Demographic and educational projections. Building an event-oriented microsimulation model with CoMICS II

Marc Hannappel

Institut für Soziologie

Universität Koblenz-Landau
56070 Koblenz, Deutschland
Email: MarcHannappel@uni-koblenz.de

Klaus G. Troitzsch
Simone Bauschke
Institut für Wirtschafts- und Verwaltungsinformatik
Universität Koblenz-Landau
56070 Koblenz, Deutschland
kgt@uni-koblenz.de

KEYWORDS

discrete-event-oriented microsimulation, survival function, graduate rates, fertility rates

ABSTRACT

German demographic or educational projections are conventionally based on macrosimulation or period microsimulation models. In this paper we introduce a discrete-event microsimulation model, which we have designed to project the graduate rates of the German population. In this model we use survival functions to calculate different fertility rates by mothers' education. This paper describes how to implement the empirical results into the discrete event microsimulation model CoMICS II and discusses further steps to create a microsimulation model which could demonstrate the development of the German graduate rate considering different strengths of social selectivity mechanisms in the German education system.

INTRODUCTION

Demographic and educational projections are an important basis for policy decisions (Mannion, et.al. 2012). The Australian National Center of Social and Economic Modeling (NATSEM) for example calculates scenarios of the future development of the australian society on the basis of numerous parameters (Kelly & King, 2001). Politicians can then use these projection results to develop political concepts or to reform tax systems or pension schemes. The German Statistical Office (GSO) uses macrosimulation techniques to calculate the development of the German society for the next 50 years. These projections focus on the prediction of the absolute number of the future population and its composition according to age and gender. For instance, the conference of the regional ministers of education uses these results to calculate the educational attainment of the future German population (Kultusministerkonferenz, 2005).

The examples focus on the political fields of application of simulations techniques. However, this view neglects the epistemological gain of simulation models. Conventionally, one assumes microsimulation models are only used to predict social phenomena whereas agent based models are used to explain social phenomena (Spielauer, 2009) (Walker, 2010).

In this paper we present a microsimulation model which will be able to generate theoretical assumptions about interaction effects between demographic processes and social selectivity mechanisms within the German education system. The construction of the model is not finished yet. Therefore, the paper is focused on the presentation of the demographic modules and results.

Correspondingly, we first give an overview of the content frame of our project. This is followed by a method section where we present our microsimulation approach, the database, the survival functions which we use to calculate the biographic events and our model structure. With an example we then show how to implement the empirical results into a simulation module. Finally we present the current simulation results and discuss further steps.

1 DEMOGRAPHIC CHANGE AND SOCIAL SELECTIVITY

Beside the political discussion how to overcome the financial crisis two developments dominate the German scientific and political discourse about the future challenges of German society: the demographic change and the inequality within the German education system. The former is characterized by a successive decrease of the fertility rates during the past decades (Peuckert, 2008). No other country of the western industrial countries has such a low fertility rate as the Federal Republic of Germany. A detailed examination of the development of the birth rates leads to an interesting result. The probability to give birth to a child varies between women with different educational status. The higher the educational level of women the lower the probability to give birth to a child (Peuckert, 2008).

The second development affects the German education system. The PISA-Studie (PISA-Konsortium Deutschland, 2007) shows a relation between the educational status of parents and the educational attainment of their offspring. Also in this issue Germany stands out from

Proceedings 26th European Conference on Modelling and Simulation ©ECMS Klaus G. Troitzsch, Michael Möhring, Ulf Lotzmann (Editors)
ISBN: 978-0-9564944-4-3 / ISBN: 978-0-9564944-5-0 (CD)

other countries. The correlation between the social background and the educational attainment of children is larger in Germany than in any other country.

We summarize: First, the German society is characterized by a negative correlation between the educational status of women and their fertility rates. Second, the German education system is characterized by a positive correlation between the educational status of parents and the educational attainment of their offspring. This leads us to the question: Wich consequences do these developments have on the future social composition of the German society?

To find an answer to this question we developed a dynamic event-oriented microsimulation model which will be described in the following.

2 METHOD

2.1 Event oriented microsimulation

Dynamic micro-analytical simulation models can either be modelled with a period approach or with an event-oriented approach (Gilbert & Troitzsch, 2005). Wheras period-oriented simulations model time as passing between equidistant points of time usually one year apart, in event-oriented microsimulation models time between two events is modeled as a random variable whose distribution is estimated from empirical data. This time interval is calculated by survival functions which parameterise the events.

This method allows to implement different kinds of calculations, which determine time as an interval until an event occurs. In contrast to period-oriented models, which import each individual in every time period into the simulation to examine whether an event will occur or not, event-oriented approaches determine the time when an event will occur (Gilbert & Troitzsch, 2005).

In event-oriented microsimulation models, individuals have to pass a module when all characteristics are available which are necessary to calculate a random value for the time until the next event of a certain type. A random number generator yields a random number from a [0,1] uniform distribution (monte-carlo experiment). A simulation algorithm compares the random number with an empirical value to calculate if an event occurs or not (for more detailed information about the process of event oriented simulation models see section 2.5)

2.2 Discrete-event microsimulation with CoMICS II

The projection of prospective German graduates in CoMICS II is a Java-based event-oriented microsimulation software. Its key elements are a consistent time system and management, a random number generator, various classes of events, an event calendar similar to the crystal ball of DYNAMOD (Abello & King, 2002) and, last but not least an external data base which keeps the model individuals and their attributes. When the simulation is initialised, all model individuals calculate the time when their next events will occur according to the survival functions applicable for these events (see the example below). The event calendar contains these calculated event dates for all individuals, and the event manager accesses this calendar and executes one event after another, and these events will change the state of the respective individual or individuals.

2.3 Data

Micro simulation models are conventionally based on representative data sets with an adequate number of individuals and attributes, so it is necessary to describe the data set which we use in the simulation. The microcensus of the German Statistical Office is the only data set in Germany which satisfies the conditions on microsimulation models: a large number of micro units, variables which include the characteristic of the micro units and a household structure which enables to link the members of a household.

The microscensus is a representative sample (annual 1% census) of the German population which includes questions about socio-economic issues. Whereas the microcensus is focused on information about labour market participation, the data set includes a number of issues about family, education and fertility. For the simulation a 70% subsample of the original microcensus – Scientific Use File (SUF) 2008 – is available, which includes a sufficient number (477,239) of individuals, attributes and dates, so that the SUF can be used for estimating the survival functions.

2.4 Survival functions

Survival functions calculate the probability of individuals to "survive" the time until an event occurs. In this context the special interest of fertility analyses consists in the calculation of the difference between the time from a starting point and the time when a woman gives birth to her first child.

The event "birth of first child" is operationalised by the variable "age at first birth". Conventionally, women are analysed by their age at first birth. The difference between their own year of birth and the age when they give birth to their first children is then used as a parameter for the time until the event "birth of first child" occurs. However, the 2008 microcensus included only information about children living in their parents' household. As a consequence it is not possible to calculate the age of first birth for women whose children have left the household. To avoid this problem, actual results could be achieved by analysing the age of first birth from the children's perspective (Spielauer, 2003), i.e. not the mothers are the subjects of the analysis, but the first born children are analysed by the age of their mothers when they were born.

To do so, we use survival functions to calculate the probability of when a woman will give birth to her first

child.

$$S(t) = P(T > t) \quad (1)$$

$S(t)$ = survival function; probability that a person "survives" (in this case: remains childless) at least until time t
T = random variable (event date)
t = specified time
$P(T > t)$ = probability that the specified time is shorter than the random variable T. The person "survives" the specified time, i.e. the event will not occur.

As mentioned above, the analyses are focused on children cohorts. Therefore the results of the survival function would be zero for 50-year-old mothers. In reality not every woman gives birth to a child, so it is necessary to adjust the survival function by (1- "rate of childlessness").

$$S_i'(t) = S_i(t)(1 - P_i(k)) \quad (2)$$

$S_i'(t)$ =survival function at time t of women with qualification i, reduced by the rate of childlessness by education
$S_i(t)$ = survival function at time T by qualification i of women
$P_i(k)$= probability of childlessness from women with qualification i

Table 1: Age specific risk $h(t)$, cumulative risk $c_{h(t)}$, cumulative survival rate $c_{S(t)}$ and adjusted cumulative survival rate c from 15–50 year-old mothers with a university degree to have a first child.

1	2	3	4	5
age	$h(t)$	$c_{h(t)}$	$c_{S(t)}$	c
15	0	0	100	61000
(...)	(...)	(...)	(...)	(...)
29	7.1	14.02	77.02	46979
30	8.4	19.13	68.64	41871
31	8.3	24.17	60.37	36826
32	9.0	29.66	51.37	31335
33	10.6	36.11	40.81	24984
34	10.0	42.23	30.78	18775
35	8.5	47.42	22.25	13575
36	5.7	50.88	16.58	10116
37	5.0	53.92	11.61	7083
(...)	(...)	(...)	(...)	(...)
50	0.0	61.00	0.00	0
$n = 1639$ weighted (unweighted: 801)				

Source: own calculation based on SUF 2005, cohorts 2002–2005 of first-born children

Table 1 shows the empirical values – risk and survival rates – of women with a university degree by age. The first column lists the age of the women and the second column lists the hazard rates by age. The risk of a first birth for women with a university degree at the age of 30 amounts to 8.4 %. The cumulative risk for these women, which is listed in column 3, amounts to 19.13 %. Compared to columns 2 and 4, columns 3 and 5 include the rate of childlessness. Women aged 50 have a 61 % cumulative risk to have given birth to their first child, unless they remain childless (39 %). The cumulative survival rate ($c_{S(t)}$) for 30-year-old women with a university degree of the analysed cohort amounts to 68.64 %. The adjusted cumulative survival rate is listed in column 5 (c), and these values will be used in the simulation, i.e. the adjusted cumulative survival rate (percentage) multiplied by 1000 – we draw uniformly distributed integer random number between 0 and 100,000. The survival rates have to match with this value range. The values of this column are the survival functions which we have converted according to the life table method and then listed in the event tables. As mentioned below, the event dates of the "age at first birth" are the results of the comparison between the empirically estimated survival function values from column 5 and a number drawn randomly by the random number generator. The idea of microsimulation is that a large number of monte-carlo experiments lead to the result that a certain percentage – in the case of the example 39% – of the random numbers will be larger than a defined "critical" value (61,000). Women with a random number larger than 61,000 will remain childless (for a more detailed description see section 2.5).

2.5 Modules

The simulation of the German graduates includes many different modules (Figure 1) with socio-demographic or socio-cultural tables (Table 1). The module "Death" is the first module the individuals have to pass through. The simulation algorithm calculates the exact time of death for every individual before the simulation starts or when an individual is born during the simulation. The time of death is a result of a comparison between the survival function value which is taken from the life tables of the German Statistical Office, and a random number, which is drawn for every individual (the example of the module "Child 1" will give a detailed overview of the procedure of the simulation algorithm). Within the module "Education" the simulation algorithm decides which kind of educational attainment will be assigned to a child considering his or her parents' education. This module takes into account the different social selectivity mechanisms within the German education system. When a woman gives birth to her first child, a partner will be matched by education and age in the "Partnership" module. As a consequence, fertility is modelled independently of partnership (Spielauer, 2003). The module "Birth" consists of three different submodules ("Child 1", "Child 2" and "Child 3"). As an example we will discuss the module "Child 1": As mentioned above, we model time as a ran-

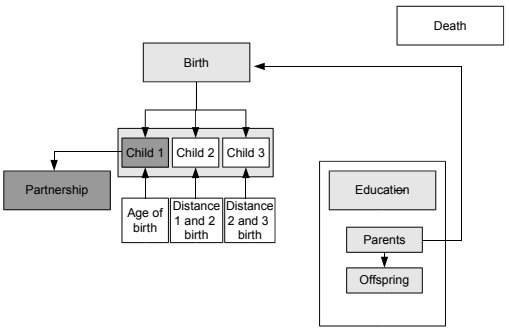


Figure 1: Overview: modules.

dom variable between the simulation start and the time when an event occurs. The module "Child 1" calculates

the day when a certain woman will give birth to her first child. First of all the simulation algorithm selects only female persons from the data set and ascertains the age of the women. Afterward the calculation process is separated in two steps. In the first step the simulation algorithm draws a random number for every woman. This number will then be compared with the first value of the event tables (61,000 for women with a university degree). If the random number is larger than this value the women will remain childless, otherwise the date of her first child's birth will be calculated. In the latter case these women have to pass the second step. A second random number in a range between the age specific value of the women's current age and zero will be drawn for every woman. This number will then be compared to the age specific value of the selected woman. If the random number r is bigger than the age specific value v_x, this woman will stay childless. If $r < v_x$, the simulation algorithm will compare r with the age specific value of women at age $x+1$ (starting age + 1). This process will be repeated until an age specific value is found which is smaller than the random number of the woman. This value represents the "year of her life" in which this woman will give birth to her first child. The year x in which the woman gives birth to her first child is determined by $(v_x > r > v_{x+1})$. The exact date of birth within this year is given by formula 3:

$$D(b)_x = \left(\frac{v_x - r}{v_x - v_{x+1}}\right) \times 365.25 \quad (3)$$

$D(b)_x$ = number of days until birth after the mother's x-th birthday
v_x = cumulative frequency of first children born before a mother's x-th birthday
v_{x+1} = cumulative frequency of first children born before a mother's $(x+1)$-th birthday

r = a uniformly distributed random number

$D(b)_x$ is the number of days after the woman's birthday until the event "birth of first child" occurs. The next step is to calculate the days from the starting age of the woman – the age at the start of the simulation – until the exact day of the event "first birth":

$$D(b) = D(b)_x + [(t_b - t_s) \times 365] \quad (4)$$

$D(b)$ = survival time (number of days until birth since the woman's starting age)
$D(b)_x$ = number of days until birth after the mother's x-th birthday
t_b = age of first birth
t_s = starting age

Finally, the date of first birth is compared to the woman's date of death ($D(d)$) which has been calculated in the module "Death" before. If $D(b) < D(d)$, the event will occur and the date of birth of the first child will be registered in the event calendar. The example in Figure 2 gives an overview of the simulation process:

The calculation processes of all other events is mainly done in the same way. Single events can trigger other events. As a consequence of this not all events will be calculated before the start of the simulation. Following the example, it is not the complete birth biography of women that is calculated at the beginning of the simulation. First, the age of first birth, operationalised as the time until first birth, as mentioned above, will be calculated and afterwards the woman who gives birth to a first

simulation start	1.1.2008	age of the woman =29
first step:		
1. random number	55,321	55,321 < 61,000
	⇒	the woman will give birth to a child
second step:		
2. random number	(age 29)	46,979 → 0
draw		
random number	17,934	$v_{34} > r > v_{35}$
of the woman		$18,775 > 17,934 > 13,575$
Value for age 34	18,775	⇒ *woman has her first*
Value for age 35	13,575	*child at the age of* 34

Calculation of the number of days between the simulation start and the date of the event:
$D(b)_x = \left(\frac{r - v_{34}}{v_{35} - v_{34}}\right) \times 365 = \left(\frac{17,934 - 18,775}{13,575 - 18,775}\right) \times 365 = 0,16 \times 365 = 58,4$

$D(b) = D(b)_x + [(t_b - t_x) \times 365] = 58,4 + [34 - 29] \times 365 = 1,883$

01.01.2008 (simulation start) + 1,883 (days until the event will occur) =27.02.2013

Figure 2: Example

child will pass the module "Child 2", which is operationalised by the time between the first and the second child. All event dates which are calculated this way are registered in the event calendar and executed in the simulation. If an event happens, the program will delete this event from the event list.

The first simulation runs have shown that the results of discrete-event-oriented microsimulation processes are similar to those of period-oriented microsimulation models. Simulations with the full number of 484,422 individuals of the 2008 microcensus will be carried out in the coming weeks.

3 RESULTS

Figure 3 shows our current simulation results and the results of the macrosimulation of the German Statistical Office (GSO) (StaBu, 2006). Although it is not possible to validate projection results, we use the comparison between the results of the GSO and our results to check whether our results are plausible or not. The German Statistical Office calculates 12 different scenarios. The results are taken from a scenario which did not consider migration processes and uses the current fertility and death rates as constant. The assumptions of this scenario are similar to the assumptions within our simulation. The dashed line represents the percentage difference between the different results. The difference between the "official" standard scenario and our simulation is in the range of 3 and -2 percentage points for the population until 80 years. Only for the population older than 80 years, we can observe an increase of the difference up to 7% which is comensated by a negative difference for the population older than 95 years.. The difference shows a very good accordance between both simulation results.

Figure 4 illustrates the number of births between 1951 and 2057. The vertical line separates the real population – the people who really exists before the simulation start – and the simulated population. The result shows a plau-

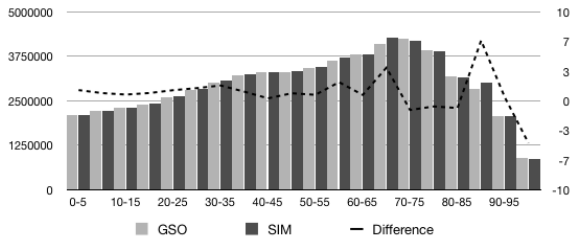


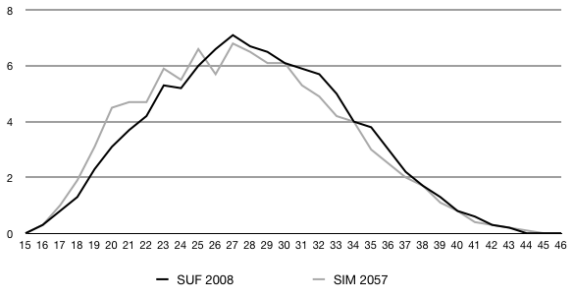
Figure 3: Results for 2057: Difference between the projection of the German Statistical Office and the results of the Simulation

sible development of the birth rates for the next 50 years. The final simulation will only project the population for the next 20 or 30 years. However, to test the simulation algorithm it is necessary to apply a longer simulation time to preclude implausible developments. The decrease of births is caused by the fact that we use a closed microsimulation model, i.e. we don't consider migration processes.

Figure 4: Number of birth between 1951 and 2057

A comparison between the results of the analyses of the age at first birth of the Scientific Use File (SUF) 2008 and the simulation results shows also consistent results (see Figure 5) . We use this comparison to control the simulation algorithm. The parameters of the action decisions which we implemented in the simulation are a result of the analysis of the SUF 2008. Consequently, the frequencies of the variable "age at first birth" of the simulated dataset should be similar to the analyses of the original dataset. The observable small difference is caused

Figure 5: Age of first birth: SUF 2008 vs. simulation results



by two reasons. First, we use a 10% subsample of the original dataset to test our simulation algorithm. As a consequence we have a small variance of the simulation results. Second, the results are based on analyses on an aggregate level. However, the survival rates of the age at first birth are calculated depending on the educational status of the women. The graph in Figure 5 shows the age of first birth for all mothers. The average age at first birth is one year smaller for the population of the simulated dataset (the grey line in Figure 5). An explanation of this development is given in Figure 6, which shows the development of the educational attainment of the population. The cohort 1962 - 1982 represents people who had completed their school training before the start of the simulation. Therefore, the development of the educational attainment of people who were born until 1982 shows the so-called "German educational expansion" which is characterised by an increase of the percentage of higher educated people (Becker, 2011). Whereas the percentage of this subpopulation stays constant during the simulation the percentage of the population with a middle secondary degree decreases and the people without a school degree increases dramatically. The development of the

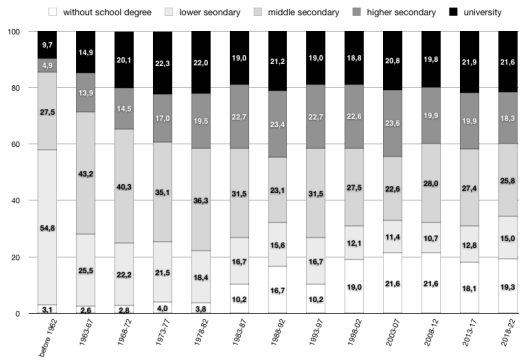


Figure 6: Educational attainment of the cohorts < 1962 to 2022

simulated educational attainment is currently not sufficiently realistic. First, we haven't yet considered transitions between vocational school and university. These transitions are significant for the German education system. Therefore, our model underestimates the development of the people with a university degree. Second, we have a large number of young people without a school degree in the original dataset. These are people who are still in school and listed as "without a school degree". These two issues will be solved in the next weeks. Currently, we are analysing a new dataset which includes information about the transition from high school/ vocational school to a university.

Furthermore, Figure 5 and 6 give an idea how to answer the question of the reciprocal influence of demographic and educational developments on the social composition of the German population. The population without a school degree is characetrised by a high fertility rate and – in contrast to other subgroups – a lower average

age of first birth. The increase of this subpopulation in our current simulation scenario leads to a decrease of the average age of first birth for the entire population. Additionally, the high fertility rate and the low social mobility of this subpopulation leads to an increase of the percentage of people without a school degree (see Figure 6).

For the final results we expect a lower increase of people without a school degree and an higher number of people with a university degree. By the use of different scenarios we are then able to demonstrate the reciprocal influence of the various developments.

4 DISCUSSION

The results are based on a 10 % subsample of the original 2008 data with only 47,000 cases to save computing time during the test phase. Because of the stochastic processes within the simulation, the sample size has an influence on the simulation results: The lower the sample size, the higher the variation of the results. However, the use of a small sample is necessary because the single simulation runs need a lot of time. With the small sample it is possible to test the simulation algorithm in much shorter time. Therefore we assume stabler results with the complete module structure and with the original dataset (484,422 cases).

The advantage of microsimulation models is the flexibility of the model assumptions. The simulated dataset includes all variables of the original empirical dataset and describes the structure of the population in any respect as it can be analysed the same way as the microcensus data set. Instead of the interpretation of the absolute number of simulated cases and very superficial view on the structure of the total population, such as in macromodels, it is possible to analyse the data in many different ways. The dataset includes information about the age at first birth, the educational status of the individuals, and all relations between members of the same family. We can analyse the data to check the particular simulation algorithms and to analyse the interaction between the implemented parameters. It is possible to show the influence of the correlation between fertility processes and the educational attainment of women, and additionally it is possible to analyze the consequences of this correlation on the future social composition of the society. Beside empirical methods, microsimulation models are a specific instrument to analyse the interaction between different strands of development.

REFERENCES

Abello, A.King, A. (2002). Demographic Projections with DYNAMOD-2. *National Centre for Social and Economic Modelling, Canberra*

Becker, R. Hadjar, A. (2011). Erwartete und unerwartete Folgen der Bildungsexpansion in Deutschland. *In Becker, R.: Lehrbuch der Bildungssoziologie. Auflage: 2. VS-Verlag, Wiesbaden. S. 195 - 214.*

Gilbert, N.,Troitzsch, K. G. (2005). Simulation for the Social Scientist. *Open University Press, London*

Kelly, S.,King, A. (2001). Australians over the coming 50 years: providing useful projections.. *National Centre for Social and Economic Modelling, Canberra*

Kultusministerkonferenz (2005). Prognose der Studienanfänger, Studierenden und Hochschulabsolventen bis 2020. *Bonn*

Mannion, O. Lay-Yee, R. Wrapson, W.Davis,P. Pearson, J. (2012). JAMSIM: a Microsimulation Modelling Policy Tool. *Journal of Artificial Societies and Social Simulation 15 (1) 8 http://jasss.soc.surrey.ac.uk/15/1/8.html*

Leim, I. (2008). Die Modellierung der Fertilitätsentwicklung. *Metropolis-Verlag, Marburg*

Peuckert, R. (2008). Familienformen im sozialen Wandel. *VS Verlag, Wiesbaden*

PISA-Konsortium Deutschland, (2007). PISA 2006. *Waxmann, Münster*

Spielauer,M. (2003). Family and Education. *Austrian Institute for Family Studies, Wien*

Spielauer, M. (2009). Microsimulation Approaches. *Statistics Canada*

Statistisches Bundesamt (2006). Bevölkerung Deutschlands bis 2050. *Wiesbaden*

Walker, L. (2010). Modelling inter-ethnic partnerships in New Zealand 1981-2006: a census-based approach. *University of Auckland*

AUTHOR BIOGRAPHIES

Marc Hannappel was born 1980 in Bad Schwalbach, Germany. He studied educational research at the University of Koblenz-Landau and obtained his degree 2006. From 4/2008 to 9/2010 he was a scholarship holder of the Hans-Böckler-Stiftung. Since 10/2010 he has been working at the Institute of Sociology at the University of Koblenz-Landau. His email is `MarcHannappel@uni-koblenz.de`

Simone Bauschke was born 1975 in Cottbus, Germany. Since 2001 she has studied computer science at the University of Koblenz-Landau. She has been working since 2009 as a research assistant at the Institute of Information Systems in Business and Public Management . Her email is `ehrenbsi@uni-koblenz.de`

Klaus G. Troitzsch was born in Lahstedt, Germany, in 1946. He studied sociology and political science at the University of Hamburg from where he received his PhD in 1979. From 1979 until de retired in 2012 he has worked at the University of Koblenz- Landau. In 1986 he was appointed professor of computer applications in the social sciences at the Institute of Information Systems in Business and Public Management where he leads research projects in the field of social simulation. His email is `kgt@uni-koblenz.de`