $r(s)$ = 1, which means that

$$\tilde{Q}(s) = \frac{q(s)}{s\cdot\tilde{p}(s)} = \frac{q_2 s^2 + q_1 s + q_0}{s\cdot(s+p_0)}$$
$$\tilde{R}(s) = \frac{r(s)}{s\cdot\tilde{p}(s)} = \frac{r_0}{s\cdot(s+p_0)} \tag{21}$$

and the polynomial $d(s)$ is from (19) of the fourth degree and it could be chosen as

$$d(s) = n(s)\cdot(s+\alpha)^2 \tag{22}$$

Parameters of the polynomial $n(s)$ which are computed from the spectral factorization are defined as:

$$n_0 = \sqrt{a_0^2},\ n_1 = \sqrt{a_1^2 + 2n_0 - 2a_0} \tag{23}$$

SIMULATION RESULTS OF CONTROL

Several simulation studies were done on this system. All of them were done in the mathematical software Matlab, version 6.5.1 and the common values for all simulations were: the sampling period was T_v = 0.3 *min*, the simulation time 1000 *min* and 5 different step changes were done during this time.

The input variables $u_1(t)$ and $u_2(t)$ were limited to the bounds <–90%; +90%>. The initial vector of parameters used for identification was $\hat{\boldsymbol{\theta}}^T = [0.1, 0.1, 0.1, 0.1]$ and the initial covariance matrix was $\boldsymbol{P}_{ii} = 1\cdot 10^7$ for i = 1,..,4.

The quality of control was evaluated by the quality criteria S_u and S_y computed for a whole time interval as

$$S_u = \sum_{i=2}^{N}\left(u(i)-u(i-1)\right)^2\ [-]$$
$$S_y = \sum_{i=1}^{N}\left(w(i)-y(i)\right)^2\ [K^2] \quad \text{, for } N = \frac{T_f}{T_v} \tag{24}$$

The first simulation study compares controllers with 1DOF and 2DOF control configutations for change of the volumetric flow rate of coolant, q_c, as an input variable. It is good to have opurtunity to affect the control output somehow. The tuning parameter in our case is position of the root α. And the effect of the different α = 0.02, 0.05 and 0.09 is studied here. Results are shown in following figures.

Courses of the output variable $y(t)$ and the input variable $u_1(t)$ presented in Figure 4 and Figure 5 show, that the decreasing value of the tuning parameter α results in slower output response and output does not reach reference signal $w(t)$ in 200 *min* which is used for each step change. On the other hand, course of the input variable is smoother for lower value of α.

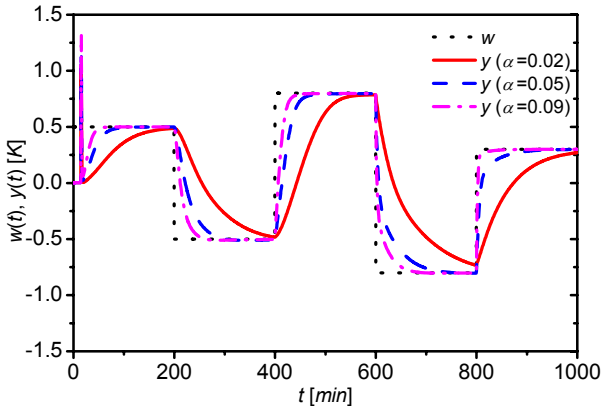


Figure 4: Output variable $y(t)$ for change of q_c as an input, 1DOF and various values of tuning parameter α

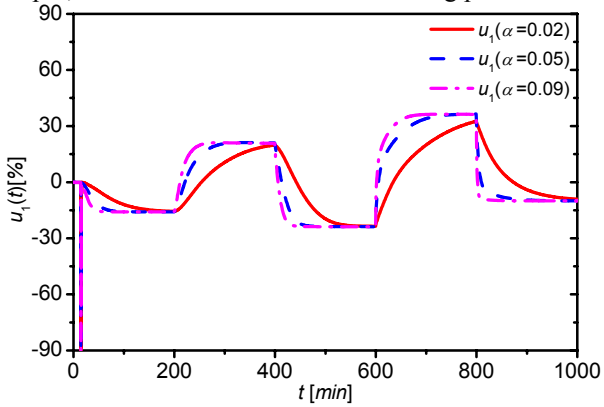


Figure 5: Input variable $u_1(t)$ for 1DOF and various values of tuning parameter α

The only control problem can be found at the very beginning which is caused by the identification, which does not have enough information and need some time to estimate parameters of the system's transfer function. However, output response in the following simulation time is smooth and without any problems – see course of the identified parameters in Figure 6 and Figure 7.

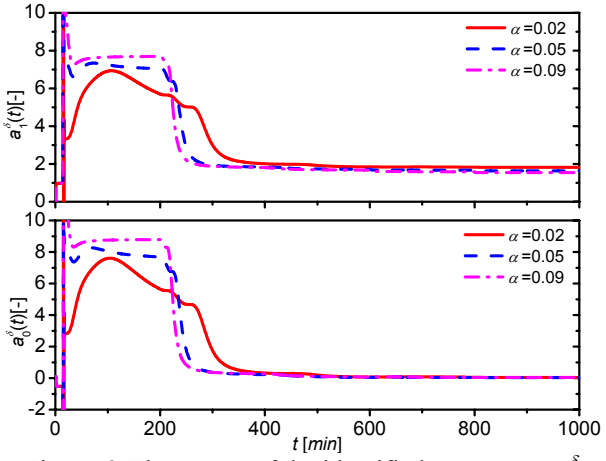


Figure 6: The course of the identified parameters a^{δ}_1 and a^{δ}_0, 1DOF configuration and various values of tuning parameter α

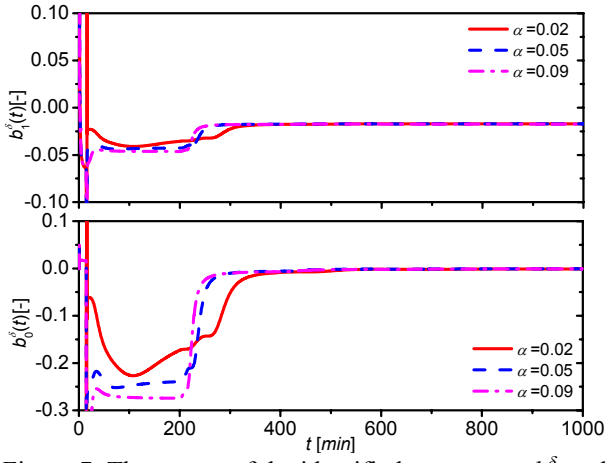


Figure 7: The course of the identified parameters b^{δ}_1 and b^{δ}_0, 1DOF configuration and various values of tuning parameter α

The second simulation study observes effects of 2DOF control configuration on the controller output. The course of the output temperature represented by the output $y(t)$ shows a bit worse results especially at the beginning of the control caused by the identification againg. The control results after approximately 100 *min* are again quite good which is displayed in Figure 8 and Figure 9.

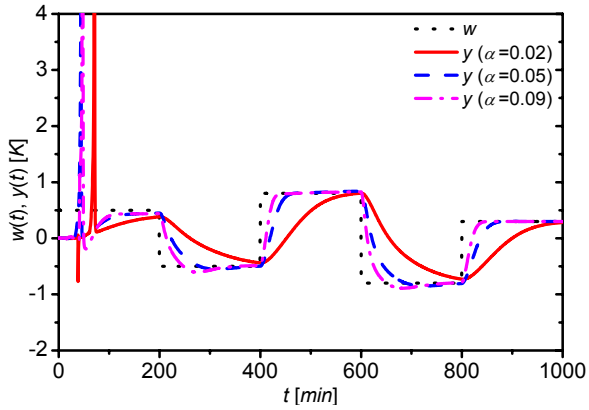


Figure 8: Output variable $y(t)$ for change of q_c as an input, 2DOF and various values of tuning parameter α

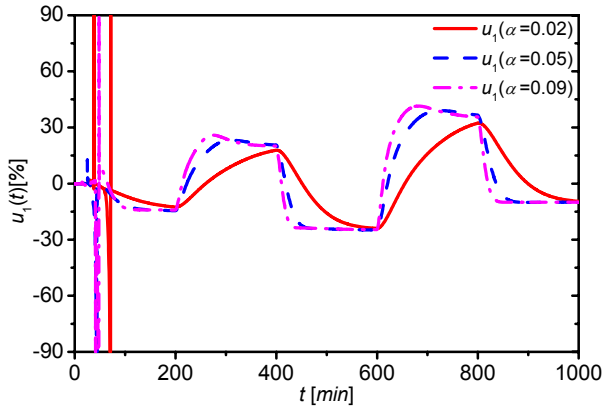


Figure 9: Input variable $u_1(t)$ for 2DOF and various values of tuning parameter α

Both, 1DOF and 2DOF, control configurations are compared in Figure 10 and Figure 11 for the tuning parameter $\alpha = 0.09$. The main difference can be found in first 100 *min* which was already mentioned and the 2DOF control configuration has small overshoots for step changes from the bigger value to lower.

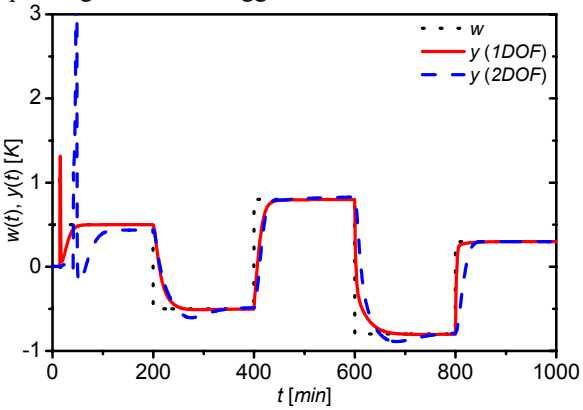


Figure 10: Comparison of 1DOF and 2DOF configuration, course of the output variable $y(t)$, change of q_c as an input and tuning parameter $\alpha = 0.09$

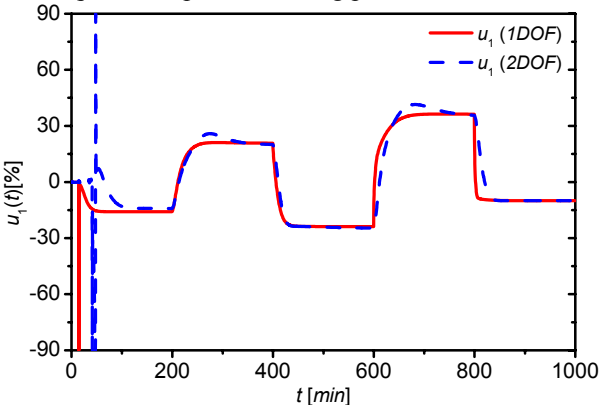


Figure 11: Comparison of 1DOF and 2DOF configuration, course of the input variable $u_1(t)$ and tuning parameter $\alpha = 0.09$

The third study, results of which are displayed in Figure 12, Figure 13 and Figure 14, have been done for the different input $u_2(t)$ which is change of the input volumetric flow rate of the reactant q.

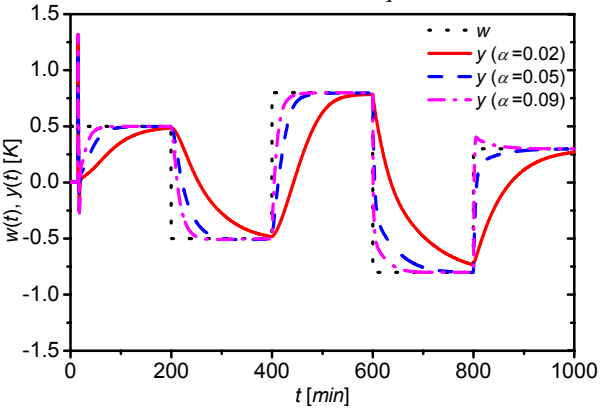


Figure 12: Output variable $y(t)$ for change of q as an input, 1DOF and various values of tuning parameter α

The results clearly implies that proposed control strategy could be used for this input too with similar results. The 2DOF configuration provides worse results especially at the beginning of the control and after the second step change too – see Figure 13 and Figure 14.

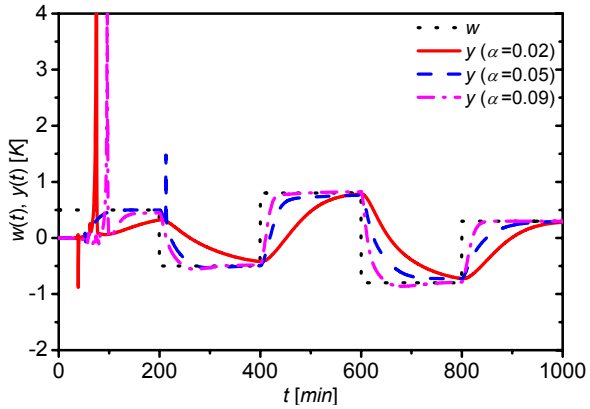


Figure 13: Output variable $y(t)$ for change of q as an input, 2DOF and various values of tuning parameter α

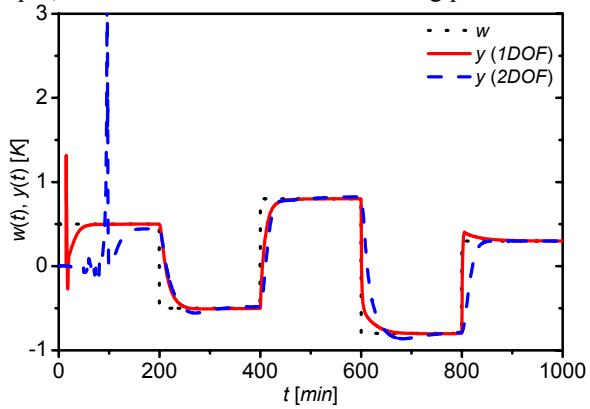


Figure 14: Comparison of 1DOF and 2DOF configuration, course of the output variable $y(t)$, change of q as an input and tuning parameter $\alpha = 0.09$

The values of control quality criterions Su and Sy for both input variables u_1 for q_c and u_2 for q, 1DOF and 2DOF control configurations are shown in Table 2 and Figure 15 and Figure 16.

Table 2: Values of control quality criteria S_u and S_y

	1DOF		2DOF	
	S_u [-]	S_y [K^2]	S_u [-]	S_y [K^2]
Results for $u_1(t)$ – volumetric flow rate q_c				
$\alpha = 0.02$	20.85	690.79	15.59	949.30
$\alpha = 0.05$	144.31	198.21	49.26	377.28
$\alpha = 0.09$	498.59	94.39	84.72	216.81
Results for $u_2(t)$ – volumetric flow rate q				
$\alpha = 0.02$	15.04	695.03	10.89	934.83
$\alpha = 0.05$	146.63	174.13	14.442	409.28
$\alpha = 0.09$	909.11	71.36	60.60	213.43

The values of criterions S_u and S_y from (24) are computed not from the very beginning, but after the first step change, i.e. from time 200 *min* due to inaccurate identification at the very beginning. This table and graphs verify previous statements, that increasing value of a results in quicker output response – see decreasing value of S_y but value of S_u representing the change of input variable increase.

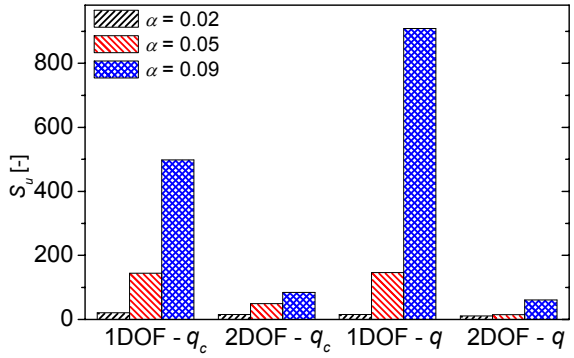


Figure 15: Values of control quality criterion S_u for all simulation studies

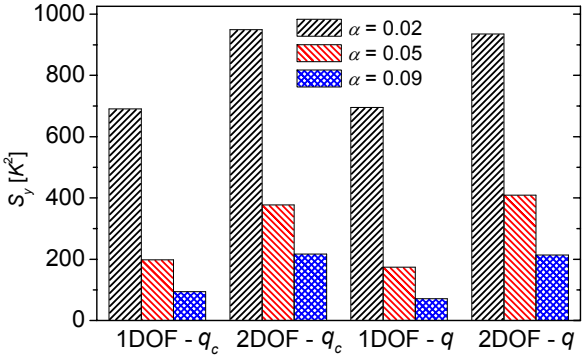


Figure 16: Values of control quality criterion S_y for all simulation studies

The proposed controller fulfills basic control requierements and one of it is disturbance attenuation. The last study provides proof to this statement. Simulation took 600 *min*, one step change of the reference signal $w(t)$ was done during this time and three step disturbances are inceted to the system – two on the input and one on the output:

- $v_1(t)$ = +10% step change of the input concentration c_{A0} for time $t \in \langle 150; 600 \rangle min$
- $v_2(t)$ = -1.5 K step change of the input reactant's temperature T_0 for time $t \in \langle 300; 600 \rangle min$
- $v_2(t)$ = +0.4 K step change of the output product's temperature T for time $t \in \langle 450; 600 \rangle min$

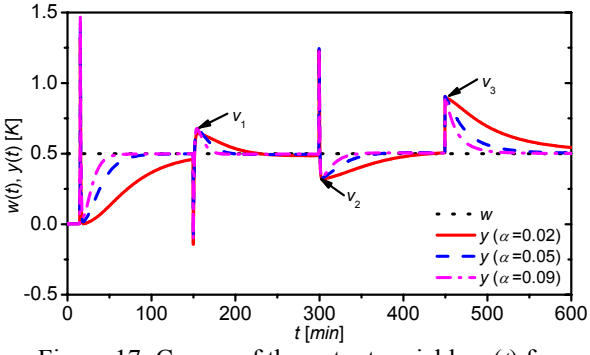


Figure 17: Course of the output variable y (t) for disturbances and various values of tuning parameter α

As you can see in Figure 17 and Figure 18. adaptive controller has no problem with the disturbance attenuation. Note that all three disturbances affects together in the last interval from 450 to 600 *min* and controller dealt with it quite good.

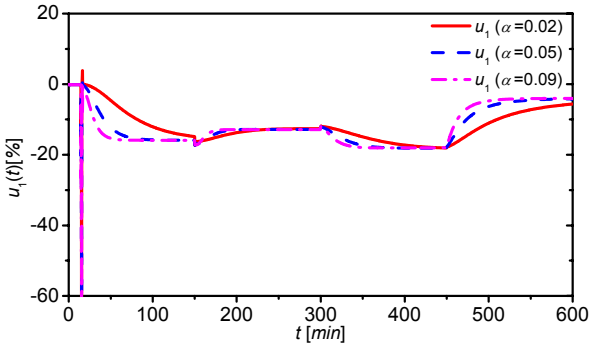


Figure 18: Course of the input variable $u_1(t)$ for disturbances and various values of tuning parameter α

CONCLUSION

This paper presents simulation studies on the CSTR focused especially on the adaptive control. Mathematical model of system is described by the set two nonlinear ordinary differential equations which are solved numerically by the Runge-Kutta's standard method. Steady-state and dynamic analyses result in the working point and the choice of the second order transfer function with relative order one as a ELM of the originally nonlinear process. The control results for both 1DOF and 2DOF control configuration presents good course of the product's temperature as an output variable. The course of this output can be affected by the choice of the root α and we can say, that increasing value of this root results in quicker response. The only disadvantage can be found in inadequate behaviour at the begging of the control which is cased by the absence of the initial information about the system. The system could have two input variables but simulations have shown that the change of the volumetric flow rate of the cooling, q_c, have a bit better control results than the change of the volumetric flow rate of the reactant, q.
The proposed control strategies satisfies basic control requirements which was proofed in the last study where three disturbances were injected into the system but the results are again quite good.

REFERENCES

Åström, K.J. and B. Wittenmark 1989. *Adaptive Control*. Addison Wesley. Reading. MA.

Bobal, V.; J. Böhm; J. Fessl; J. Machacek 2005. *Digital Self-tuning Controllers: Algorithms. Implementation and Applications*. Advanced Textbooks in Control and Signal Processing. Springer-Verlag London Limited

Fikar, M.; J. Mikles 1999. *System Identification*. STU Bratislava

Gao, R.; A. O'dywer; E. Coyle 2002. A Non-linear PID Controller for CSTR Using Local Model Networks. *Proc. of 4th World Congress on Intelligent Control and Automation*. Shanghai. P. R. China. 3278-3282

Ingham, J.; I. J. Dunn; E. Heinzle; J. E. Přenosil 2000. *Chemical Engineering Dynamics. An Introduction to Modeling and Computer Simulation*. Second. Completely Revised Edition. VCH Verlagsgesellshaft. Weinheim.

Kucera, V. 1993. Diophantine equations in control – A survey. *Automatica*. 29. 1361-1375

Middleton, R.H.; G. C. Goodwin 2004 *Digital Control and Estimation - A Unified Approach*. Prentice Hall. Englewood Cliffs.

Mukhopadhyay, S.; A. G. Patra; G. P. Rao 1992. New class of discrete-time models for continuos-time systeme. *International Journal of Control*. vol.**55**. 1161-1187

Stericker, D.L; N. K. Sinha 1993. Identification of continuous-time systems from samples of input-output data using the δ-operator. *Control-Theory and Advanced Technology*. vol. 9. 113-125

Vojtesek, J; P. Dostal 2008. Simulation analyses of continuous stirred tank reactor. In: *Proceedings of the 22nd European Conference on Modelling and Simulation*. Nicosia, Cyprus. p. 506-511.

ACKNOWLEDGMENT

This work was supported by the Ministry of Education of the Czech Rep. under grants No. MSM 7088352101 and MSM 7088352102.

AUTHOR BIOGRAPHIES

JIRI VOJTESEK was born in Zlin. Czech Republic and studied at the Tomas Bata University in Zlin. where he got his master degree in chemical and process engineering in 2002. He has finished his Ph.D. focused on Modern control methods for chemical reactors in 2007. His contact is `vojtesek@fai.utb.cz`.

PETR DOSTAL studied at the Technical University of Pardubice. He obtained his PhD. degree in Technical Cybernetics in 1979 and he became professor in Process Control in 2000. His research interest are modeling and simulation of continuous-time chemical processes. polynomial methods. optimal. adaptive and robust control. You can contact him on email address `dostalp@fai.utb.cz`.

ROMAN PROKOP was born in Hodonín, Czech Republic in 1952. He graduated in Cybernetics from the Czech Technical University in Prague in 1976. He received post graduate diploma in 1983 from the Slovak Technical University. Since 1995 he has been at Tomas Bata University in Zlín, where he presently holds the position of full professor of the Department of Automation and Control Engineering and a vice-dean of the Faculty of Applied Informatics. His research activities include algebraic methods in control theory, robust and adaptive control, autotuning and optimization techniques. His e-mail address is: `prokop@fai.utb.cz`.

SIMULATION OF WATER USE EFFICIENCY TO TACKLE THE DROUGHT

Asha Karunaratne
Faculty of Agricultural sciences
Sabaragamuwa University
Belihuloya, 70140
Sri Lanka
E-mail:asha.karunaratne@yahoo.co.uk

Neil Crout
Sayed Azam-Ali
School of Biosciences
University of Nottingham
Nottingham, NG7 2RD, UK
E-mail:Neil.Crout@nottingham.ac.uk

KEYWORDS
Climate change, drought, bambara groundnut, water use efficiency, root growth

ABSTRACT

Rainfed environments are characterised by unpredictable and highly variable seasonal rainfall and hence highly variable yields. Water use efficiency (WUE, yield per unit of water use) is commonly used for agricultural production with limited water resources. Expertise working towards the water resources need to address the multitudinous aspects in which cropping systems and amounts, timing and methods of irrigation, and fertilizer applications may be changed to improve WUE while maintaining yield and harvest quality goals. Since experimentation cannot address all scenarios accurate simulation models may fill in the gaps. Crop simulation models are used widely to predict crop growth and development in studies of the impact of climatic change. The present paper explains the model for WUE for an underutilised crop, bambara groundnut under drought as a sub-module of BAMGRO main model (Karunaratne, 2009). This quantitative model explains the root growth, root distribution and water uptake on daily basis under variable climatic conditions. The model links the size and distribution of root system to the capture of water over the growing period. The model was calibrated using glasshouse experimental data, Nottingham, UK and published information. It was validated against 2 years of independent data sets (2007, 2008) from Nottingham and field site at Notwane, Botswana. Although the limited information on root growth is available, validation of soil moisture against glass house and field reported satisfactory results.

INTRODUCTION

Although global biomass resources are vast and underutilised, over the coming decades, in the face of a growing population and a changing climate, there is likely to be increased pressure on plant resources for food, fuel and other plant products as we move from an oil-based to a bio-based economy. The consequences are the increase in pressure on global agricultural productivity. Plant biologists, agronomists, crop modellers and breeders should therefore consider the future of crop production in a changing climate.

The marginal nature and adverse climatic conditions in the world's arable lands challenge the existence of major crop species. Around the world, species that are little used, or which were grown traditionally but have fallen into disuse, are being brought out of the shadows and put to use, especially in the hands of the poor. Over 7,000 plant species have been grown or collected for food. But worldwide, less than 150 have been commercialised and just three crops - maize, wheat and rice - supply half of our daily proteins and calories. Yet a large number of crops that are now overlooked have the potential to play a much more important role in sustaining livelihoods and enhancing environmental health. The underutilised crops are surely the crops of the future (http://www.cropsforthefuture.org/) that will survive under extreme climatic conditions. Bambara groundnut (*Vigna subterranea* (L.) Verdc) is one such indigenous legume with significance as a source of protein in sub-Saharan Africa where it is mainly grown by women farmers in subsistence agricultural systems in, despite the lack of any major research effort until recently. Its nutritional composition (protein content is 16-25%) is highly comparable or superior to other legumes (Linnemann and Azam-Ali 1993), providing an important supplement to cereal-based diets.

The unpredictable variability of climate especially with erratic distribution of annual rainfall in sub-Saharan Africa routinely causes severe yield losses. The undoubted importance of water conditions for crop growth and development has been identified on many occasions (Roose and Flower 2004). The rate of water uptake by the root system and the factors affecting the process of root growth are of fundamental interest in determining economic yield of a crop.

The drought tolerance capabilities of bambara groundnut have been characterised in many instances with commonly used growth indices such as, Leaf Area Index (LAI), Total Dry Matter (TDM) and yield (Collinson et al., 1996; Mwale et al., 2007a). Since monitoring root growth and distribution is both labour-intensive and expensive, attempts have been made to develop models to simulate the water use by the root system (King et al. 2003; Manschadi et al. 1998). Thus research work has prioritised the quantification of environmental factors through suitable experimental and modelling approaches. Physiologically-based mechanistic crop models are frequently employed to

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali
David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

estimate crop yields in variable climate studies, as they attempt to represent the major processes of crop environmental response.

The first dynamic crop model of bambara groundnut, BAMnut (Bannayan, 2001) followed the approaches of the CERES, family of models in which soil profile is divided into 3 layers and the root system restricted to the top and second layer. BAMFOOD project model (Cornelissen, 2005) used the water routine of the PALM model (Matthews, 2005) since no data were available on soil water content in the experiments used. It calculates the ratio between water supply and potential transpiration. The water supply component is influenced by the actual water content of the soil layers and the depth and distribution of the root system.

The present model, BAMGRO uses a simple approach to simulate the root growth, root distribution and soil water uptake of the crop under variable climatic conditions using the starting framework of wheat root model applied for pre anthesis period described by King et al., (2003) and modified for whole crop cycle in bambara groundnut. The model was primarily calibrated with glasshouse experiment (Tropical Crops Research Unit-2003), Nottingham, UK and validated for glasshouse experiments (TCRU-2007, 2008) and field experiment in Notwane, Botswana (2007-2008 season).

Therefore the objective of the present study were to simulate (1) root growth and root distribution, (2) the water uptake by the root system and (3) the soil water balance of the profile under variable climates in controlled environments and in the field.

MODEL DATA SETS

Briefly, datasets from glasshouse experiments in University of Nottingham, UK (TCRU) in 2003 and published information in King et al., (2003) were used to calibrate BAMGRO. The model was validated against independent data sets from Nottingham, UK (TCRU-2007, TCRU-2008) and field site in Notwane, Botswana.

The details of experimental design, plant sampling procedures, irrigation treatments and standard measurements for TCRU-2003 experiment have been previously explained in Mwale et al. (2007a) and TCRU-2007, TCRU-2008) and field experiments in Botswana in Karunaratne et al. (2010a 2010b).

Over the summer months of 2007 and 2008 (April to September), two contrasting bambara groundnut landraces; Uniswa Red (Swaziland) and S19-3 (Namibia) were grown in five glasshouses with each house having one Uniswa Red and one S19-3 plot under controlled temperature regimes. Two temperatures 23 ± 5^0 C (LT) and 33 ± 5^0 C (HT) were imposed in the five glasshouses. Soil moisture in each house was non-limiting with weekly irrigation to field capacity until harvesting up to 77 DAS in 2007 and up to 33 DAS in 2008. The treatments were allocated according to a split-plot design that combined two bambara groundnut landraces (Uniswa Red, S19-3) and two different temperature regimes (LT, HT) with each treatment replicated twice and thrice at low and high temperature, respectively, due to limited number of glasshouses.

During 2007 and 2008 growing season in TCRU experiments, soil moisture content in the soil profile was monitored in all plots using a PR2 probe. The PR2 probe measures the soil moisture at 10 cm, 20 cm, 30 cm, 40 cm, 60 cm, and 100 cm. Each plot has four access tubes, the average of the access tubes readings represent the mean amount of water in the soil for each plot. Measurements were taken weekly starting from 55 DAS. Unfortunately during TCRU- 2007 experiment, the PR2 broke down after 119 DAS so the measurements are unavailable between 119 and 168 DAS.

The experimental farm, Notwane (Botswana College of Agriculture) performed field experiments for set of landraces at three sowing dates: December 21, January 18 and February 1 in 2006/2007 growing season where a range of environmental conditions were considered. The experiment was conducted in a single split plot with three sowing dates in main plots and the landraces in sub plots, replicated four times. Neutron probe was used to determine the soil moisture content in the profile (up to 100 cm depth) of Botswana field experiments (2007-2008). However this paper explains the model validation results for Uniswa Red sown in December 2007 only.

MODEL DESCRIPTION

The BAMGRO-soil water module uses daily time-steps to simulate root growth, root distribution, root water uptake and soil water balance from sowing until maturity for different bambara groundnut landraces. A summarised detail of soil water module is described below.

The soil is represented as a one dimensional profile; it is homogeneous horizontally and consists of a number of soil layers. The total soil depth is assumed to be 1.5 m, divided into 15 soil layers each of 10 cm depth. This model computes the daily changes to root length and balance of soil moisture content for each soil layer due to rainfall and irrigation, vertical drainage, soil surface evaporation and root water uptake processes.

The root distribution is simulated according to Gale and Grigal (1987) as in Equation (1)

$$Y = 1 - \beta^d \quad (1)$$

Where, Y is fraction of root system accumulated from soil surface to depth d and β, parameter to describe root distribution with depth.

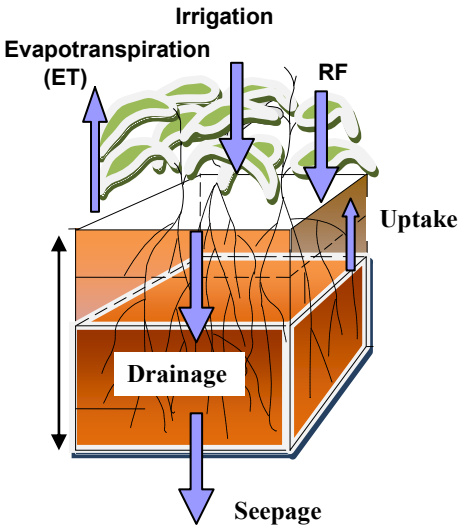


Figure 1: Inward and outward water flows for the different layers considered in the soil water balance.

The BAMGRO model follows the cereal root model (King et al., 2003) to calculate root length density (L_v; cm cm^{-3}). The value of β, σ, RW and Y (Equation (1)) are used to estimate the root length density of each 10 cm layer of the soil profile at each stage of crop growth according to Equation (3). King et al., (2003) reported that total root length (L m) is related to root dry weight (RW g m^{-2}) by the specific root weight (σ g km^{-1}). According to experimental evidence the value of σ significantly varies among landraces.

$$L_v = \left(Y_d - Y_{d-10}\right) \times \frac{RW}{\sigma} \qquad (2)$$

Where, L_v is root length density at 10 cm of soil layer at depth d (cm cm^{-3}); Y_d, cumulative fraction of roots at depth d; Y_{d-10}, cumulative root fraction at depth (d-10); RW, root weight (g m^{-2} d^{-1}) and σ, specific root weight (g km^{-1}).

Potential water extraction from the soil by roots equals potential transpiration. Its magnitude depends on the depth and density of the root system, and on the available soil water. This maximum uptake rate can be realized in a soil that is at FC and fully exploited by roots. When either soil moisture or root density is below optimum the actual water uptake is reduced relative to potential. Following (King et al., 2003), a generic function (Equation (4)) is used to predict water uptake as a fraction of total available water which is potentially available to uptake over the day. Thereby the potential water uptake for each 10 cm soil layer is estimated by Equation (4) based on the maximum available water in each layer on a daily basis.

$$U_{pot(t)} = \theta \times \left(1 - Exp\left(-k_w \times L_v\right)\right) \times E \qquad (3)$$

Where, k_w is 'root water capture coefficient' (cm^2); L_v, root length density of the soil layer (cm cm^{-3}); E, water capture parameter; θ, fraction of available water in soil layer and $U_{pot(t)}$, change in potential water uptake in layer i (cm d^{-1}).

The root water capture coefficient (k_w) is related to the resource uptake physiology especially molecular mechanism of water and nutrient transport across membranes and soil water transport mechanisms (King et al., 2003). Due to the lack of available data BAMGRO uses the value of two for k_w, similar to the value used for dry land barley (Gregory and Brown, 1989) and wheat (King et al., 2003). However, BAMGRO reduces k_w when the crop is exposed to temperatures below optimum ($T_{mean} < T_{opt}$).

The potential uptake by the whole root system is the accumulated capture by roots in each layer (1 to 15), assuming maximum possible rooting depth (1.5 m).

$$U_{pot(soil)} = \sum_{1}^{15} \frac{dU_{pot(i)}}{dt} \qquad (4)$$

Then actual water uptake is calculated using the potential values as given by Equation (4) considering Water Limited Growth (WLG) as in Equation (5) and Light Limited Growth (LLG). The actual water uptake from individual layer is calculated as a proportion of $U_{pot(soil)}$ and U_{actual} (Equation (7)).

$$WLG = \frac{\left(U_{pot(soil)} \times TE\right)}{SD} \qquad (5)$$

$$U_{actual} = \left(\min\left(\frac{LLG}{WLG}, 1\right)\right) \times U_{pot(soil)} \qquad (6)$$

$$U_i = \left(\frac{U_{actual}}{U_{soil}}\right) \times U_{pot(i)} \qquad (7)$$

Where, U_{actual} is actual rate of water uptake by roots in profile (mm d^{-1}); $U_{pot(soil)}$, potential rate of water uptake by roots in profile (mm d^{-1}); U_i, actual rate of water uptake by roots in layer i (mm d^{-1}) and TE, transpiration efficiency (g mm^{-1}).

As mentioned earlier, the model assumes 15 soil layers of 10 cm. Soil moisture is calculated separately for each of these (Figure 1). Layer 1 is the topmost layer dealing with calculation of potential evaporation from soil, addition from rainfall and irrigation, water extraction from crop component and vertical drainage. The subsequent layers deal with water extraction from roots and vertical drainage.

To estimate infiltration the model takes the simplified approach in which the top layer takes up water until it is at field capacity. Subsequent water is added directly to the second layer (Equation (9)). The drainage component and FC_1 are estimated according to Equation (9) and (10) respectively.

$$\frac{dWATER_1}{dt} = IRRI + RAIN - DR_1 - EVAPO - UP_1 \quad (8)$$

$$DR_1 = \max((WATER_1 - FC_1), 0) \quad (9)$$

$$FC_1 = FC_{soil} \times d_1 \quad (10)$$

Where, $WATER_1$ is soil moisture in layer 1(mm); *IRRI*, irrigation (mm d^{-1}); *RAIN*, precipitation (mm d^{-1}); UP_1, actual water uptake by roots in layer 1(mm d^{-1}); DR_1, drainage in layer 1(mm d^{-1}); d_1, depth of layer soil layer (cm) and FC_1, field capacity in layer 1 (cm).

If the soil moisture in the adjacent upper layer exceeds its *FC* the excess water flows to the subsequent layer of the soil profile. The major component in soil water balance in layers 2 to 15 is due to the uptake of water by the crop (Figure 2). In addition, excess water is directed to the next lower layer as the drainage fraction as explained in layer 1 (Equation (9)). The soil water balance in layer 2 to 15 is given by Equation (11).

$$\frac{dWATER_i}{dt} = DR_i - DR_{i+1} - UP_i \quad (11)$$

Where, $WATER_i$, is soil moisture in layer *i* (mm); UP_i, actual water uptake by roots in layer *i* (mm d^{-1}) and DR_i, drainage in layer *i* (mm d^{-1}).

SIMULATION RESULTS

The model was validated only for soil moisture due to the unavailability of root growth data. Mainly the model was compared with experimental soil moisture data sets from glasshouse experiments in 2007 and 2008 and the Botswana 2007/2008 season. The available soil moisture is assumed to be the net remaining after water uptake, vertical drainage and evapotraspiration. Therefore the simulation results of root growth and distribution though the profiles are shown as they are connected to soil water uptake component (Figures 2 and 3) at LT and HT in 2007 and 2008 respectively.

Comparison between simulated and observed soil moisture content (mm) for four soil layers (layer 1, 10 cm; layer 2, 20 cm; layer 3, 30 cm; layer 10, 100 cm) in glasshouse experiments during summer months of 2007 and 2008 are shown in Figure 2 and 3 respectively for UniswaRed.

The model was able to simulate the reduction in soil moisture content (mm) correctly due to the drought (2007, 77 DAS; 2008, 33 DAS). However the predicted soil moisture content (mm) in deeper layers was heavily under estimated, particularly under high temperature (33 ± 5 ^{0}C) thus indicating over estimation of losses of the water from the layer 10 (100 cm). A similar trend was observed for the variation of soil moisture content (mm) for S19-3 (data not shown). However the model generally over estimated the soil moisture content in the 2008 glasshouse experiment in which the drought was imposed at 33 DAS.

BAMGRO-soil water module simulates the soil moisture variation the soil profile of field sites in Notwane, Botswana with less deviation from measured values (MAE is ± 22.8 mm) (Figure 4). However, there is an over estimation especially towards the end of the growing season.

DISCUSSION

The BAMGRO-soil water module provides a framework for predicting root growth, water uptake and soil water balance for bambara groundnut landraces grown under drought stress conditions. Generally the model over estimates the soil moisture content in upper soil layers and it is heavily under estimated at deeper layers. There are several possibilities for these discrepancies.

According to the model, the vertical distribution of roots (*Y*) as described by β, can influence the water uptake capacity of the crop. The general over estimation of the simulation results from the present study indicates that the parameter values used for β are higher. As this was derived from the crop grown at optimum temperature condition (28 ± 5 ^{0}C), a general reduction of β can be hypothesised under heat and cold stress. However, this has not been considered within the model due to the lack of information on changes of β under different temperature stress conditions. In addition, the use of a single value of β from sowing to harvesting, does not consider the root distribution with age. The value used for (k_w) is also not very specific to bambara groundnut and this may contribute towards the poor correlation of model simulations with the measured data. In addition, several soil physical factors influence root growth and distribution that are not considered in BAMGRO (eg. hydraulic conductivity, soil porosity).

The model clearly indicates the relationship of root length density (L_v) and water uptake (Equation (6). However the variation of L_v under drought, heat and cold stress for bambara groundnut is unknown. This provides a weak link within the model. Husain et al. (1990) indicates that both L_v and rooting depth of faba bean (*Vicia faba* L.) grown under drought stress were significantly higher than regularly irrigated crops. A study focussed on investigation of L_v and water uptake revealed that some cereal species consistently had five to ten times the total root length of grain legumes and a higher correlation with maximum rooting depth than the root length density (Hamblin, 1987).

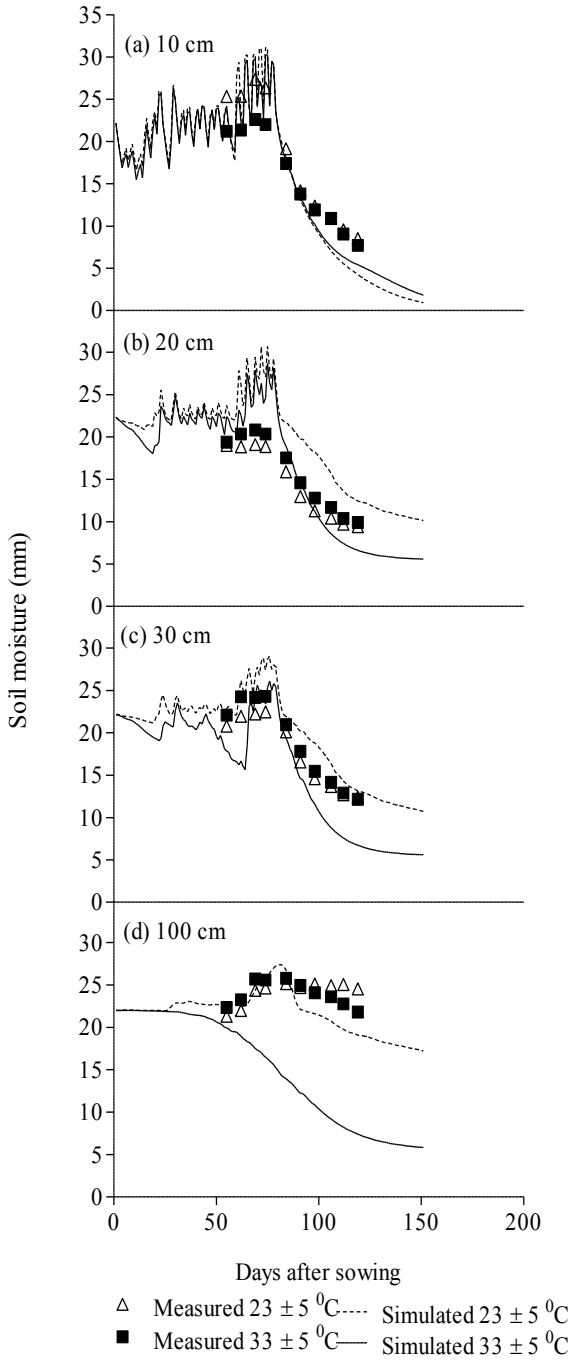


Figure 2. Soil moisture variation with days after sowing at top 10 cm (a), 20 cm (b), 30cm (c) and 100 cm (d) layers for Uniswa Red grown under 23 ± 5^0 C and 33 ± 5 ^{0}C in Glasshouse experiments during 2007. Model data sets were provided by Ibraheem Alshareef.

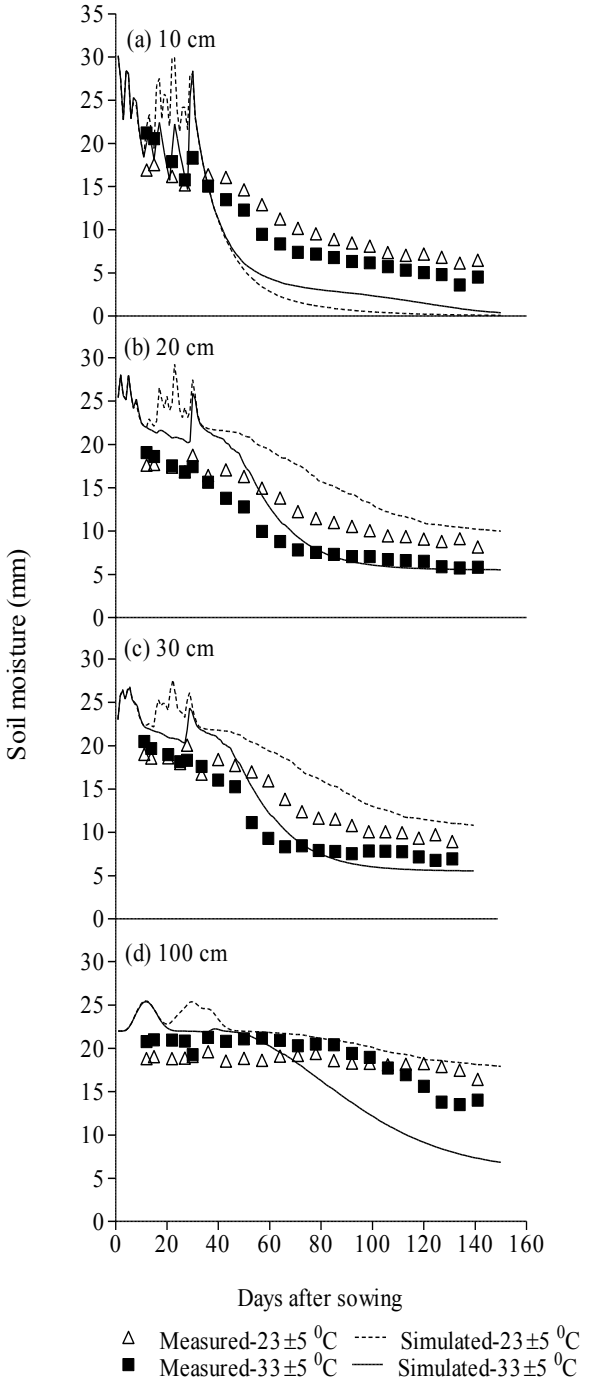


Figure 3. Soil moisture variation with days after sowing at top 10 cm (a), 20 cm (b), 30cm (c) and 100 cm (d) layers for Uniswa Red grown under 23 ± 5 ^{0}C and 33 ± 5 ^{0}C in Glasshouse experiments during 2008. Model data sets were provided by Stanley Noah.

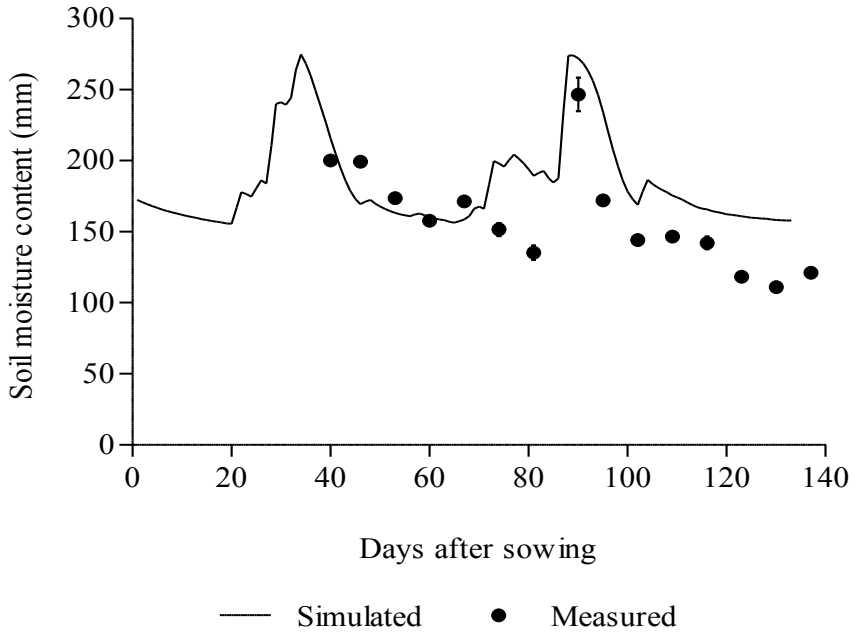


Figure 4. Soil moisture variation of Uniswa Red with days after sowing through the soil profile in Botswana field site during the growing season 2007-2008. The soil moisture was measured using neutron probe. Model data sets were provided by, Abu Sasey Botswana College of Agriculture, Botswana).

CONCLUSIONS AND FUTURE WORK

A model for simulation of root growth, distribution and plant water uptake developed by this study is a modified approach of a simple wheat model. The functions and relationships were derived from the glasshouse experiments at TCRU, University of Nottingham, UK. The testing the model performance was primarily done with the experimental observations from glasshouse experiments with early and late drought and field trials in Botswana. BAMGRO-soil water model predicts the soil water content for two bambara groundnut landraces (Uniswa Red, S19-3) realistically but needs further improvement in calibration of k_w, β and ε_s.

REFERENCES

Bannayan, M. 2001. "BAMnut: a crop simulation model for bambara groundnut". *Agricultural Sciences and Technology*, 15, 101-110.

Gale, M. R. and D. F. Grigal 1987. "Vertical root distribution of northern tree species in relation to successional status". *Canadian Journal of Forest Research,* 17, 829-834.

Collinson, S. T.; S. N. Azam-Ali; K. M. Chavula; and D. A. Hodson, 1996. "Growth, development and yield of bambara groundnut (*Vigna subterranea*) in response to soil moisture". *Journal of Agricultural Sciences (Cambridge)*, 126, 307-318.

Cornellisen, R. L. E. J. 2005. "Modelling variation in the physiology of bambara groundnut (Vigna subterranea (L) Verdc.)". PhD thesis School of Biosciences. Sutton Bonington, University of Nottingham, United Kingdom.

Karunaratne, A. S.; S.N. Azam-Ali; N. M. J. Crout; S. Mayes; P. Steduto; and G. Izzi, 2009. "Modelling the response of bambara groundnut: A key underutilized crop in agricultural systems". Proceedings of 23rd European Conference on Modelling and Simulation: ECMS, Madrid, Spain, 841-847.

Karunaratne, A.S.; S. N. Azam-Ali; I. Al-Shareef; A. Sesay; S.T Jørgensen;. and N. M. J Crout 2010a. "Modelling the canopy development of bambara groundnut". *Agricultural and Forest Meteorology*, (in press).

Karunaratne, A.S.; S. N., Azam-Ali; I. Al-Shareef; A. Sesay; and N. M. J. Crout 2010b. "Modelling the dry matter production and partitioning of bambara groundnut for abiotic stress". (Under review; European Journal of Agronomy).

King, J., A.; R. G. Gray; I. Bradely; Binham; J. Foilks; P. Gregory; and D. Robinson (2003). "Modelling cereal root system for water and nitrogen capture: Towards an economic optimum". *Annals of Botany*, 91, 383-390.

Linnemann, A. R. and S. N. Azam-Ali 1993. "Bambara groundnut *(Vigna subteranea)*".In: Underutilised crops Pulses and Vegetables". London, Chapman and Hall.

ManschadiI, A.M.; J. Aurborn; H. Stutzel; W. Gobel; and M.C. Saxena 1998. "Simulation of faba bean (*Vicia faba* L.) root system development under

Mediterranean conditions". *European Journal of Agronomy*, 9, 259-272.
Matthews, R. B. 2006. "PALM: An agent-based spatial model of livelihood generation and resource flows in rural households and their environment". *Ecological modelling,* 194, 329-343.
Mwale, S. S.; S. N. Azam-Ali; and F. J. Massawe 2007a. "Growth and development of bambara groundnut (*Vigna subterranea*) in response to soil moisture 1. Dry matter and yield". *European Journal of Agronomy*, 26, 345-353.
Roose, T. and A. C. Flower 2004. "A model for water uptake by plant roots". *Journal of Theoretical Biology*, 228, 155-171.

AUTHOR BIOGRAPHIES

ASHA KARUNARATNE studied Crop Science at Peradeniya University, Sri Lanka. She is a lecturer to the Faculty of Agricultural Sciences, University of Sabaragamuwa, Sri Lanka. While she was teaching at Sabaragamuwa, University, completed M,Phil in crop physiology in tea plantation. In 2005, she was awarded a commonwealth scholarship for PhD leading to modelling bambara groundnut at University of Nottingham, UK. Currently, contributes to FAO-AquaCrop program through consultation of calibrating of bambara groundnut. Her email is asha.karunaratne@yahoo.co.uk.

NEIL CROUT is a Professor in Environmental Modelling in the Agriculture and Environmental division at Nottingham University, UK. His research interests are working in the development and application of simulation models of environmental systems, in particular (spatial) models of the transfer of (elemental) trace contaminants. He is also interested in the predictive performance of environmental models, especially in relation to their complexity and has exciting projects investigating the development of 'parsimonious' models. His email address is Neil.Crout@Nottingham.ac.uk .

SAYED AZAM-ALI graduated in Plant Science from the University of Wales, Bangor and then completed his PhD in Environmental Physics at the University of Nottingham in 1983. His research interests include the ecophysiological responses of tropical crops to abiotic stresses and the agroecological evaluation of underutilised crops. He was until recently Director of the University of Nottingham, Tropical Crops Research Unit and is now Vice-President (Research) University of Nottingham Malaysia Campus (sayed.azam-ali@nottingham.edu.my).

A STUDY ON LAMARCKIAN AND BALDWINIAN LEARNING ON NOISY AND NOISELESS LANDSCAPES

Cheng Wai Kheng
University of Nottingham, Malaysia Campus
School of Computer Science
Jalan Broga 43500 Semenyih, Selangor, Malysia.
E-mail: cheng-wai.kheng@nottingham.edu.my

Meng Hiot Lim
Nanyang Technological University
School of Computer Engineering
Block N4, 2b-39, Nanyang Avenue, Singapore 639798
E-mail: emhlim@ntu.edu.sg

Siang Yew Chong
ASAP Research Group
University of Nottingham, Malaysia Campus
School of Computer Science
Jalan Broga 43500 Semenyih, Selangor, Malysia.
E-mail: siang-yew.chong@nottingham.edu.my

KEYWORD
Optimisation, Landscape, Memetic Computing, Baldwinian effect, Lamarckian evolution.

ABSTRACT

Lamarckian evolution and Baldwinian effect are two typical replacement schemes of the individual learning process in Memetic Computing. In this paper, we perform a comprehensive study on the behaviour of Lamarckian evolution and Baldwinian effect in noisy and noiseless continuous optimisation problems. The output of this comprehensive study shows that Lamarckian lifetime learning performs better on noiseless problem, while Baldwinian learning performs better on noisy problems.

INTRODUCTION

Memetic Computing (MC) represents one of the prominent emerging approaches of Computational Intelligence introduced in the last decades. Within the field, one of the most successful methodologies is the memetic algorithm (MA) that has materialised in the form of hybrid evolutionary search.

Typically, MA involves some lifetime learning within the Darwinian Evolution Cycle. MA may involve single or multiple lifetime learning procedures (Ong and Keane 2004). (Krasnogor and Smith 2005). When designing MA, several issues are often considered, from how much computational budget to allocate for lifetime learning at individual and generation level, which individuals of the population that should undergo lifetime learning to what form of lifetime learning methodology to employ (Nguyen, Ong et al. 2009). Among the design issues, one that has received little attention over the years is the type of learning scheme to use, i.e., Lamarckian or Baldwinian learning. In black box optimisation, the underlying problem characteristics are generally unknown beforehand. Hence, it is well established today according to the "No Free Lunch" theorem that no single algorithm, including, i.e., Lamarckian or Baldwinian learning, has clear advantage over another on all classes of problems except random walk (Wolpert and Macready 1997). In fact, the performance of MA is highly dependent on the problem landscape. Information regarding MA and connectivity analysis can be found in (M. N. Le, Y.S. Ong et al. 2009).

The motivation of this paper is to study the effect of Lamarckian and Baldwinian learning in MA for solving noisy and noiseless optimisation problems. We first analyse the behaviour of Lamarckian learning in the context of line search, which leads to the hypothesis that Lamarckian performs poorly in noisy problem, and then validate the hypothesis with simulation results.

The remaining paper is organised as followed. The paper starts with problem definition. Next, brief definitions of Lamarckian and Baldwinian learning and their comparisons are presented. Subsequently, the definition of Multiple Evaluation Method in the context of Baldwinian learning and an analysis of Lamarckian learning in the context of non-linear programming, followed by some empirical results and discussions, and finally ends with conclusion and future work.

LITERATURE REVIEW

A. Problem Definition

Stochastic global search such as evolutionary algorithm (De-Jong 2006) has been used in solving non-linear programming problem for several decades, which takes the form of:

$$\begin{aligned} &\arg\min_{\mathbf{x}} \ f : \mathbf{x} \to \mathbf{R}, \mathbf{x} \in \mathbf{R}^d \\ &\text{subjected } \ \mathbf{low} \le \mathbf{x} \le \mathbf{high} \end{aligned} \tag{1}$$

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

where **x** is a parameter vector with size *d* and each variable in the **x** is restricted within the range of **low** and **high**.

In this study, the noise introduced during the optimisation process itself. This situation can arise when vector **x** refers to an output of devices, and the optimisation is real-time. Noise may be introduced during the transmission between the device and the computer, loss of precision when converting from analogue to digital signals, or even the inconsistency of the device itself. Such noises are usually assumed to be normally distributed.

We simulated this noise in standard benchmark functions through perturbations of the input vector **x**. Optimisation problem in Eqn. (1) is then transformed into:

$$\begin{aligned} &\min f(\mathbf{x}+\boldsymbol{\delta}) \\ &\text{subjected } \mathbf{low} \le \mathbf{x} \le \mathbf{high} \\ &\delta^{(i)} \in \gamma, \boldsymbol{\delta} = \begin{bmatrix} \delta^{(1)} \\ \vdots \\ \delta^{(d)} \end{bmatrix} \end{aligned} \tag{2}$$

δ_i is generated randomly according to certain density function γ. In this paper, Gaussian density function is used.

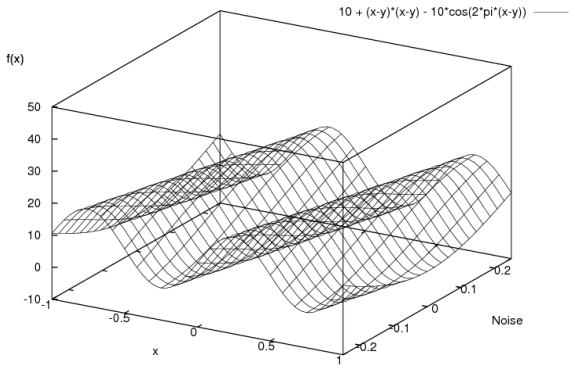


Fig 1: Noisy Rastrigin

Such perturbations transform the underlying problem landscape from static to dynamic. As an evaluation, the same input vector **x** generates different responses from object function *f* when it is to be evaluated several times. In this paper, we refer Eqn. (1) as noiseless problem and Eqn. (2) as noisy problem. Figure 1 demonstrates the noise effect shifting the landscape with noise range between -0.2 to +0.2.

B. Lamarckian and Baldwinian Lifetime Learning

Lamarckian lifetime learning involves the solution (both vector **x** and fitness) to be replaced during the individual learning process whereas Baldwinian replaces only the fitness of the solution instead of vector **x**. Let y be the response of **x**, $y = f(\mathbf{x})$. A solution space is then defined as an unordered pair of solution and its response.

$$\mathrm{S} = \mathrm{XY} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \ldots . (\mathbf{x}_n, y_n)\} \tag{3}$$

The individual improvement process defines the neighbouring structure, X^N.

$$\begin{aligned} \mathrm{X}^N = N(\mathbf{x}) = \{&\mathbf{x}_1, \mathbf{x}_2 \ldots ., \mathbf{x}_p \mid \\ &\forall_{i \le p} \mathbf{x}_i = \mathbf{x} + \boldsymbol{\delta}_i\} \end{aligned} \tag{4}$$

In general, the neighbourhood can be defined by any random generator with neighbourhood size σ, or special neighbourhood that is bound to specific function.

Subsequently, the replacement policies (with loss of generality, assuming a minimisation problem) proceeds as follows:

Lamarckian replacement policy:

$$\begin{aligned} L(\mathbf{x}, y) \to (\mathbf{a}, f(\mathbf{a})), \{\mathbf{a}, \mathbf{a}' \in \mathrm{X}^N \mid \\ f(\mathbf{a}) < f(\mathbf{a}'), \mathbf{a} \ne \mathbf{a}'\} \end{aligned} \tag{5}$$

Baldwinian replacement policy:

$$\begin{aligned} B(\mathbf{x}, y) \to (\mathbf{x}, f(\mathbf{a})), \{\mathbf{a}, \mathbf{a}' \in \mathrm{X}^N \mid \\ f(\mathbf{a}) < f(\mathbf{a}'), \mathbf{a} \ne \mathbf{a}'\} \end{aligned} \tag{6}$$

Note that the fitness function is not changed. Baldwinian lifetime learning only alters the function responses.

To date, numerous studies have been conducted on Baldwinian effect and Lamarckian evolution. (Julstrom 1999) studied the effect of applying different frequencies of Baldwinian or Lamarckian on a four-cycle combinatorial optimisation problem. Their empirical results showed that Baldwinian performed worst on average compared to pure Darwinian and Lamarckian. Mayley discussed on the same issue in (Mayley, G. 1996) and commented that if the solution in local improvement cannot be found by genetic operation within certain steps, the Baldwinian effect may mislead the search.

While the above discussed on the drawbacks of Baldwinian learning, Whitley et. al suggested that Baldwinian learning brings about the advantage of escaping from local optima since Lamarckian may possess the drawback of leading to premature convergence (Whitley, Gordon et al. 1994). However, Baldwinian learning generally converged slower and poorer results compared to Lamarckian. Castillo et al studied the same approach in Evolutionary Neural

Networks and concurred Whitley's finding (Castillo, Arenas et al. 2006).

METHODOLOGY

Although there have been many comparisons done between Lamarckian and Baldwinian, they are limited to problem in the absence of noise. The motivation of this paper is to extend the investigation of Baldwinian effect and Lamarckian evolution in both noisy and noiseless landscape.

A. Multiple Evaluation Method as Baldwinian learning

When an underlying optimisation problem is plagued with noise, the use of resampling in the search like the Multiple Evaluation Method (MEM) (Tsutsui and Ghosh 1997) (Tsutsui and Ghosh 1997) may be useful, which takes form as followed:

$$B(\mathbf{x}, y) \to (\mathbf{x}, \mu), \{\mathbf{x}_i \in \mathrm{X}^N \mid \mu = \frac{1}{p}\sum_{\mathbf{x}_i \in \mathrm{X}^N} f(\mathbf{x}_i)\} \quad (7)$$

Since the process only replaces the fitness with the mean, which in general, qualifies as a form of Baldwinian lifetime learning. Instead of taking the average, there are other forms of MEM, such as taking the worst fitness in the neighbouring solutions. More information regarding MEM can be found in (Y. S. Ong, P. B. Nair et al. 2006)

B. Analysis of Lamarckian in Noisy Optimisation Problem

Deterministic individual learning scheme such as Quasi-Newton, Gradient Descent, etc. that depends on the gradient information as guidance would be misled by the noise, especially when Lamarckian learning is used. When noise is introduced to the landscape, Lamarckian spends the resources trying to adapt to the noise instead of the actual landscape, which would cause the algorithm to converge slower for non convex problem. Nevertheless, Lamarckian should always converge on better results if the global stochastic search manages to bring the solution to the global optimum basin. After all, the basin that consists of the global optimum itself is a convex problem.

On the other hand, by countering the noise with re-sampling, Baldwinian replacement policy gains advantage by having fitness value closer to the actual function. This increases the selection pressure on "good" parents while teaming with fitness proportional selection mechanism, such as stochastic universal sampling.

Without hesitation, the following section analyses how noise could mislead the heuristics and reduce the efficiency of Lamarckian learning in noisy multimodal optimisation problem with a series of simple respond graphs.

A line search is chose as a deterministic search in this paper because it is the core component in most of the gradient-based search, thus, this would be a solid starting point to investigate. With the assumption of non-repeating space travelling, the line search takes form of:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k \quad (8)$$

$\mathbf{x}_{k+1}$, the next solution to be evaluated, is determined by the current solution $\mathbf{x}_k$, with the increment of constant value α_k according to the searching direction $\mathbf{d}_k$.

$$\mathbf{d}_k = \begin{bmatrix} \frac{f(\mathbf{z}_1)}{\varepsilon} \\ \frac{f(\mathbf{z}_2)}{\varepsilon} \\ \vdots \\ \frac{f(\mathbf{z}_d)}{\varepsilon} \end{bmatrix}. \quad (9)$$

Let $z_m^{(j)}$ indicates the j^{th} dimension of the vector $\mathbf{z}$, and $x_k^{(j)}$ indicates the j^{th} dimension of vector $\mathbf{x}$ at step k. $z_m^{(j)}$ is defined as:

$$z_m^{(j)} = \begin{cases} x_k^{(j)} + \varepsilon \text{ iff } j = m \\ x_k^{(j)} \text{ otherwise} \end{cases} \quad (10)$$
subjected $0 < m \le d$

It is possible that the noise changes the gradient information. As the effect, the searching direction is not determined by the gradient of $\mathbf{x}_k$, instead, it is determined by $\mathbf{x}_k+\boldsymbol{\delta}_k$, where each dimension of $\boldsymbol{\delta}_k$ is drawn independently according to a density function. $z_m^{(j)}$ are then transformed to:

$$z_m^{(j)} = \begin{cases} x_k^{(j)} + \delta_k^{(j)} + \varepsilon \text{ iff } j = m \\ x_k^{(j)} + \delta_k^{(j)} \text{ otherwise} \end{cases} \quad (11)$$
subjected $0 < m \le d$

For the sake of simplicity, we illustrate the misleading searching direction by assuming every respond is drawn with a single δ noise value in a single dimension problem. Figure 2 illustrates the searching direction has been reversed by the noise. Point A and B are the same solution vector where A is evaluated with the absent of noise, and B with the present of noise. Suppose the searching starts with point A. The gradient information is decreasing towards the right. Given the same solution vector, the noise shows the gradient is decreasing towards left.

Second, assuming the noise is large and is divisible by α, it is possible that a line search starts with position

that is m steps before the current solution $\mathbf{x}_k$, providing that $\mathbf{x}_{k-m}=\mathbf{x}_k+\boldsymbol{\delta}_k$, which violates the assumption of non-repeating space travelling.

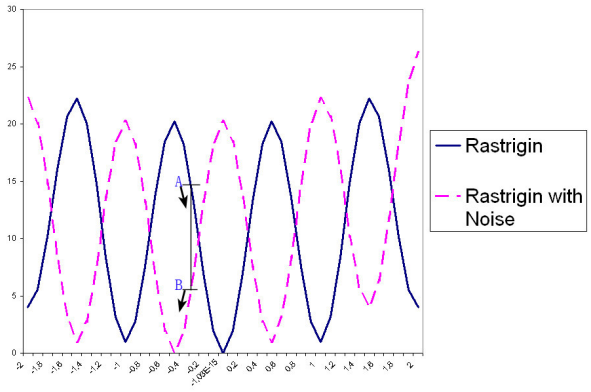


Fig 2: Noise in Searching Direction Determination

Lastly is the live lock effect, which causes the line search to oscillate between the same local optimum that causes by the noise. The term live lock is introduced in concept of concurrency, where two processes are given resources without making any progress. However, this effect happens when the noise is periodic only. In Figure 3, point A is a local maximum that moves decreasingly to reach local minimum point B in step k. In k+1, the respond of B is raised up to point C with the effect of noise. The line search moves decreasingly from point C to point D, which in fact, is the local maximum point A when the noise is removed.

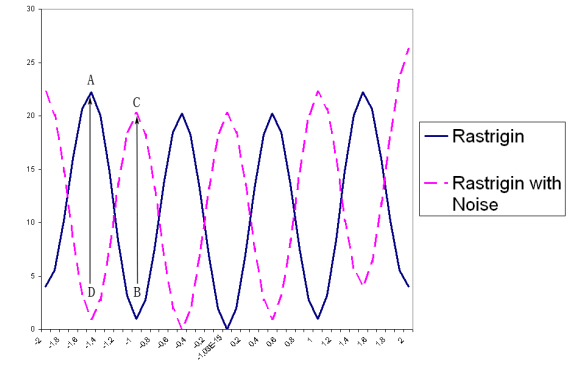


Fig 3: Line search oscillating between local optima, and the same local optimum causes by noise.

EMPIRICAL RESULTS AND DISCUSSIONS

We investigated the above analysis by running simulations on two sets of benchmark functions. First set is the noiseless function. Second set is the noisy functions with 0.001 noise level. The experimental setting is given in Table 1. Table 2 outlines the benchmark functions we use.

The choice of Lamarckian lifetime improvement budget is based on the size of k in Eqn. (8). Broyden-Flecher-Goldfarb-Shanno (BFGS) takes two gradient evaluations in every increment of k. Thus, the budget b formula is given as:

$$b = k[(2 \times d) + l] \tag{12}$$

where d is the problem dimension and l is a variable to determine step size α.

Table 1: Experimental Parameters

Global Search: Real Coded Genetic Algorithm
Lamarckian Property: Local Improvement Method: Broyden–Fletcher–Goldfarb–Shanno (BFGS) Lifetime Improvement Frequency: 0.1 Lifetime Improvement Budget (FFE): 130
Baldwinian Property: Local Improvement Method: Multiple Evaluation Method stated in Eqn. (7) Lifetime Improvement Frequency: 1.0 Lifetime Improvement Budget (FFE): 13
Crossover Rate: 1.0
Mutation Rate: 0.03
Mutation Type: Gaussian Mutation with 1.0 Radius
Selection: Stochastic Universal Sampling
Replacement: Ranking
Termination Condition: 300,000 Function Evaluations
Population Size: 50

For the fair comparison, summation of function evaluation on each generation is equal for both approaches. The noise level σ is set to 0.001 * (function upperbound - function lowerbound). Each problem is repeated for 30 independent runs. The mean best over function calls are recorded in Figures 4 and 5.

The simulation results agree with our findings. On one hand, Lamarckian converges on better solutions on average for noiseless problems as shown in Figure 4. On the other hand, Figure 5 shows that Baldwinian is significantly better than Lamarckian on noisy problems. Baldwinian's convergence speed is significantly higher compared to Lamarckian in all test cases on average. However, in the case of Rastrigin, the simulation result shows that Baldwinian converges faster than Lamarckian, but not as optimal as its converged solutions. The convergence speed for Baldwinian is faster than Lamarckian because the re-sampling process increases the chances of selecting good parents to reproduce. However, towards the end of the search, the effect of Baldwinian diminishes as the selection pressure is not as important as it is at the beginning of the search. Since the global search narrows down the searching region in the global optimum basin, a line search is more appropriate in this situation.

Table 2: Benchmark Functions

Name	Range	Equation	Dim (d)
Ackley	[-32, 32]	$F_{Ackley}(\mathbf{x}) = -20 \cdot \exp(-0.2\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2}) - \exp(\frac{1}{d}\sum_{i=1}^{d}\cos(2\pi x_i)) + 20 + e$	30
Hybrid F8F2	[-3, 1]	$F_{HybridF8F2}(\mathbf{x}) = \sum_{i=2}^{n-1}(\frac{z(x_{i-1}, x_i)}{4000}) - \cos(z(x_{i-1}, x_i)) + 1.0 + (\frac{z(x_n, x_1)}{4000}) - \cos(z(x_n, x_1)) + 1.0$ $z(x, y) = 100(x^2 - y)^2 + (x-1)^2$	30
Rastrigin	[-5, 5]	$F_{Rastrigin}(\mathbf{x}) = 10n + \sum_{i=1}^{d} x_i^2 - 10 \cdot \cos(2 \cdot x_i)$	30
Griewank	[-600, 600]	$F_{Griewank}(\mathbf{x}) = \frac{1}{4000} \cdot \sum_{i=1}^{d}(x_i - 100)^2 - \prod_{i=1}^{d}\frac{x_i - 100}{\sqrt{i}}) + 1$	30
Schwefel	[-3.142, 3.142]	$F_{Schwefel}(\mathbf{x}) = \sum_{i=1}^{d}(\sum_{j=1}^{i} x_j)^2$	30

CONCLUSION AND FUTURE WORK

In this paper, we analysed the behaviour of Lamarckian evolution and Baldwinian effect in the absence and presence of noise in continuous optimisation problems. The comprehensive study shows that Lamarckian performs well in noiseless problems whereas Baldwinian performs better in noisy problems. The empirical results are consistent with our previous analysis of line search. Results from our studies have the implication on the specific use of replacement policy for problem having specific characteristic (i.e. noisy or noiseless continuous optimisation problem).

This study is at first step towards further understanding of MC for problem solving. Future studies would include investigations of Lamarckian and Baldwinian approach to other problems such as combinatorial optimisation and multi-objective optimisation.

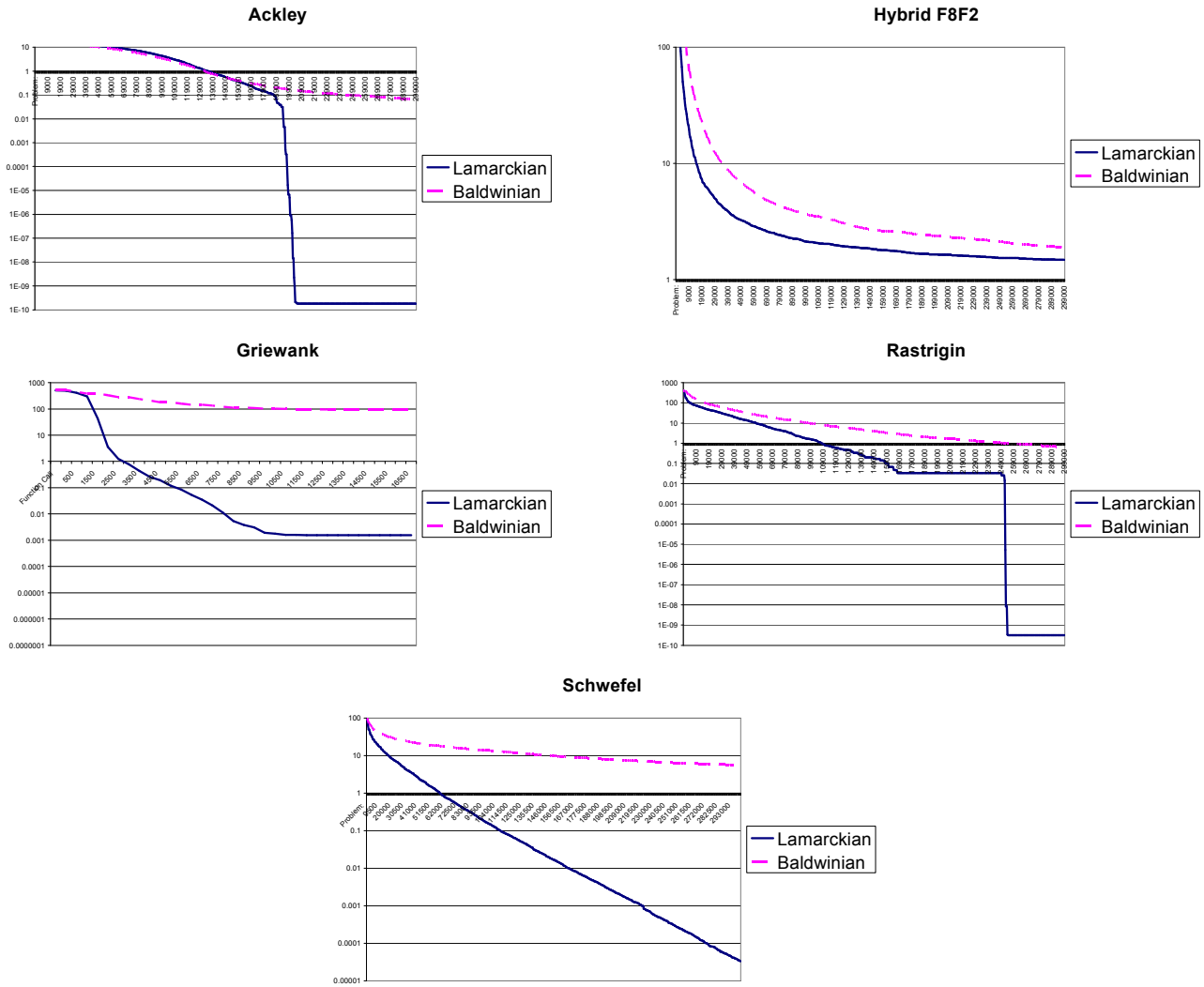


Figure 4: Convergence Speed between Baldwinian and Lamarckian on Noiseless Problem

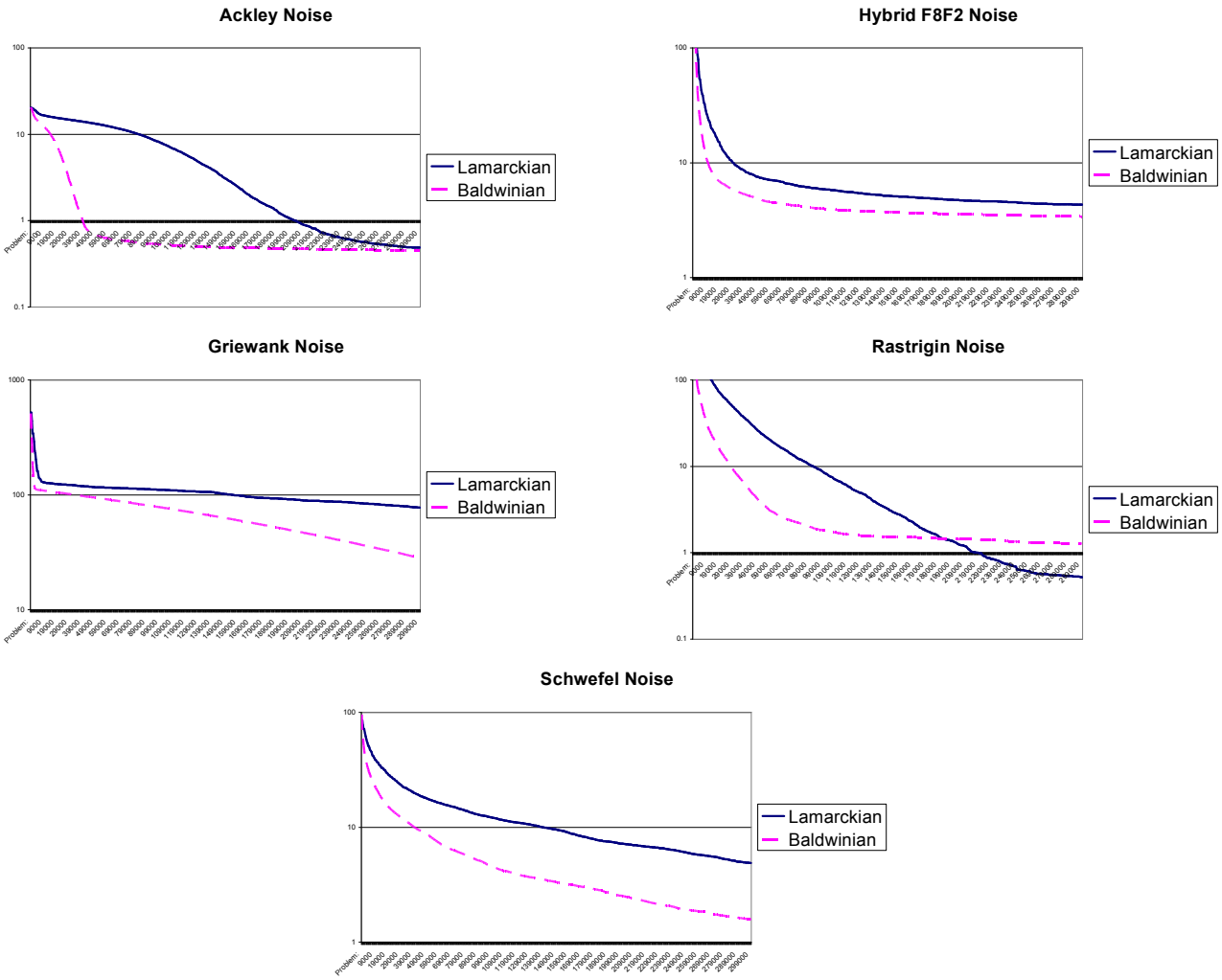


Figure 5: Convergence Speed between Baldwinian and Lamarckian on Noisy Problem with θ=1E-03.

ACKNOWLEDGEMENT

The authors acknowledge the funding support of Singapore Technologies (ST) Engineering Pte Ltd.

REFERENCES

Castillo, P. A., M. G. Arenas, et al. (2006). Lamarckian evolution and the Baldwin Effect in Evolutionary Neural Networks.

De-Jong, K. (2006). Evolutionary Computation: A Unified Approach, Prentice Hall of India Private Limited.

Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization & Machine Learning, Addison-Wesley.

Julstrom, B. A. (1999). Comparing Darwinian, Baldwinian, and Lamarckian Search in a Genetic Algorithm for the 4-Cycle Problem. Late Breaking Papers at the 1999 Genetic and Evolutionary Computation Conference**:** 134--138.

Krasnogor, N. and J. Smith (2005). "A tutorial for competent MAs: model, taxonomy, and design issues." Evolutionary Computation, IEEE Transactions on **9**(5): 474-488.

Mayley, G. (1996). "Landscapes, Learning Costs and Genetic Assimilation." Evolutionary Computation **4**: 213-234.

Moscato, P. (1989). On Evolution, Search, Optimization, Genetic Algorithms and Martial Arts: Towards MAs, Caltech Concurrent Computation Program.

M. N. Le, Y. S. Ong, Y. Jin & B. Sendhoff. (2009). 'Lamarckian MAs: local optimum and connectivity structure analysis', Memetic Computing Journal, Vol. 1, No. 3, pp. 175-190.

Nguyen, Q. H., Y. S. Ong, et al. (2009). "A Probabilistic Memetic Framework." Evolutionary Computation, IEEE Transactions on **13**(3): 604-623.

Ong, Y. S. and A. J. Keane (2004). "Meta-Lamarckian learning in MAs." Evolutionary Computation, IEEE Transactions on **8**(2): 99-110.

Tsutsui, S. and A. Ghosh (1997). "Genetic algorithms with a robust solution searching scheme." Evolutionary Computation, IEEE Transactions on **1**(3): 201-208.

Whitley, L. D., V. S. Gordon, et al. (1994). Lamarckian evolution, The Baldwin Effect and Function Optimization.

Proceedings of the International Conference on Evolutionary Computation. The Third Conference on Parallel Problem Solving from Nature: Parallel Problem Solving from Nature, Springer-Verlag.

Wolpert, D. H. and W. G. Macready (1997). "No Free Lunch Theorems for Optimization." IEEE Transactions on Evolutionary Computation **1**(1): 67-82.

Y. S. Ong, P. B. Nair and K. Y. Lum (2006). 'Max-Min Surrogate-Assisted Evolutionary Algorithm for Robust Design', IEEE Transactions on Evolutionary Computation, Vol. 10, No. 4, pp. 392-404.

AUTHORS BIBLIOGRAPHY

CHENG WAI KHENG is a Postgraduate Student of School of Computer Science, University of Nottingham, Malaysia Campus. He is a member of Intelligent Decision Support Systems Research Group, University of Nottingham, Malaysia. His current research interests include the application of Hyper/Meta Heuristic, Evolutionary Computation and Optimisation. He received his B.Sc in Computer Science from University of Nottingham, Malaysia in 2007.

MENG HIOT LIM is currently an Associate Professor at the School of Electrical & Electronics Engineering, Nanyang Technological University. He received his B.Sc., M.Sc. and PhD from the University of South Carolina. He holds concurrent appointment as Deputy Director of both the Centre for Financial Engineering, a multi-disciplinary research centre anchored at the Nanyang Business School and the highly regarded M.Sc. in Financial Engineering program. His research interests include memetic computation, computational intelligence, combinatorial and continuous domains optimization, evolvable hardware systems, computational finance and graph theory. He is the Managing Editor-in-Chief of Memetic Computing Journal, the Editor-in-Chief of a Book Series on Adaptation, Learning, and Optimization published by Springer-Verlag and an Associate Editor of IEEE Transactions on SMC – Part C. Dr Lim has also guest-edited several special issues in international journals. His research has been funded by agencies such as the NSTB (now known as A*Star), Temasek Defense Strategic Initiatives and industries which include ST Engineering, ST Yellow Pages, Seiko Instruments, and Boeing Phantom Works. He played major role in international conferences, such serving as INISTA-2009 General Chair, ISIC-09 Publicity Chair, CEC-7 Logistic Chair, etc.; and currently serving as the General Chair of ICCP-2010 in China and SEMCCO-2010 in India.

SIANG YEW CHONG is an Assistant Professor with the School of Computer Science, University of Nottingham, Malaysia Campus, a member of the Automated Scheduling, Optimization and Planning (ASAP) Research Group, University of Nottingham, UK, and an Honorary Research Fellow with the School of Computer Science, University of Birmingham, UK. He was a Research Associate with the Centre of Excellence for Research in Computational Intelligence and Applications (CERCIA), University of Birmingham, UK, in 2007. He is an Associate Editor of IEEE Transaction on Computational Intelligence and AI in Games. His major research interests include evolutionary computation, neural networks, and evolutionary game theory. Dr. Chong received the Ph.D. degree in Computer Science from the University of Birmingham, UK, in 2007, and was awarded the 2009 IEEE Computational Intelligence Society Outstanding Ph.D. Dissertation Award for his work on theoretical frameworks for performance analysis in co-evolutionary learning.

MULTILAYERED DEVS MODELING AND SIMULATION IMPLEMENTATION VALIDATION ON A CONCRETE EXAMPLE: PREDICTION OF THE BEHAVIOR OF A CATCHMENT BASIN

Emilie Broutin
Paul Bisgambiglia
Jean-François Santucci

University of Corsica, SPE
CNRS, UMR N°6134
Campus Grimaldi
20250 CORTE
Email : {broutin | bisgambi | santucci}@univ-corse.fr

KEYWORDS
DEVS, Multilayer, Python-DEVS, reusability, catchment basin, level of abstraction

ABSTRACT

Modeling complex natural system is a difficult task requiring the cooperation between several specialists. Owing to these specialists we can obtain separate precise models but it is often necessary to interconnect them in order to solve a given problem. However this interconnection raises two kinds of problems: (a) data may be expressed into different units according to the models; (b) time units may be different when running the different models. We solve these problems by using a special component.

1. INTRODUCTION

Modeling a complex system is a collaborative work between specialists from various expertise areas which results in the integration of a set of detailed models which have to deal with a great number of data. No problems occur until we need to connect these models to each other. But serious problems may appear during data exchange between models because data from a model may not fit to another model; temporal problems may also arise because of models working with the different time units. We have proposed a framework allowing to efficiently connect models in [1]: using the DEVS (Discrete Event Specification) formalism we have been able to define a new component called Assembly Model component. It contains conversion and temporal functions allowing to solve the two previously introduced problems. This paper will deal with a PythonDEVS [3] software implementation of the concepts of multilayered modeling and simulation already presented [1] and with the validation of the implemented software using a concrete application concerning the prediction of the behavior of a catchment basin. Such an application requires the interconnection of models developed separately and will be a nice validation example in order to describe and validate the concepts associated with the newly introduced Assembly Model component. The outline of the paper is the following. After an introduction which will set up the problem, the second section will briefly summarize the DEVS formalism, section 3 will presents the multilayer concept already presented in [1] which allow to interconnect different models using the Assembly Model component. Section 4 will detail the main classes involved in order to implement the previously summarized concepts before a brief presentation of the PythonDEVS implementation software of the multi-layered modeling and simulation. In section 5 we will present in detail how the Assembly model component involved in multilayered simulation can be used in order to deal with the prediction of the behavior of a catchment basin. We will first briefly describe two models developed by specialists required in order to perform precise simulations. The first one [2] defined by specialists belonging to the hydrological domain allows to model the behavior of a catchment basin according to rainfall. However in order to take into account the presence of snow a second model [2] defined by climatologists is also needed. We will then describe the software implementation of the two models, their interconnection using an Assembly Model component and the obtained results. Finally concluding remarks will permit to summarize the presented work as well as the future work we envision.

2. DEVS FORMALISM

DEVS formalism was created by Professor Zeigler [1] [2] [3] [4] [5] allowing to model a discrete event system. Two kinds of models are defined: 1) basic models from which larger ones are built, and 2) coupled models which describe how these models are connected together in hierarchical fashion. Basic models (called atomic models) are defined by the following structure:
CA = < X, S, Y, δint, δext, λ, ta> where:

(i) X is the set of input values;
(ii) S is the set of sequential states;
(iii) Y is the set of output values;
(iv) δint is the internal transition function dictating state transitions due to internal events ;
(v) δext is the external transition function dictating state transitions due to external input events ;
(vi) λ is the output function generating external events at the output, and
(vii) ta is the time-advance function which allows to associate a life time to a given state.

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

The behaviour of an atomic model is illustrated as follows: the external transition function describes how the system changes state in response to an input. When an input is applied to the system, it is said that an external event has occurred. The next state s' is then calculated according to the current state s. The internal transition function describes the autonomous (or internal) behaviour of the system. When the system changes state autonomously, an internal event is said to have occurred. The next state s' is therefore calculated only according to the current state s. The output function generates the outputs of the system when an internal transition occurs. The time advance function determines the amount of time that must elapse before the next internal event occurs, assuming that no input arrives in the interim.
An atomic model enables us to specify the behaviour of a basic element of a given system.
A coupled model indicates how to couple (connect) several component models together to form a new model. This latter model can itself be employed as a component of a larger coupled model, thus giving rise to hierarchical construction. A simulator is associated with the DEVS formalism in order to execute a coupled model's instructions so as to actually generate its behaviour. The architecture of a DEVS simulation system is derived from the abstract simulator concepts (Zeigler and al. 2000) associated with the hierarchical and modular DEVS formalism. The abstract simulator allows the definition of a simulation tree whose root element is dedicated to the time advance management.

3. MULTILAYERED DEVS ARCHITECTURE

One of the most difficult tasks in the field of modeling and simulation of complex systems is to choose a good level of detail. In all domains, models are built at a precise abstraction level. The abstraction level of a model determines the amount of information that is contained in the model (figure 1). As presented in figure 1 the quantity of information in a model decreases with the abstraction levels: a model described at a low abstraction level will contain more information than a model described at a higher abstraction level [16].

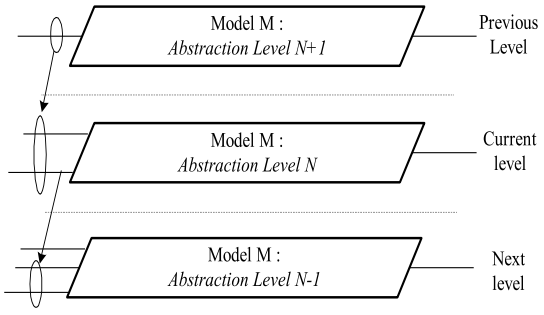


Figure 1: Abstraction Hierarchy

Well defining the abstraction level is an important step in modeling. A model described according to several abstraction levels is called a "hierarchical model". Let us call abstraction hierarchy such a notion of hierarchy.
We have to point out that this concept of abstraction hierarchy is quite different to the hierarchy of description inherent to the DEVS formalism. The notion of coupled models involving atomic or coupled models allows the definition of models using a hierarchy of description. Such a notion of hierarchy is only a mean to easily define models. In that case a flattened model can be easily generated from this hierarchy of description. However it is quite impossible to easily derive a flattered model when dealing with models involved in an abstraction hierarchy since data transfers have to be performed.
The presented work will allow to deal with models which have been designed according to different levels of abstraction. We have been able to define a multilayered DEVS architecture in order to perform the communication between models defined at different abstraction levels.

First of all we have studied HLA [15] and GDEVS [6]. These distributed architectures helped us to set up the time management of the simulation due to the similarity of the problems (especially in the next version of the multilayered architecture that will propose a new definition of the simulator including the modifying time management algorithm). However we cannot use HLA or GDEVS for our problem because one of our objectives is reusability of models. With the two-mentioned architectures we must modify and re-write the models because of the data unit problem.

The multilayered DEVS architecture [1] is based on the DEVS formalism [8]. This is an extension of the classical DEVS formalism that leans on the definition of a coupled model called Assembly Model involving two kinds of atomic models. The goal of our work is to allow an easy interconnection of DEVS models which have been defined separately and which were not dedicated to communicate with each other.

Because we have been able to solve the interconnections of models defined by specialists coming from different domains using classical DEVS atomic and coupled we did not need to modify the DEVS abstract simulator [9,10,11].

We differentiate two kinds of models:

- Behavioural models, which describe the dynamic of the systems required in order to solve a given problem i.e. the models defined by different specialists.
- Assembly model: it is the central point of the proposed formalism. Each behavioural model is linked to the assembly model and all the shared data are processed by the Assembly Model in order to performed data exchange between behavioural models.

During the data exchange three kinds of problems may occur:

- Temporal type of problems: the models involved in the study of a given phenomenon may not require data at the same time unit. For example a model may deal with seconds as time units while another one may require hours. This involve that a conversion will be necessary in order to keep the consistency of the data during the simulation

- Spatial type of problems: data issued from a geographic area associated with a model involved in the study of a given phenomenon may correspond to a larger or a smaller one in another model depending on the scales that have been used in both cases.
- And finally abstraction type of problems: Data involved in the different models required in order to study a given phenomenon may be at different levels of details. For example a model may deal with variables expressed with meters as units while another one may require variables expresses in kilometers.

In order to solve these problems, the Assembly model provides some conversions functions. In [1] we have presented a detailed version of the architecture with a formal representation.
Therefore we have slightly modified our original proposition. The most important element in the assembly model is the model called DRIV described more precisely in [1]. This component has three features:

- Time management: this feature is currently managed by the Root component; we plan to delegate this feature to the DRIV component in order to resolve the temporal problem. Because of the complexity of this feature, it will be detailed later in another paper.
- Management of the Input/Output : The DRIV component is in charge of the read and write order on the STO model. The conversion functions are called in the DRIV model instead of in the STO model as it is planed at first [1]. The DRIV component also centralizes the declarations of the data type and unit made by the authors of the different behavioural models. These declarations are required for the definition of the conversion functions.
- Management of the conflicts: The DRIV component contains a priority list for the resolution of the conflicts that may occurs during the data exchange.

The next section is devoted to the software implementation of the Assembly model while in section 4 we will detail the software implementation of a concrete example.

4. PYTHON-DEVS AND DEVSIMPY IMPLEMENTATION OF THE ASSEMBLY MODEL

In sub-section 4.1 we first briefly present the Python-DEVS software and the DEVSIMPY framework which has been used in order to implement the Assembly Model concepts. The sub-section 4.2 describes the implementation of the Assembly Model.

4.1. Python-Devs and DEVSIMPY

Python-Devs is a software implementation [3,12] of the DEVS formalism made by Jean-Sébastien Bolduc and Hans Vangheluwe using the python language. It provides two files:

- DEVS.py that contains the definition of the DEVS models i.e. the atomic and the coupled model.
- Simulator.py that implements the simulator engine.

Four classes are described:

- Port Class contains the common elements used by a DEVS port.
- BaseDEVS Class contains the elements use by both the coupled and the atomic model: the input and output port for example.
- AtomicDEVS and CoupledDEVS classes respectively describe the atomic and coupled DEVS model elements. Methods are added in these two classes. For instance the peek and poke methods of the AtomicDEVS class are used for sending and receiving a message on a port. In CoupledDEVS model, methods were also added for adding models or connecting ports. Furthermore a select method was added in order to define the priority between models if two events occurred at the same date.

A complete explanation of the Python-Devs

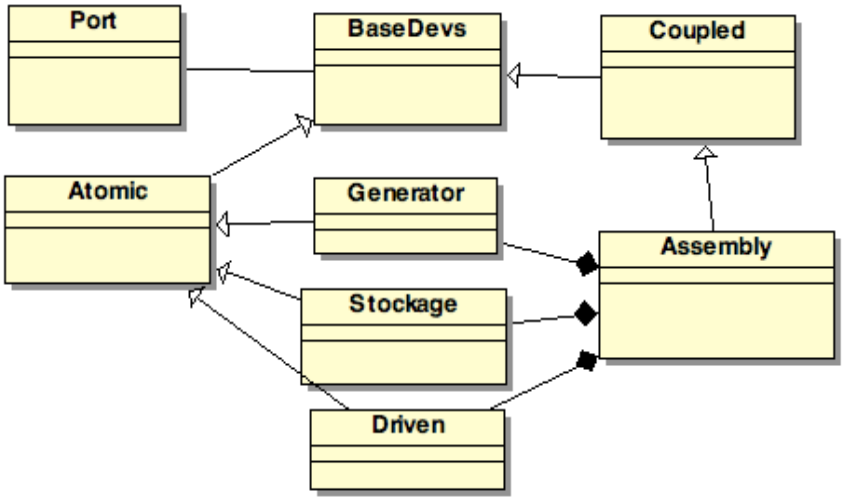


Figure 2 Class diagramm of the multilayered architecture

implementation can be found in [12] and the package can be downloaded in [3].
From Python-Devs Laurent Capocchi created a graphical environment. This environment provides a simple way for creating DEVS models : with a simple drag and drop an atomic or a coupled model can be created and clicking on it allows a window to appear for writing or modifying methods. The simulation is performed by running the Python-DEVS simulator.
A complete description of this graphical interface will be published soon.

4.2. Implementation of the ASSEMBLY MODEL using PYTHON-DEVS and DEVSIMPY

We added four classes to the original Bolduc and Vangheluwe definition of Python-Devs. Figure 2 presents the new class diagram.
We added a Generator class; its role is the creation of events. This class inherits from the Atomic Class.
We also implemented a DRIVEN class that inherits from atomic Class; this class represents the Driven model and its role is the reception and the transmission of data from/to the stockage models or the behavioural models.
A Stockage class has been also added; its role is the stockage of data. It inherits from the Atomic Class.
Finally we added a class called Assembly; this class inherits from the CoupledDEVS Class. This coupled model will allow to implement the interconnection of instances of DRIVEN, STOCKAGE and Generator classes.
Figure 2 present the class diagram of the multilayered architecture.

5. VALIDATION THROUGH A CONCRETE EXAMPLE

We choose a concrete example for the validation: the prediction of the hydrologic behavior of a catchment basin.
We will present in sub-section 5.1 the models involved in order to describe the hydrologic behavior of a catchment basin. Sub-section 5.2.

5.1. Hydrologic behavior of a catchment basin.

The hydrologic behavior of a catchment basin can be obtained by interconnecting two models: a snow model and a watershed model.
Among all of the existing watershed modeling we have select the GR3J model which has the advantages to be quite simple but also quite precise. GR3J is an hydrological model for the study of catchment. It performs good results by using a representation of the rainfall-runoff process as simple as possible and depending on very few parameters.

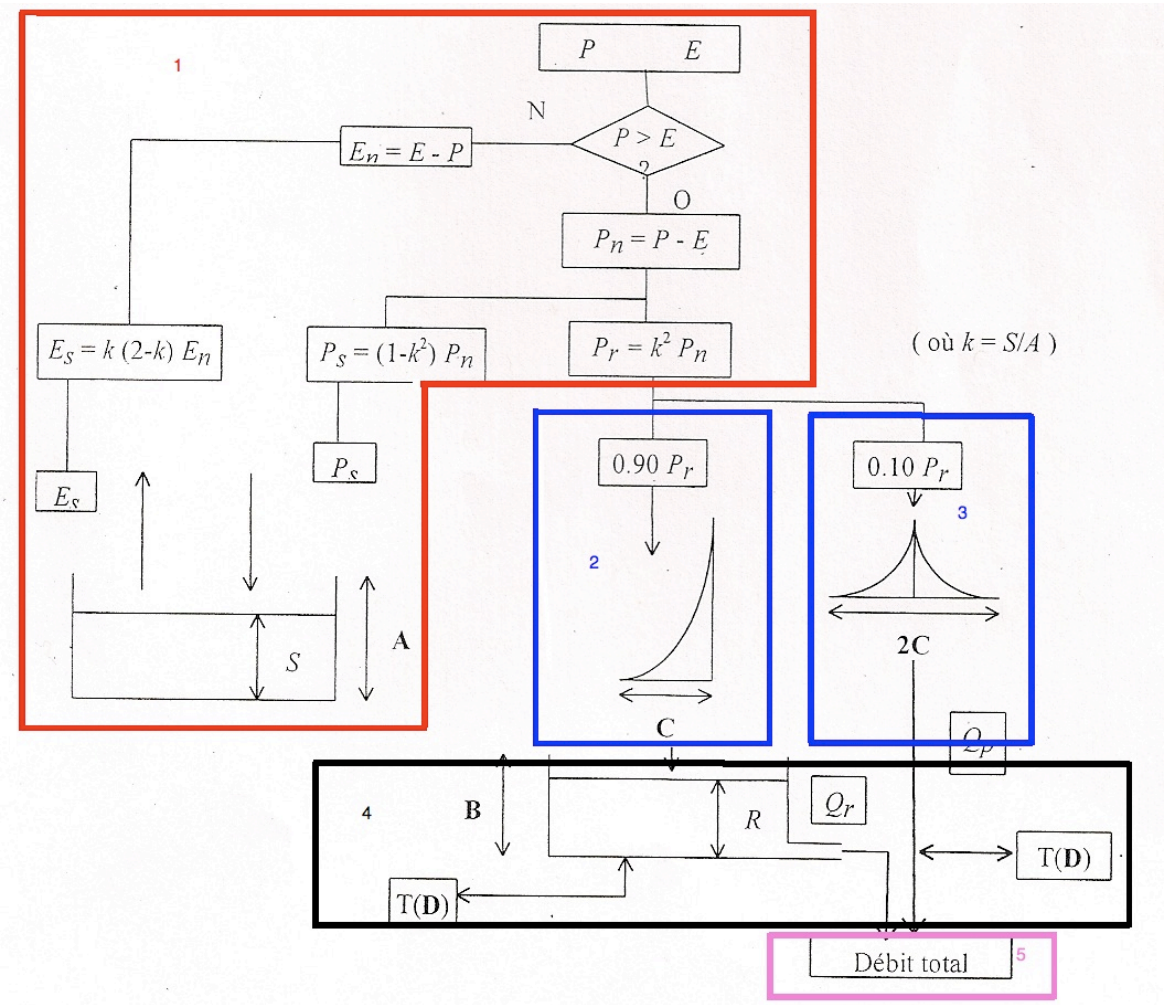


Figure 3 GR3J model

Figure 3 describes the GR3J model. We may point out that variables P and E represent respectively the precipitations and the potential evaporation. If the precipitations are greater than the evaporation the first transformation (Pn) is performed. Then the following two functions are computed:

- The production function: a part of the water goes through the soil reservoir which is defined by its capacity (noted down here A) and its real level S. S evolves according to the rain Pn and the evaporation En. The input (Ps) and output (Es) flows are taken into account when Pn and En are positive.
- The transfer function: the water that does not go into the soil reservoir represents the available water for runoff (called Pr in figure 3). This quantity of water is divided into two parts: the most important quantity is described in the left part in figure. The reader may notice in figure 3 that 90% is transformed by a 1-day unit hydrograph (let us call it SH1) while 10% is transformed by a second unit-hydrograph (let us called SH2). The first part, after routing by SH1 is given as input to a reservoir whose level is R and maximum capacity is B. This reservoir allows to obtain a first output called Qr in figure 3. The second hydrograph SH2 allows to generate a second output called Qp in figure 3.

  After passing through these two hydrographs the total of the two outputs (Qr and Qp) of the two hydrographs represents the total steam flow (called Debit Total in figure 3).

We have coded this scheme into DEVS models. We have been able to represent the behavior pointed out in figure 4 by the interconnection of the following five DEVS atomic models:

- The first atomic model (called ModelInterception) allows to model the part of the algorithm of figure 3 which corresponds to the red part and the number 1. The model receives precipitations as input, checks if the rain is greater than the evaporation and then sends the appropriate part to the soil reservoir and the other to the next model.
- ModelSH1 (in blue and noted 2 in figure 3) represents the first hydrograph; its take 90% of the water and send it to the reservoir model depending on a given formula.
- ModelSH2 (in blue and noted 3 in figure 3) represents the second hydrograph; the water quantity that is received as input is almost immediately sent on the output port.
- ModelRoutage (in black and noted 4 in figure 3) is the reservoir in which water goes after passing through SH1.
- ModelSomme (in pink and noted 5 in figure 3) allows to compute the sum of the output of the two hydrographs and therefore send on the output port the final output.

The interconnection of these five models allows to define a coupled model representing the GR3J model.

The snow model receives the temperatures and the precipitations as input to compute a result using a formula defined by a climatologist.

A complete description of these models can be found in [2] but we have to point out that the GR3J model presented in this paper was slightly modified in order to obtain better results.

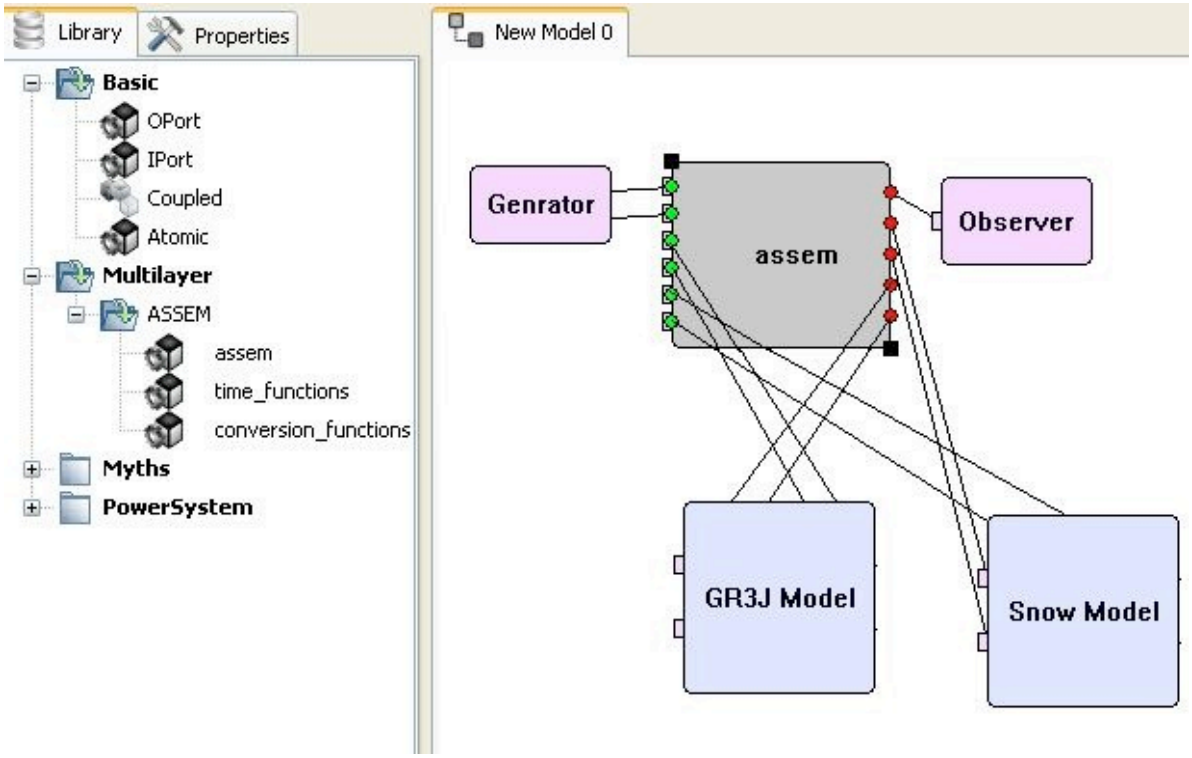


Figure 4 Multilayered architecture in DEVSIMPY

The next section describes how we have been able to interconnect these two models using a Python-DEVS and DEVSIMPY Assembly model implementation.

5.2 Python-DEVS and DEVSIMPY implementation

We describe in this sub-section how we have been able to interconnect the previously introduced two models using the assembly model. We present in figure 3 the coupled model allowing to interconnect the Assembly model (called ASSEM coupled model in figure 3) with the two models (GR3J and Snow atomic models).

In Figure 4 the left part of the window presents the domains available in DEVSIMPY. We add the multilayer domain, which contains the description of the assembly model, the generator, the observer, the GR3J Model and the Snow model. With a simple drag and drop we can add one of these models to the new Model, which is going to be created in order to perform simulations of the hydrologic behaviour of a catchment basin.

The hydrological and snow model are here considered as black box, so we are not able to change anything on it since specialists have separately defined them. We only know which kind of data (as well as its unit) is going to be sending on a given port. Owing to this information we have been able to create a kind of library which will help us in order to perform an automatic treatment of the data: let us call RoutageLibrary this library.

The conversion functions allowing the two models (GRJ3 model and snow model to communicate) are stored into a file and may be called by the stockage model when a data is sent on or sent to a given port. The stockage model receives a read or write order. Depending on this order it can execute two actions:

- Write order: the stockage model receives a value on one of its ports with a write order. Using the RoutageLibrary the unit of the considered data is obtained and the corresponding conversion function is invocated before storing the obtained data.
- Read order: the stockage model receives a read order on one of its ports. Using the RoutageLibrary we are able to find which kind of unit associated to the considered data and the corresponding conversion function is invocated.

5.3. Results

This part presents the results of this work. We have pointed out a comparison between real measurements taken by a specialist on a river chosen as example and the output of the simulation process. Figure 5 presents these results for a year and figure 6 between the year 1969 and 1972.

We can see that the obtained results fit well with the real measurements except for few parts of the two graphs. These differences are due to the GR3J model; this

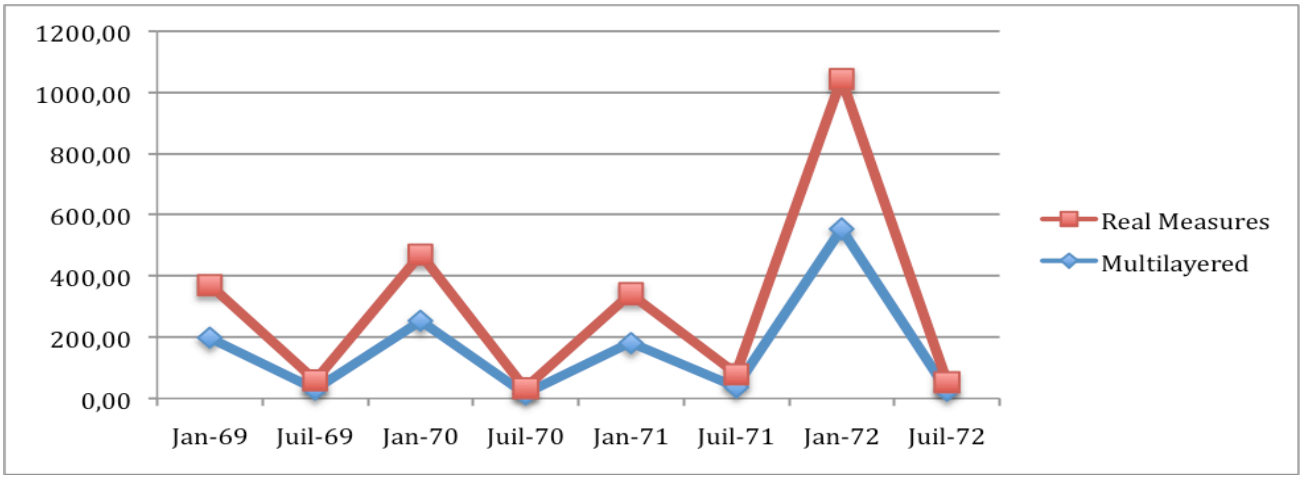


Figure 5 Comparision between real measures and results of our simulation for a year

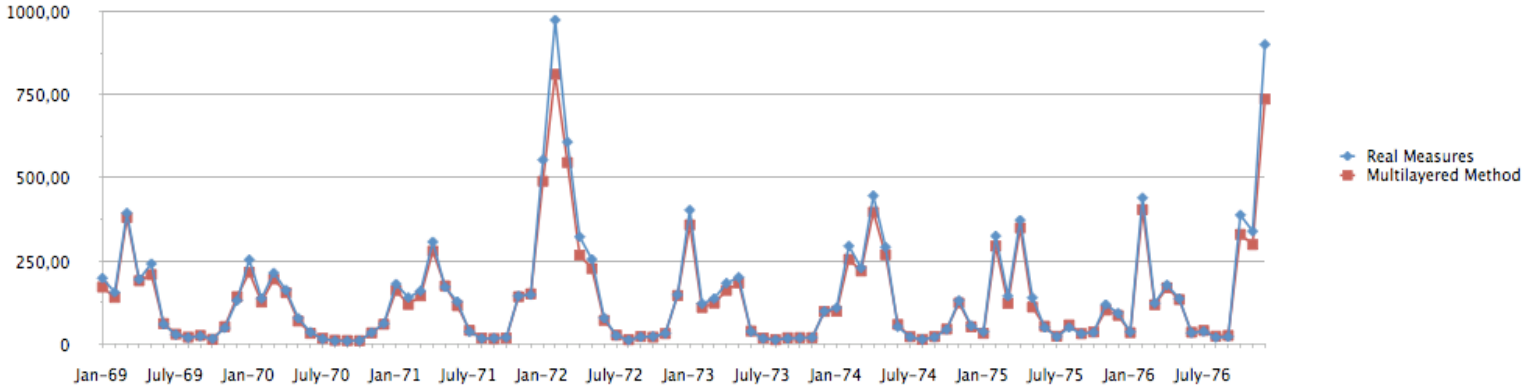


Figure 6 : Comparison between real measures and results of our simulation for several years

model is not suitable for the treatment of flood events. The highest peaks in the first graph (figure 5) correspond to flood periods.
We can see on the second figure (figure 6) that the differences correspond to the winter and spring period. We may point out that these periods correspond to a combination of rain periods with snowmelt periods which produces huge flood.
We finally can conclude that the obtained results are pretty good in comparison to the real measurement except when flood periods occur.

6. CONCLUSIONS AND PERSPECTIVES

We have presented here a validation of the multilayered formalism presented in [1]. We use a validation example dealing with the prediction of the behaviour of a catchment basin. This example perfectly fits to our problematic because of its complexity. By complexity we mean the use of different units of data used in order to describe the different models. The next validation example we envision will be the progression of a fire forest, which is a little more complex than the behaviour of a catchment basin.
We use Python-Devs and its extension DEVSIMPY for the implementation. We choose the python language for its simplicity but obviously not for its rapidity which is not here the objective.
The last part of this paper presented results obtained by running the combination of two separately defined models using the Assembly model. The obtained results show the efficiency of our approach. The next step will be the integration of a solution for dealing with temporal problems. The time management will be delegated to the Assembly model, so we will need to redefine the abstract simulator. In a first approach we plan to use a conservative time management method.

REFERENCES

[1] Broutin E, Bisgambiglia P, Santucci JF, “Simulation of heterogeneous DEVS models; application to the study of natural systems”. In proceeding of the Spring Simulation Conference 2009, San Diego CA.

[2] Broutin E, Bisgambiglia P, Santucci JF, « A PYTHON VALIDATION OF THE MULTILAYER DEVS THEORY: CASE OF A CATCHMENT BASIN, European Simulation and Modelling Conference, OCT 26-28, 2009 Holiday Inn, Leicester, UK, EUROPEAN SIMULATION AND MODELLING CONFERENCE 2009, Pages: 15-19

[3] http://moncs.cs.mcgill.ca/MSDL/reasearch/DEVS

[4] Liu B, Yao Y, Toma J, Wang H, “Implementation of Time Management In Runtime Infrastructure”.

[5] Praehofer H; Sametinge J, Stritzinger A, 2000. "Building Reusable Simulation Components," *presented at WebSIM2000, Web-Based Modelling and Simulation, San Diego, CA, USA.*

[6] Zacharewicz G, Frydman C, Giambiasi N, “G-DEVS/HLA Environment for Distributed Simulations of Workflows”, in: Simulation, n° 84, pp. 197-213, 2008

[7] Zeigler, B.P; Hall S.B; Sarjoughian H.S. 1999. "Exploiting HLA and DEVS To Promote Interoperability and Reuse in Lockheed's Corporate Environment," *Simulation*, vol. 73, number 4.

[8] Zeigler, B.P. 1975. “Theory of Modelling and Simulation” *Academic Press*.

[9] Zeigler, B.P. 1976. “Theory of Modeling and Simulation.” *New York, Wiley.*

[10] Zeigler, ,B.P. 1984. “Multifaceted Modelling and Discrete Event Simulation”. *London, Academic Press.*

[11] Zeigler, B.P. 1990. “Object-Oriented Simulation with Hierarchical, Modular Models”.

[12] Bolduc J.S. and Vangheluwe H, PythonDEVS: a modeling and simulation package for classical hierarchical DEVS. Technical Report, MSDL, McGill University, 2001

[13] Calvin, J.O, Weatherly R, “An Introduction To The High Level Architecture Runtime Infrastructure (RTI)”.

[14] Fujimoto R.M. “Time Management In The High Level Architecture”.

[15] Frederick Kuhl , Richard Weatherly , Judith Dahmann, Creating computer simulation systems: an introduction to the high level architecture, Prentice Hall PTR, Upper Saddle River, NJ, 1999_

[16] P. Benjamin,. Erraguntla JM, Delen D, Mayer R, "Simulation Modeling at Multiple Levels of Abstraction," presented at the 1998 Winter Simulation Conference, 1998

BIOGRAPHIES

Emilie Broutin is a PhD student in computer science in the University of Corsica. Her current research focuses on DEVS multilayered modeling and simulation.

Paul Bisgambiglia is a professor in Computer Sciences at the University of Corsica. He is responsible of the modeling and simulation team of the UMR CNRS 6134. His research activities concern the techniques of modeling and simulation of complex systems and the test of systems described at high level of abstraction. He makes his researches in the laboratory of the UMR CNRS 6134.

Jean-Francois Santucci is Full Professor in Computer Sciences at the University of Corsica since 1996. His main research interest is Modelling and Simulation of complex systems. He has been author or co-author of more than 100 papers published in international journals or conference proceedings. He has been the scientific manager of several research projects corresponding to European or industrial contracts. Furthermore he has been the advisor or co-advisor of more than 20 PhD students and since 1998 he has been involved in the organization of more than 10 international conferences. He is conducting newly interdisciplinary researches involving computer sciences, archaeology and anthropology: in the one hand he is performing researches in the archeaoastronomy field (investigating various aspects of cultural astronomy throughout Corsica and Algeria using tools issued from Computer Sciences) and on the other hand he is applying computer sciences approaches such as GIS (Geographic Information Systems) or DEVS (Discrete EVent System specification) to anthropology.

A NEUROALGORITHMIC INVESTIGATION OF THE OUTER RETINA

Tomás Maul, Andrzej Bargiela and Lee Jung Ren
School of Computer Science
University of Nottingham Malaysia Campus
Malaysia
Email: Tomas.Maul@nottingham.edu.my

KEYWORDS

Retinal Models; Multi-Resolution Recurrent Neural Networks; Optimization; Image Processing.

ABSTRACT

In this paper a new model of the Outer Plexiform Layer (OPL) of the human retina is presented. The model, which is a multi-resolution Linear Recurrent Neural Network (LRNN) defined by 31 parameters, was subjected to several optimization experiments targeting different low-level visual functions involving the control of noise, brightness, contrast, saturation and color. Our results indicate that the model can indeed implement the above image processing functions and that the solutions can be modulated in different ways. The model provides a good starting point for extensions targeting real world applications and for the generation of testable biological hypotheses.

INTRODUCTION

The retina is arguably one of the most extensively studied neural structures in the primate nervous system. In spite of this, many open problems regarding both the structure and function of the retina remain to be addressed. For example, the precise number of different horizontal cell (HC) types in the primate retina is still a matter of debate (Kolb et al., 1994). Furthermore, the precise connectivity patterns between different photoreceptor and HC types is still unclear. Added to this, the retina's representational, coding and computational functions have yet to be completely revealed (Field and Chichilnisky, 2007). In a similar vein, but now in the context of computer applications, the field of Image Processing is one where the maturity of the field does not entail any shortage of open problems. It is hard to identify an image processing function (e.g. denoising or colour correction) that can't benefit from further improvement, be it in terms of accuracy, speed or adaptability. Refer to (Egmont-Petersen et al., 2002) for a useful review of Artificial Neural Network (ANN) solutions to Image Processing problems.

The primate retina consists of five main cell types, i.e.: photoreceptor (PR), horizontal, bipolar, amacrine and retinal ganglion cells. Each one of these types consists of several sub-types, such that the retina can be seen to consist of 55 distinct neurons subtypes in total (Masland, 2001). Two main layers of neuronal interconnections, the Outer Plexiform Layer (OPL) and the Inner Plexiform Layer (IPL) carry out the bulk of retinal processing. The current paper focuses on the OPL seeing that, in spite of its greater simplicity and thus modeling viability, it still presents us with several open problems. Furthermore, accurate modeling of the OPL output (and thus IPL input), can be seen as a pre-requisite for modeling the latter.

Efforts related to the modeling or emulation of biological retinas is not new and can be classified into the following three categories, ordered from the most biologically concerned to the most engineering based: 1) computational neuroscience modeling (Wohrer et al., 2006), 2) software based image processing (Ebner, 2006) and 3) hardware based image processing (Zaghloul and Boahen, 2006). Although these efforts have contributed to addressing some biological questions and advancing image processing to some extent, as we have mentioned challenges remain. To the best of our knowledge, our model of the OPL is the first one to include detailed chromaticity-specific connection patterns between PRs and HCs, with the possibility of distinct patterns for different HC types.

Our approach lies at the intersection of Computational Neuroscience and Natural Computation. For the sake of simplicity, this intersection will be hereafter referred to as Neuroalgorithmics. This relatively old but newly named field is concerned both with the modeling of biological neural systems and the extraction of biologically inspired algorithms. It is hoped that the explicit pursuit of this dual concern will provide us with a new framework and methodologies for solving some of the remaining open problems in vision science.

Because retinal models are generally considered large scale models it was not practical to use multi-compartmental modeling techniques. Furthermore, because the OPL is generally interpreted as a linear filter and contains feedforward, lateral and recurrent connections, we decided to simplify its representation via a Linear Recurrent Neural Network (LRNN) (Haykin, 2008). The LRNN model is represented by a set of 31 parameters. As mentioned in (Schutter, 2009) some modeling studies are motivated by unknown input/output relationships, while others are motivated by unknown structural elements (e.g. connectivity). In the OPL case, there

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

is partial and incomplete understanding of both aspects. The parameterization allows for flexible generation of different OPL network configurations which potentially can help us reveal structural aspects. The optimization of the network parameters viz a viz different low-level visual functions, helps address questions related to input/output relationships.

The primary objective of the work reported in this paper is to extract a useful neural solution from nature, that is accurate, compact, efficient, parallelizable, flexible (i.e. implements different visual functions), adaptable (i.e. allows for simple modulation), extensible and theoretically tractable. The secondary objective of the work is to generate reasonable hypotheses or useful questions pertaining to the biological retina.

The following section will discuss the methods adopted in more detail (e.g. the model and optimization framework). Experimental results will be presented in the subsequent section demonstrating several properties of the LRNN. The final section will analyze the results, evaluate the LRNN with reference to the work's objectives, and offer several conclusions and directions for future work.

METHODS

The Model

The model reported in this paper is essentially an OPL-inspired LRNN. The main OPL elements incorporated in the model include: different types of cone photoreceptors (PR), inter-cone gap junctions (GJ), different types of horizontal cell, inter-HC gap junctions, feedforward connections from PRs to HCs, feedback connections from HCs to PRs, local HC receptive fields and variable proportions of connections between different PR types and HCs. Omitted OPL elements include: rods, nonlinearities of cell potentials and complex temporal dynamics. Other simplifications include the representation of light as image pixels and the assignment of three cones per image pixel. The abstraction choices that were made reflect our intention to generate plausible hypotheses pertaining to the functional consequences of GJs and cone/HC connectivity patterns and to efficiently test the functional diversity of the network.

The structure and function of the LRNN is represented by a set of 31 parameters. Table 1 depicts these parameters and provides brief descriptions for each one. The value of each parameter belongs to the range $[0, 1]$. For attributes whose values belong to the set of natural numbers (e.g. maximum iterations), another parameter is used to represent the actual maximum value of that attribute. In this case, the value in Table 1 (e.g. parameter 31) represents a proportion and the actual value of the attribute is computed via the function $r(p_i \times max_i)$, where r is the rounding function, p_i represents the value of parameter i and max_i represents the actual maximum value of attribute i. Maximum values are defined for the following three attributes: maximum number of iterations, HC input radius and HC output radius. The notation HC_n denotes the horizontal cell with index n. The multi-resolution property of the LRNN is a result of the fact different HC types may receive/send information from/to the PR layer at different resolutions.

Table 1: LRNN parameter descriptions.

P1	Weight of the influence of light on cones
P2	Weight of a cone's influence on itself
P3	Cone GJ weight: intra-positional and inter-chromatic
P4	Cone GJ weight: inter-positional and intra-chromatic
P5	Cone GJ weight: inter-positional and inter-chromatic
P6	Radius of input from the PR layer onto HC1 cells
P7	Radius of output from HC1 cells to the PR layer
P8	Proportion of synapses from red cones onto HC1 cells
P9	Proportion of synapses from green cones onto HC1 cells
P10	Proportion of synapses from blue cones onto HC1 cells
P11	Relative connection weights from PRs onto HC1 cells
P12	Relative connection weights from HC1 cells onto PRs
P13	Radius of input from the PR layer onto HC2 cells
P14	Radius of output from HC2 cells to the PR layer
P15	Proportion of synapses from red cones onto HC2 cells
P16	Proportion of synapses from green cones onto HC2 cells
P17	Proportion of synapses from blue cones onto HC2 cells
P18	Relative connection weights from PRs onto HC2 cells
P19	Relative connection weights from HC2 cells onto PRs
P20	Radius of input from the PR layer onto HC3 cells
P21	Radius of output from HC3 cells to the PR layer
P22	Proportion of synapses from red cones onto HC3 cells
P23	Proportion of synapses from green cones onto HC3 cells
P24	Proportion of synapses from blue cones onto HC3 cells
P25	Relative connection weights from PRs onto HC3 cells
P26	Relative connection weights from HC3 cells onto PRs
P27	Weight of an HC's influence on itself
P28	HC GJ weight: same position, different HC type
P29	HC GJ weight: different position, same HC type
P30	HC GJ weight: different position, different HC type
P31	Number of iterations

Equation 1 encapsulates the update rule for the PR layer. As the equation shows, the main inputs to any PR cell originate from light, the cell itself and nearby PR and HC cells:

$$b_{ip}^{t+1} = w_1 a_{ip}^t + w_2 b_{ip}^t + w_3 \sum_{j \neq i} b_{jp}^t + w_4 \sum_{p' \in n(p)} b_{ip'}^t + \\ + w_5 \sum_{\substack{j \neq i \\ p' \in n(p)}} b_{jp'}^t \quad \sum_{h \in H} w_h \left(\sum_{(hp') \in v(ip)} c_{hp'}^t \right) \tag{1}$$

where a, b and c represent light intensity, PR cell activity, and HC cell activity respectively, t denotes time, i and

j index chromaticity, p and p' index position and h represents HC type. The set of neighbours $n(p)$ is defined in Equation 2 while the set of connections $v(ip)$ is defined in Equation 3. Influence weights are denoted by w_i where the subscript refers to the corresponding parameter i in Table 1. The weight w_h represents the influence of HC cells of type h on PR cells. Although weights are based on their corresponding parameters, actual weights are normalized by the total sum of weights used in a particular cell update.

$$n(p) = \{p' \mid p' \neq p \wedge d(p'_x, p_x) \leq r \wedge d(p'_y, p_y) \leq r\} \quad (2)$$

where $d(p'_x, p_x)$ represents the distance between positions p and p' along the x-axis and r represents the *radius* of GJ influence.

$$v(ip) = \{(h, p') \mid con(PR_{pi}, HC_{hp'}) \equiv 1\} \quad (3)$$

where $con(c_1, c_2) \equiv 1$ denotes that a feedforward connection exists from cell c_1 to cell c_2.

Equation 4 encapsulates the update rule for the HC layer:

$$c^{t+1}_{hp} = w_{h'} \left(\sum_{(ip') \in v(hp)} b^t_{ip'} \right) + w_{27} c^t_{hp} + w_{28} \sum_{k \neq h} c^t_{kp} + \\ + w_{29} \sum_{p' \in n(p)} c^t_{hp'} + w_{30} \sum_{\substack{k \neq h \\ p' \in n(p)}} c^t_{kp'} \quad (4)$$

where b, c, t, i, p, p', w_i and $n(p)$ are defined as before, k represents HC type, $v(hp)$ is defined in Equation 5 and $w_{h'}$ represents the weighted influence of PR cells on HC cells of type h.

$$v(hp) = \{(i, p') \mid con(PR_{ip'}, HC_{hp}) \equiv 1\} \quad (5)$$

The connectivity between PRs and different HC types is characterized by distinct weights, input/output radii and proportions of red, green or blue cones (see Table 1: P6 to P26). The model makes use of connectivity templates (one for each HC type), which are repeatedly used over the whole space of cells.

Initialization of the network is performed through a simple sequential activation process, whereby first the inputted image (i.e. "light") is copied to the PR layer and then the activities of the cells in the HC layer are computed by taking into account PR activities and PR-to-HC connections only (i.e. HC GJs are ignored at this stage).

Optimization

The simplified model of the OPL as characterized by the 31 LRNN parameters, was subjected to several optimization experiments. In each experiment the model (or parameter set) was optimized relative to a different low-level visual function, i.e.: ↓ noise, ↑↓ brightness, ↑↓ contrast, ↑↓ saturation and ↑↓ redness, where ↑ denotes *increase* and ↓ denotes *decrease*. In general, if a visual function involves increasing measure m (e.g. contrast) then optimization with regards to that function is implemented via a cost function that takes as input a source and a target image, where the latter corresponds to the former with increased m, and computes the distance between the output of the LRNN (as defined by the parameter set) and the target image. All images used in the optimization were of resolution 100×100. After optimization, the LRNN targeting a new visual function, is almost invariably a new set of parameters and thus a new neural network structure/function.

Optimization was performed using a Global Stochastic Optimization (GSO) algorithm, combining elements of Differential Evolution Storn and Price (1997) and Genetic Algorithms Goldberg (1989). At each iteration, the population of solutions is first expanded through mutation and cross-over, then ranked, after which new solutions are generated based on the differences between strong and weak solutions. Solutions are then ranked and trimmed keeping both quality and diversity in mind. Optimization continues until either a certain maximum number of iterations or a particular cost has been reached. For lack of space we will not delve into further details of the GSO. Note that although our optimization approach is a hybrid, we have no reason to believe that other GSO approaches should be less successful in finding useful parameter configurations.

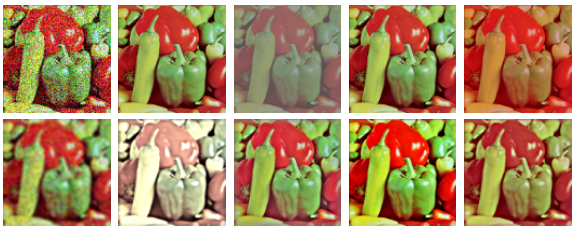


Figure 1: Processing examples.

RESULTS

Figure 1 contains examples of the processing capabilities of the OPL model. The top row depicts input images whereas the bottom row depicts corresponding output (processed) images (extracted from the PR layer). From left to right images were processed with network configurations that were optimized for the following functions: ↓ noise, ↑ brightness, ↑ contrast, ↑ saturation and ↓ redness.

Although the optimization cost functions were designed around single images, Table 2 demonstrates that the resulting network configurations are general enough to suitably handle new images. For each visual function (e.g. ↓ noise), the corresponding solution was tested on

Table 2: Some generalization results.

Function	% Correct	Mean	Max.
↓ Noise	100	3.04	4.18
↑ Brightness	60	0.02	0.04
↓ Brightness	70	0.05	0.15
↑ Contrast	100	0.75	0.94
↓ Contrast	100	0.12	0.35
↑ Saturation	100	0.12	0.16
↓ Saturation	100	0.1	0.13
↑ Redness	70	0.0003	0.0011
↓ Redness	60	0.0003	0.0010

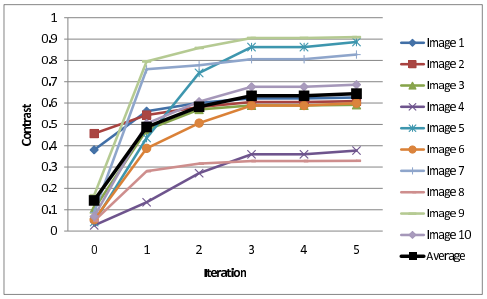


Figure 2: Multiple iterations of contrast enhancement.

ten different images. The second column from the left depicts the percentage of images that were correctly processed for each function. The definition of "correct processing" is relative to a non-reference measure for the visual function under consideration. For example, in the ↓ noise case, if the *noise* of the output image is lower than that of the input, then the processing is considered to be correct. The third column from the left depicts the mean measure differences ($output \quad input$) for each corresponding function, whereas the last column depicts the most significant difference.

Table 3 depicts parameter configurations optimized for five different functions. The functions from left (↓N) to right (↓S) are: decrease noise, increase contrast, decrease contrast, increase saturation and decrease saturation.

The fact that the LRNN is stable can be observed in Figures 2 and 3. These graphs also demonstrate how different quantities (e.g. contrast) can be fine-tuned by applying multiple iterations. Iteration 0 depicts the contrast (or saturation) measures of the input images. Notice how in the saturation case, the processing measure at the last iteration is more predictable from iteration 0 than in the contrast case.

Apart from the application of multiple iterations (e.g. Figure 2), fine-graded control over visual functions can also be obtained from parameter modulation. Figures 4 and 5 show several relevant parameters, whose modulation leads to predictable changes in terms of contrast. The first figure depicts parameters that are directly pro-

Table 3: Parameters optimized for different functions.

Param.	↓N	↑C	↓C	↑S	↓S
P1	0.12	0.78	0.00	0.87	0.95
P2	0.84	0.53	0.00	0.32	1.00
P3	0.11	0.00	0.00	0.00	0.34
P4	0.44	0.03	0.00	0.05	0.00
P5	0.00	0.00	0.57	0.00	0.00
P6	0.17	0.57	0.00	1.00	0.27
P7	0.66	0.94	0.28	0.62	0.00
P8	0.28	0.68	0.62	0.63	0.92
P9	0.00	0.88	0.39	0.82	0.00
P10	0.90	0.96	1.00	0.75	0.56
P11	0.33	0.50	0.84	0.21	0.00
P12	0.78	1.00	1.00	0.60	0.00
P13	0.42	0.00	0.13	0.41	0.00
P14	0.00	0.97	0.46	0.00	1.00
P15	0.16	0.35	0.45	0.36	0.89
P16	0.92	0.97	0.81	0.05	0.14
P17	0.00	0.00	0.88	0.18	0.28
P18	0.05	1.00	0.63	0.00	0.77
P19	0.62	0.05	0.80	0.27	0.30
P20	1.00	0.69	0.01	0.50	0.48
P21	1.00	0.17	0.39	0.17	0.00
P22	0.35	1.00	0.00	0.70	0.51
P23	1.00	0.68	0.83	0.45	0.00
P24	0.06	0.45	0.82	0.59	0.00
P25	0.15	0.46	0.67	0.49	0.04
P26	0.74	1.00	0.91	0.68	0.00
P27	0.94	0.55	1.00	1.00	0.94
P28	1.00	0.00	1.00	0.34	0.00
P29	0.97	0.37	0.08	0.00	0.27
P30	0.10	0.69	0.00	1.00	0.53
P31	0.95	0.08	0.50	0.46	0.07

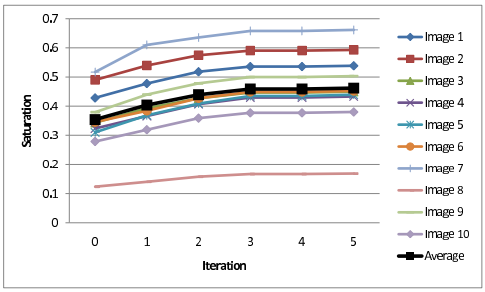


Figure 3: Multiple iterations of saturation enhancement.

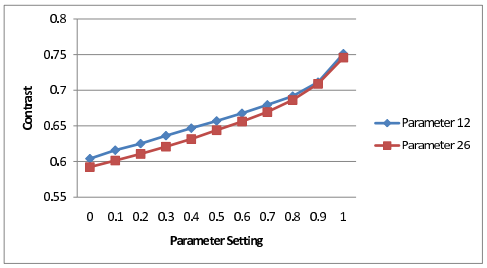


Figure 4: Contrast parameters with direct proportionality.

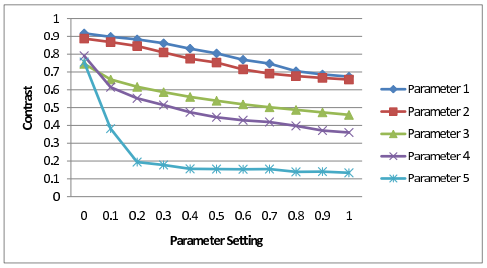


Figure 5: Contrast parameters with inverse proportionality.

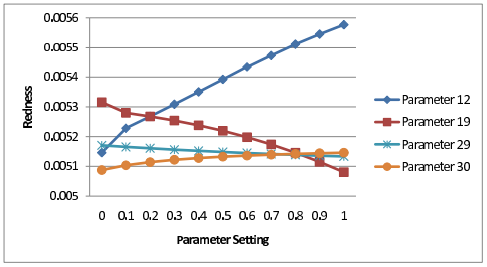


Figure 6: Redness parameters.

portional to contrast, whereas the second one depicts parameters that are inversely proportional to contrast. Notice how different parameter/contrast curves exhibit different degrees of linearity and steepness.

Figure 6 presents several parameters whose modulation affects the level of redness in processed images. The different curves reveal different types of relationship (i.e. direct or inverse proportionality), degrees of linearity and steepness.

DISCUSSION

The results indicate that the OPL-inspired LRNN is stable (see Figures 2 and 3) and does indeed exhibit functional versatility (see Table 2). The model is capable of modulating noise, brightness, contrast, saturation and colour. Table 2 demonstrates that optimized solutions can generalize to new images. According to these results, the strongest functions are ↓ noise, ↑↓ contrast and ↑↓ saturation, whereas the weakest ones are ↑↓ brightness and colour correction. It is possible that a slightly more complex model (e.g. incorporating further biological nonlinearities) might allow for more effective brightness normalization and color correction.

Due to space constraints, an in-depth analysis of the parameter variations in Table 3 is not possible. However some observations can be made. Notice how parameters P3 and P4 tend to be larger in the ↓noise solution relative to the other solutions. This is due to the averaging (and thus denoising) effect of the corresponding GJs. P3 is also relatively large in ↓saturation because this parameter directly contributes to the pulling of chromaticities towards a central grayer value. In the ↓saturation solution, notice how the HC layer seems to be effectively switched off (e.g. the HC3-PR weight (P26) is zero and the HC1 output radius (P7) is zero). Further tests of the degree of impact of different HC types on the PR layer confirmed that in the ↓saturation solution only HC2 had an impact. In the ↓noise solution the desaturating effect of P3 is counteracted by a saturating effect implemented by the HC layer (all HC types have an impact on the PR layer). This saturating effect mirrors what is found in the ↑saturation solution. First of all, notice how in the ↑saturation solution P3 is zero (in contrast to ↓saturation). Secondly, notice how HC2 (i.e. the second HC type, counting from the top of the table) is effectively unused: the input weight is zero and the output radius is zero. The two remaining HC types are differentiated in terms of receptive field size (one large and one small) and chromatic properties (the order of RGB proportions in one HC type is the reverse of the order in the other). Thirdly, notice how inter-positional HC GJs are *zero* for the same HC type but *one* for different types. The ↑contrast solution is quite similar to the ↑saturation solution. HC impact tests revealed that all HC types had an effect on the PR layer, although HC2 did have a significantly weaker impact than the other HC types. HC1 and HC3 exhibit different receptive field sizes and inverse orders of RGB proportions. The HC input/output weights are approximately $0.5/1.0$ for both HC types. The HC GJ configuration in the ↑contrast solution is relatively distinct from the ↑saturation solution, however since both solutions do their processing in a single iteration (see P31), thus making the effects of HC GJs not apparent at the PR layer until iteration 2, nothing can be concluded from their different settings. The simple ↓contrast solution appears to be mostly a consequence of parameters 1 to 4 being zero.

The results also show that the visual functions can be modulated either by running variable numbers of iterations (e.g. Figure 2) or by modulating specific parameters. Figure 4 shows that by increasing/decreasing the weight of connections from HC to PR cells, we can increase/decrease contrast. Ideally we would like to have a minimal subset of parameters, which is amenable to logical manipulation, such that any of the functions (e.g. color correction) can be activated to any degree (e.g. large gradations) and in any way (e.g. remove blue illuminant). The current paper demonstrates initial steps in this direction.

The fact that saturation related solutions seem to re-

quire fewer HC types than contrast related solutions might contribute to answering the question of why the human retina seems to exhibit 3 HC types contrary to many other mammalian retina. The experiments also revealed that even in cases where HCs might have no direct impact on the PR layer they may still be crucial in an indirect manner, e.g.: if a HC does not receive input from the PR layer it may do so (indirectly) from neighboring HCs and then output this information to the PR layer or if a HC does not output information to the PR layer, it may receive information from it and then feed it back (indirectly) via neighboring HCs.

CONCLUSION

Some of the strengths of the OPL-inspired multi-resolution LRNN include: generality, parallelizability, simplicity, versatility, adaptability, extensibility and biological relevance. Generality is verified by the fact that the LRNN performs as expected on new images. Parallelizability is an automatic consequence of the LRNN being a neural solution. Because the LRNN consists of a mere two layers we can say that the solution is simple. Versatility results from the fact that the LRNN can be optimized for several different functions involving noise, brightness, contrast, saturation, color and possibly other properties. Adaptability ensues from the fact that the LRNN can be effectively modulated in different ways. The fact that the LRNN can be easily modified to incorporate new properties (e.g. salience dependence) suggests that it is extensible.

As the results demonstrate, currently the weakest functions seem to involve color correction and brightness control. These weaknesses might be linked to another limitation, which refers to the chosen level of abstraction. By incorporating other OPL details and thus expanding the currently restricted parameter space, the above limitations might be solved and new functions might be realizable (e.g. saliency mapping).

Future work thus includes: expansion of the model, improvement of brightness and color correction, exploration of new low-level visual functions, deeper analyses of parameter settings and their effects on different functions, inclusion of more realistic temporal properties and development of a theory of the network's dynamics, analysis of patterns of indirect HC effects, exploration of algorithmic extensions, generation of biological hypotheses with broader implications and refinement of the function control space. The last point refers to the ability to efficiently modulate parameters such that different functions and function combinations can be activated in different ways and degrees using a minimal parameter subset. The ultimate objective is to automate the modulation of this parameter subset and relate this to plausible OPL functions such as contrast gain control.

REFERENCES

Ebner, M. (2006). Evolving color constancy. *Pattern Recognition Letters*, 27(11):1220–1229.

Egmont-Petersen, M., De Ridder, D., and Handels, H. (2002). Image processing with neural networksa review. *Pattern Recognition*, 35(10):2279–2301.

Field, G. and Chichilnisky, E. (2007). Information processing in the primate retina: circuitry and coding.

Goldberg, D. (1989). *Genetic Algorithms in Search and Optimization*. Addison-wesley.

Haykin, S. (2008). *Neural networks: a comprehensive foundation*. Prentice Hall.

Kolb, H., Fernandez, E., Schouten, J., Ahnelt, P., Linberg, K., and Fisher, S. (1994). Are there three types of horizontal cell in the human retina? *Journal of Comparative Neurology*, 343(3):370–386.

Masland, R. (2001). The fundamental plan of the retina. *nature neuroscience*, 4:877–886.

Schutter, E. (2009). *Computational Modeling Methods for Neuroscientists*. MIT Press.

Storn, R. and Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359.

Wohrer, A., Kornprobst, P., and Vieville, T. (2006). From light to spikes: A large-scale retina simulator. *Proceedings of the IJCNN 2006*, pages 8995–9003.

Zaghloul, K. and Boahen, K. (2006). A silicon retina. *Journal of Neural Engineering*, 3:257–267.

AUTHOR BIOGRAPHIES

TOMÁS H. MAUL was born in Madeira, Portugal and did a BSc. in Biological Psychology at the University of St. Andrews, an MSc. in Computer Science at Imperial College and a PhD. in Computational Neuroscience at the University of Malaya. He worked for two years at MIMOS Bhd. as a Senior Researcher in the fields of Pattern Recognition and Computer Vision. He is currently an Assistant Professor at the University of Nottingham Malaysia Campus, where he conducts research in the areas of Neural Computation, Optimization and Computer Vision. His e-mail address is `Tomas.Maul@nottingham.edu.my` and his Web-page is `http://baggins.nottingham.edu.my/~kcztm/`.

ANDRZEJ BARGIELA is Professor and Director of Computer Science at the University of Nottingham, Malaysia Campus. He is a member of the Automated Scheduling and Planning research group in the School of Computer Science at the University of Nottingham. Since 1978 he has pursued research focused on processing of uncertainty in the context of modelling and simulation of various physical and engineering systems.

His current research falls under the general heading of Computational Intelligence and involves mathematical modelling, information abstraction, parallel computing, artificial intelligence, fuzzy sets and neurocomputing.

LEE JUNG REN is currently pursuing his PhD in Computer Science at The University of Nottingham, Malaysia Campus. He did a Bachelors of Computer Science in the same university in 2007 where he developed a Medical Diagnosis Aiding application with Semantic Web for his Final Year Dissertation. His current research is focused on a neuroalgorithmic approach to retinal modeling.

STATISTICAL EXTRACTION OF PROTEIN SURFACE ATOMS BASED ON A VOXELISATION METHOD

Ling Wei Lee and Andrzej Bargiela
School of Computer Science
The University of Nottingham Malaysia Campus
Jalan Broga, 43500 Semenyih
Selangor Darul Ehsan, Malaysia

KEYWORDS
Protein Surface Atoms, Protein Surface Analysis, Space Voxelisation

ABSTRACT

Proteins play a vital role in maintaining the balance of bodily functions in all living beings. However their functional properties are difficult to predict since they depend not only on the sequence of constituent amino acids but also on the 3D folding of the protein. This paper presents a new statistical method for extraction of surface atoms of a protein. The method is based on space-voxelisation and generalizes our previous deterministic method by repeating the surface extraction process for various orientations of a protein; so as to achieve a statistical consensus about surface atoms. Based on the experimental study we have established an optimal range of values of voxel occupancy for the selection of surface atoms; with optimality defined as a maximum coincidence of extracted surface atoms from the protein presented to the algorithm in 13 different orientations. The results show that the voxel occupancy threshold of between >40% and >50% allows our algorithm to extract surface atoms with high degree of confidence.

INTRODUCTION

Proteins stemming from the same ancestor may undergo evolutionary change with some parts of the protein remaining unaffected while other parts modifying their structure. The assessment of the affinity of so evolved proteins is assessed through two broad classes of methods: a) methods that analyse amino acid sequences; and b) methods that analyse 3D structures of proteins.

Protein sequence comparison takes the sequence strings of proteins and attempts to align the two strings so as to find the largest common subsequence between the strings. There are many algorithms available to perform such alignment. Some algorithms are single sequence based like the Smith-Waterman algorithm [1] or Needleman-Wunch algorithm [2], while others attempted to perform multiple protein sequence alignments [3]. Pei [4] presented a paper reviewing methodologies and advances within this field. Systems like PSI-BLAST and CLUSTALW are commonly used by researchers for large scale comparison of protein sequences with each system having its own advantages.

Sequential comparison – depending on the type of algorithm implemented – can be very efficient computationally. However, what such comparisons cannot achieve is the identification of inheritance of functional properties between the compared proteins. A sequence order can only provide information such as the type of elements contributing to the protein and one has to make assumptions (or perform additional study) on how these changes affect the folding of a protein. However the derivation of protein folding from the basic sequence information is a very challenging problem due to the many degrees of freedom of the molecular structure.

Consequently, an alternative approach of experimental, crystallography-based discovery of 3D structures has established itself as a practical method of identifying overall protein structure. Consequently a new field of proteomic research has emerged, that of identifying common sub-structures in proteins that have similar functional properties. This research is now supported by the availability of databases containing a complete hierarchy of existing proteins classified based on their structures. The SCOP (Structural Classification Of Proteins) database [5] gives "a comprehensive ordering of all proteins of known structure according to their evolutionary and structural relationships". SCOP orders protein entries based on the following levels : Species, Protein, Family, Superfamily, Fold and Class. Another existing database that has been widely in use is the CATH (Class, Architecture, Topology, Homologous superfamily) database [6]. Class describes the secondary structure compositions of domains, Architecture defines the shape, Topology gives the sequential connectivity while Homologous superfamily groups together proteins with structures in the same Topology. Such classification databases are helpful to analysts and bioinformaticians studying the relationship between the proteins and the conservation of unique features in the structures.

However there have been only few attempts to analyse proteins in terms of their surface atoms structure and composition. According to [7], "protein surface comparison is a hard computational challenge and

Proceedings 24th European Conference on Modelling and Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

evaluated methods allowing the comparison of protein surfaces are difficult to find". Structural and sequence comparisons may reveal patterns that remain the same throughout evolution. Still, these signatures do not necessarily guarantee the same functions for the same evolutionary line of proteins. Protein surface comparison on the other hand may not be able to reveal inherited traits; however it is able to detect similar areas that provide the same reaction to external agents regardless of whether the protein comes from the same family. There may exist proteins from the same family containing binding sites with different characteristics while proteins from different ancestors may evolve to contain sites with similar features. Figure 1 shows the comparison for the binding sites.

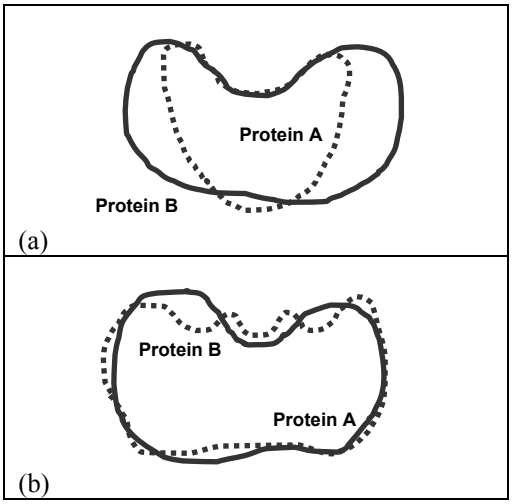


Figure 1 : Comparison of binding sites for proteins of
(a) Different ancestors with similar binding sites
(b) Same ancestor with different binding sites

As can be seen from the illustrations there is a need for protein surface analysis to determine the characteristics of potential dock sites. The following sections provide more detailed information on the algorithm developed and a brief description of past implementations and experiments.

BACKGROUND

One of the most commonly used methods for the study of protein surfaces is the Connolly method [8] whereby a water-molecule-sized probe is used to inspect the surface of the protein. Wherever the probe fits in the cavity it identifies with a high probability a potential binding site. In [9] Jiang and Kim used cube representations for the soft docking of proteins. By using two specific molecular complexes the authors show that geometric docking alone with conformational changes is sufficient to determine the correct binding agents. However this method has its limitations when applied to various protein complexes since there are proteins that do not form rigid docking bonds.

Cubic grid representations of proteins have been used in the past due to the ease of implementation and manipulation of the 3D structures. In this research we adopt this representation and attempt to overcome the inherent limitations of this topological construct. The basic algorithm for extracting surface voxels of a protein may be summarized as follows:

1. Pre-process the protein (details given in the next section) and compile the required information in a new file.
2. Impose the protein into the experimental cubic grid space using the compiled data.
3. Tessellate the experimental space until the smallest unit voxel size is achieved (in this case the value is 40 units).
4. Extract all voxels containing the presence of any of the atoms within the protein. This stage also includes a checking for the occupancy percentage of each voxel. For example, a >20% value means that voxels in which >20% of its space is occupied by an atom/groups of atoms.
5. Filter all voxels having 1 or more faces fully exposed. These are categorised as the surface voxels.
6. Based on the surface voxels the surface atoms are then extracted. This is done by compiling the atoms contained within the surface voxels.

The main problem encountered when using cubic grids relates to the orientation of the protein within the experimental space. Any arbitrary rotation within the grid space leads to a different set of voxels being chosen post-tessellation. At one particular orientation an atom may occupy an entire voxel while at another orientation of the protein the same atom may take up two or more voxels. Figure 2 provides an illustration.

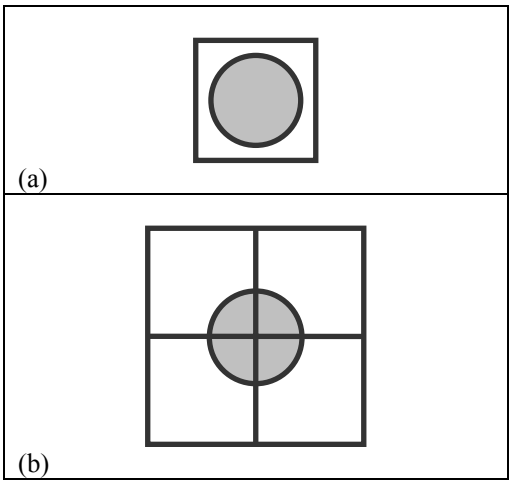


Figure 2 : Positioning of an atom before rotation (a) and after rotation (b). Initially the atom occupies 1 voxel and after rotation it is partially contained in 4 voxels.

As a consequence of the above, a deterministic identification of surface atoms for one orientation of a

protein typically returns results that are quite different from those obtained for another orientation. Hence, in this work, we relax the notion of deterministic identification of surface-atoms and introduce an alternative notion of statistical identification. This requires consideration of several orientations of the protein (which allows building of appropriate statistics) and the optimisation of the threshold value that determines when an atom can be considered as belonging to a specific voxel. Selection of a representative set of orientations of a protein is a compromise between the computational effort and the quality of the accumulated statistics. In the following section we provide a rationale for our selection. The determination of a suitable threshold for voxel occupancy has been performed as an empirical investigation (described in the following section) and the conclusions are deemed to be general.

THE ALGORITHM

The algorithm for the extraction of protein surface atoms extends a previous development [10]. Input files for the program are obtained from the RCSB Protein Data Bank (PDB) in PDB format. A pre-processing stage is carried out to extract the required information including the spatial coordinates, the atom element and all residues data following which these are stored into a new file together with additional information of Van der Waals radius and the electronegativity of the atoms. Compiling these data into a single file is important as the program has been configured and optimised to process all input files in the predefined format. In the original PDB file, the spatial coordinates provide the actual positioning of the atoms in the protein. However for the purpose of this research these coordinates have been scaled to larger values for ease of processing and analysis. For example, an atom is given the coordinates of (-8.371, 23.633, 40.487) in the PDB file. Upon pre-processing the new values for the atom are (139.35, 310.78, 541.72). Scaling up the values makes visual analysis easier while at the same it also ensures that important specifics do not get overlooked – small details that play important contributions are then easily identified. The Van der Waals radiuses for the elements lie between 1Å and 3Å and these values were also scaled up by ten-fold. Therefore the diameter range for the atoms is valued from 20Å to 60Å. By taking the estimated average of the diameter range it is thus concluded that a value of 40Å will be used as the value for the smallest unit voxel size. For easier understanding all future references to this smallest unit voxel size shall be termed as 40 units instead of 40Å so as to coincide with the notion of grid spaces and voxels.

In the algorithm the protein is first imposed into a cubic grid experimental space bound by Cartesian coordinates. Any object casted into this space is defined by a set of (x, y, z) coordinates. A checking is carried out to determine if any of the spatial coordinates of the atoms are in the negative range. If there exists any negative values, the protein is then translated into the positive regions. The spatial coordinates data are then stored into temporary memory for easy reference.

Following that the protein will be rotated based on a set of defined rotations in all X, Y and Z dimensions. The selected range of angles includes 20˚, 30˚, 45˚, 60˚ in all dimensions. These angles were chosen based on the understanding that any atom will experience the largest change of orientation at the aforementioned angles. The affine transform is implemented for the rotation of the protein. For a detailed explanation of the matrices involved the reader is referred to [10]. A reference point is first taken from the protein for use in the affine transform whereby all atoms are translated to the origin based on this reference point, rotated and re-translated back into the test space. However it should be noted that some post-rotation atoms may occupy voxels in the negative regions. Again, another round of checking is carried out to ensure that all the atoms are translated to the positive domains.

The transformed protein with all positive spatial coordinate values is then ensured that it is confined within the experimental space. A space tessellation algorithm based on the bisection is carried out to slice the experimental space into cubes of equal sizes with the smallest size being 40 units. This value is chosen based on a general analysis of the diameter of protein atoms and it was found that the average value agrees at a value of 40. A significant modification has been made to the original implementation to cater for the process of voxel occupancy checking. Previously any voxel containing the presence of any atom – regardless of the percentage of occupancy of the atom within that voxel – is shortlisted. For the purpose of extracting the surface voxels, a series of voxel occupancy percentage is introduced ranging from >0% to 100% with increments of 5%, therefore giving a total of 21 cases. The following table gives the list of experimental rotational conditions.

Table 1 : Various Rotations Used for Protein Experiments for Each Voxel Occupancy

Rotation Angles (x, y, z)
(0, 0, 0)
(20, 0, 0)
(30, 0, 0)
(45, 0, 0)
(60, 0, 0)
(0, 20, 0)
(0, 30, 0)
(0, 45, 0)
(0, 60, 0)
(0, 0, 20)
(0, 0, 30)
(0, 0, 45)
(0, 0, 60)

By multiplying the 21 cases with the list of 13 rotations there are thus a total of 273 experiments to be performed. In each iteration, the algorithm checks the percentage occupancy of voxels by protein atoms. All voxels meeting the specific threshold value are marked and stored for subsequent analysis.

To extract the surface atoms voxels are checked if they have exposed surfaces. If any of the surfaces of a voxel is not connected to any other voxel then it is categorised as a surface voxel. Figure 3 gives an illustration.

Surface atoms are then extracted by reference to the surface voxels identified in the previous step. The surface atoms information is then stored for the final stage of compiling statistics. Complete execution of all the experimental conditions is followed by a post-processing stage which collects all the generated output for analysis. The main challenge lies in ensuring the consistency of the extracted surface atoms across all different orientations. Common atoms shared by all rotation sets based on percentage of occupancy are shortlisted and the extraction accuracy is then calculated.

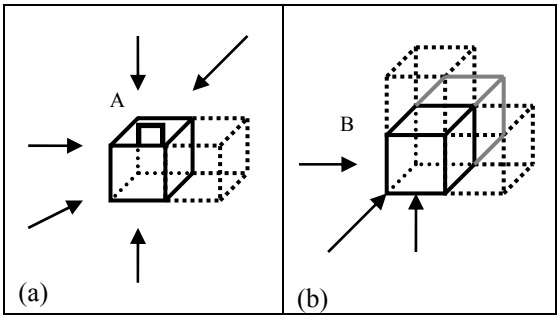


Figure 3 : (a) 5 faces exposed for voxel A. Therefore the exposure value is 5. (b) 3 faces exposed for voxel B. Therefore the exposure value is 3.

RESULTS

The algorithm is iterated 273 times to generate sets of output for each experimental condition. Each occupancy percentage contains 13 sets of results – one for every rotation in all X, Y and Z dimensions as well as for the original non-rotated file.

For each occupancy percentage, atoms that are common across all 13 orientations are first extracted. The common atoms are then compared to the total number of extracted surface atoms for each rotation and the percentage is calculated. Figure 4 gives the plot of the percentage of common atoms versus the occupancy percentage of voxels. There are 13 lines altogether in the graph – each line represents each of the different orientations.

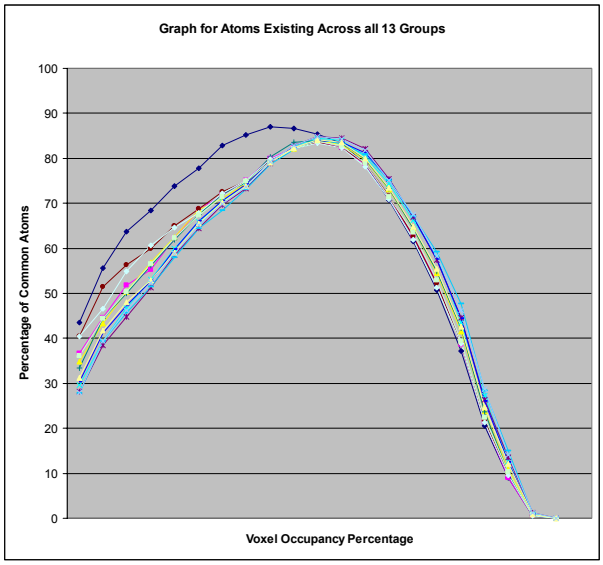


Figure 4 : Plots for each rotation giving % of common atoms identified in all 13 orientations

The horizontal axis begins with 0% voxel occupancy to 100% voxel occupancy. As can be seen from the graph the extracted common atoms peaks at about 50%. The following table gives a list of all the rotation cases with their corresponding peak percentages.

Table 2 : List of all rotation cases with their peak percentages and extraction percentages for common atoms in all 13 orientations

Rotation Case (x, y, z)	Peak Percentage (%)	Extraction Percentage (%)
(0, 0, 0)	40	87.06
(20, 0, 0)	50	83.55
(30, 0, 0)	50	83.96
(45, 0, 0)	50	84.52
(60, 0, 0)	50	84.84
(0, 20, 0)	50	83.55
(0, 30, 0)	50	83.89
(0, 45, 0)	50	84.00
(0, 60, 0)	50	84.24
(0, 0, 20)	50	83.45
(0, 0, 30)	50	83.72
(0, 0, 45)	50	84.03
(0, 0, 60)	50	84.84

From the graph and the given table it can be concluded that at >50% voxel occupancy the extraction of the protein surface atoms reaches its optimum. At the same time, the voxels processed are fewer in number and hence this increases the processing speed. Thus it can be said that a compromise has been reached at >50% for the minimum number of voxels used while retaining the best surface atoms extractions.

However there are still some surface atoms that may be left out due to differences in the orientations. The condition for clustering of the surface atoms is therefore relaxed and atoms existing across 12 and even 11 cases are taken into consideration as these are atoms with high probabilities of existing on the surface and which may play significant roles. Atoms that have been marked as existing across all 13 cases are classified as definite entries. The following tables show the output obtained for two relaxed scenarios.

Table 3 : List of all rotation cases with their peak percentages and extraction percentages for common atoms in 13 and 12 out of 13 orientations

Rotation Case (x, y, z)	Peak Percentage (%)	Extraction Percentage (%)
(0, 0, 0)	40	97.02
(20, 0, 0)	45	95.16
(30, 0, 0)	50	95.11
(45, 0, 0)	50	95.41
(60, 0, 0)	50	95.80
(0, 20, 0)	50	94.72
(0, 30, 0)	45	95.93
(0, 45, 0)	50	95.39
(0, 60, 0)	50	95.17
(0, 0, 20)	45	94.85
(0, 0, 30)	50	95.28
(0, 0, 45)	50	95.76
(0, 0, 60)	50	96.18

Table 4 : List of all rotation cases with their peak percentages and extraction percentages for common atoms in 13, 12 and 11 out of 13 orientations

Rotation Case (x, y, z)	Peak Percentage (%)	Extraction Percentage (%)
(0, 0, 0)	40	99.05
(20, 0, 0)	45	97.99
(30, 0, 0)	45	97.67
(45, 0, 0)	50	97.90
(60, 0, 0)	50	98.38
(0, 20, 0)	45	97.71
(0, 30, 0)	45	98.38
(0, 45, 0)	45	98.04
(0, 60, 0)	45	97.88
(0, 0, 20)	45	97.87
(0, 0, 30)	45	98.16
(0, 0, 45)	45	98.65
(0, 0, 60)	45	98.68

The relaxation of the condition leads to a higher number of atoms being classified as surface atoms. From the tables it can be gathered that the best extractions for all rotation cases exist between >40% and >50% of voxel occupancy. This corresponds well to the strict filtering of the atoms existing across all 13 cases whereby the optimum occupancy percentage is >50%.

The experiment has been repeated for conditions with the inclusion of atoms existing across 13, 12, 11, 10 and 9 cases and the optimum voxel occupancy percentage for the best extractions remains the same. Also, it can be concluded that the relaxation of the condition leads to higher extraction percentages, of which will gradually becomes closer to 100% as more and more atoms are included.

DISCUSSION

The use of voxels in the extraction of protein surface atoms proves to be an effective and efficient method. The tessellation of the experimental space is computationally efficient and it is easy to determine whether a voxel contains any atom. Rotation was used to generate various orientations of the protein and this was chosen over translation due to the challenge posed. Affine transformations are involved in calculating the new positions of atoms when a protein has been rotated. The working hypothesis was that a rotational transform affects various sections of the protein to different degrees and this is largely dependent on the distance of atoms from the center of the protein. Future work will verify this hypothesis by comparing its performance against alternative transformations, e.g. discrete translation along all x, y, and z axes.

To determine the surface atoms one needs to identify surface voxels which involves only checking 6 surfaces of individual voxels and can be implemented very efficiently. It was found that the best extraction of voxels occurs with the occupancy threshold set to between >40% and >50%. This value is deemed to be a function of the topology of the voxels, so as long as one uses cubical voxels this threshold can be used in subsequent studies to produce the optimum surface atom extraction rates.

As the filtering condition is relaxed more atoms are included thus increasing the number of surface atoms being shortlisted. The extraction percentages show consistency in between different orientations. This correlates well with the relaxation of the condition. The algorithm executes efficiently with each iteration taking less than 10 seconds. Generally the surface atoms of a protein can be determined in under 2 minutes on a 2GHz single core PC when the optimum occupancy percentage is used.

The algorithm does not yield a 100% accurate extraction but it does provide satisfactory results with up to 97% accuracy. The atoms extracted are highly probable surface atoms but at the same time, there is also a possibility that an extracted atom may belong to the layer below the surface layer. However this does not pose any major problems to the analysis stage. At this stage the results are validated through visual inspection

of the protein images since the PDB does not hold the surface atoms information.

A good extraction of the surface atoms aids greatly in the study of the features and characteristics of a protein. By applying methods that determine relationships between the atoms, the motifs and key signatures can be identified especially the composition that contributes to the attributes of protein docking sites. Predictions can be made of proteins with similar motifs that may bind to the same ligands or binding agents. Furthermore, a classification system can be constructed that clusters proteins with similar signatures into the same groups. Such a system is able to provide new discoveries and insights on the functions of proteins. Bioinformations will find this system useful in the development of new drugs as the system is able to indicate side effects on other proteins that may caused by a particular drug.

This approach is to be implemented in full on protein sets obtained from the PDB. The output obtained will then be used for the analysis of protein surfaces in upcoming research work.

CONCLUSION

A new approach for the identification and extraction of protein surface atoms is presented here in which voxels are used as the main tool. A range of experimental conditions were used whereby the protein was tested in different orientations. The occupancy percentages of the voxels were checked and it was found that the optimal extraction occurs at somewhere between >40% and >50% of occupancy. Atoms existing across all 13 orientation cases display a fairly high extraction percentage at about 84%. As the condition is relaxed and atoms existing across both 12 and 11 orientations are included it can be seen that the extraction percentages increased to about 95% and 97% respectively. This approach is efficient and is able to produce a good output through the use of voxels. All extractions of surface atoms are to be used for analysis in ongoing research work.

REFERENCES

[1] Smith T., Waterman M. 1981. "Identification of Common Molecular Subsequences". *Journal of Molecular Biology*, No.147, 195-197.

[2] Needleman S., Wunsch C. 1970. "A General Method Applicable to The Search for Similarities in The Amino Acid Sequence of Two Proteins". *Journal of Molecular Biology*, No.48, 443-453.

[3] McClure M.A., Vasi T.K., Fitch W.M., 1994. "Comparative Analysis of Multiple Protein-Sequence Alignment Methods". *Molecular Biology and Evolution*, No.11 571-592.

[4] Pei J. 2008. "Multiple Protein Sequence Alignment". *Current Opinion in Structural Biology*, No.18, 382-386.

[5] Andreeva A., Howorth D., Chandonia J-M., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G., 2007. "Data Growth and Its Impact on The SCOP Database: New Developments". *Nucleic Acids Research*, 1-7.

[6] Orengo C.A., Michie A.D., Jones S., Jones D.T., Swindells M.B., Thornton J.M., 1997. "CATH – A Hierarchic Classification of Protein Domain Structures". *Structure*, No.5, 1093-1108.

[7] Via A., Ferre F., Brannetti B., Helmer-Citterich M., 2000. "Protein Surface Similarities: A Survey of Methods to Describe and Compare Protein Surfaces". *Cellular and Molecular Life Sciences*, No.57, 1970-1977.

[8] Cao, J., Pham D.K., Tonge L., Nicolau D.V., 2002. "Predicting Surface Properties of Proteins on The Connolly Molecular Surface". *Smart Materials and Structures*, No.11, 772-777.

[9] Jiang F., Kim S.H., 1991. "'Soft Docking': Matching of Molecular Surface Cubes". *Journal of Molecular Biology*, No.219, 79-102.

[10] Lee L.W., Bargiela A., 2009. "Space-Partition Based Identification of Protein Docksites". *Proceedings of the 23rd European Conference on Modelling and Simulation (ECMS 2009)*, 848-854.

AUTHOR BIOGRAPHIES

LING WEI LEE was born in Kuala Lumpur and studied at the University of Nottingham Malaysia Campus where she took up Computer Science and obtained her honours degree in 2007. She worked for about a year as an analyst programmer with a local company before deciding to pursue her postgraduate studies. Her current research focuses on the applications of granular computing methods in the area of proteins. She is currently in her second year of research. She can be reached at Ling-Wei.Lee@nottingham.edu.my.

ANDRZEJ BARGIELA is Professor and Director of Computer Science at the University of Nottingham, Malaysia Campus. He is member of the Automated Scheduling and Planning research group in the School of Computer Science at the University of Nottingham. Since 1978 he has pursued research focused on processing of uncertainty in the context of modelling and simulation of various physical and engineering systems. His current research falls under the general heading of Computational Intelligence and involve mathematical modelling, information abstraction, parallel computing, artificial intellingence, fuzzy sets and neurocomputing.

AUTHOR INDEX