INVESTIGATION OF ACCIDENT BLACK SPOTS ON LATVIAN ROADS USING SCAN STATISTICS METHOD

Vitalijs Jurenoks, Vladimirs Jansons, Konstantins Didenko
Faculty of Engineering Economics
Riga Technical University
Kalku iela 1, Riga, LV1658, Latvia
vitalijs.jurenoks@rtu.lv
vladjans@latnet.lv
ief@rtu.lv.

KEYWORDS
Logistics system, accident black spots, scans statistics, modelling, Monte-Carlo method, system optimization.

ABSTRACT

Scan statistics is an instrument of research of statistical dynamics of the development of the object under investigation in space and time [1] – [5]. Research is made of statistics of road accidents distributed in space and time in the territory of Latvia. Alongside with the Monte-Carlo method when calculating the p-value, it is possible to use the direct method of scan modelling, taking into consideration the empirical rules of distribution of road accidents on the highways of Latvia. In this case both parametric and non-parametric methods may be used to describe the distribution of road accidents in Latvia. The methodology of research of road accidents in space and time enables to detect the most dangerous places on the highways of Latvia, check the efficiency of the decisions taken to improve the quality of roads and allocate the resources of the socio-economic system in a more rational way.

BACKGROUND

In most cases the possibility of traditional statistical conceptions and methods for investigation of real socio-economical objects is limited (the grey zone in Figure 1). Traditional statistical methods are more appropriate for investigation of the influence of internal factors – for the investigation of localized (grey zone) objects. Scan statistics allows investigating the socio-economical problems having extremely complex, i.e. synergic structure of interrelations (structure of the open systems). The analytical description of such systems in a simplified way enables to consider the likelihood scenarios of development of the object under investigation, but rarely of an object as a whole.

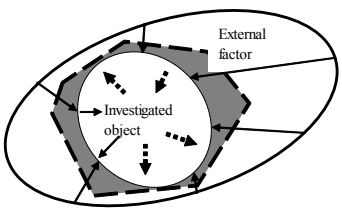

Figure 1: Illustration of area (grey zone) using traditional statistical methods

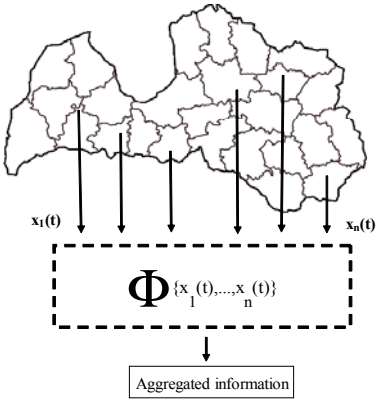

Figure 2: Data aggregation process

Traditional methods of research of industrial and economic objects is the analysis of the received statistical supervision of investigated objects $x(t_1)$, $x(t_2)$, ... ,$x(t_n)$ in time. As a result of processing the received statistical information, as a rule, is generalized information of the object as a whole, where $\Phi\{x_1(t), x_2(t),\ldots, x_n(t)\}$ is the generalized statistical information after processing in the aggregation block (Figure 2). Thus, use of traditional methods of research does not allow receiving necessary information about the behavior of separate parts of the object investigated. The illustration of accidents on Latvian roads, divided by administrative regions of Latvia is presented in Figure 3.

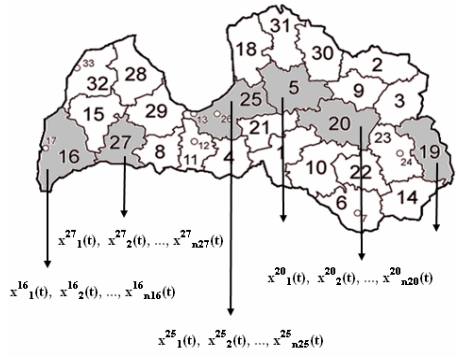

Figure 3: Illustration of accidents on roads of Latvia (divided by administrative regions)

The statistics of accidents on roads of Latvia is presented in Table 1.

Proceedings 22nd European Conference on Modelling and Simulation ©ECMS Loucas S. Louca, Yiorgos Chrysanthou, Zuzana Oplatková, Khalid Al-Begain (Editors)
ISBN: 978-0-9553018-5-8 / ISBN: 978-0-9553018-6-5 (CD)

Table 1: Traffic accident statistics on roads of Latvia

Time period, years	Traffic accidents	Persons injured	Persons killed
1993	3389	3721	670
1994	3814	4380	717
1995	4056	4903	611
1996	3709	4326	540
1997	3925	4674	525
1998	4540	5414	627
1999	4442	5244	604
2000	4482	5449	588
2001	4766	5852	517
2002	5083	6300	518
2003	5368	6634	483
2004	5081	6416	516
2005	4466	5600	442
2006	4302	5404	407

The diagram of traffic accidents on roads of Latvia by type per year is presented in Figure 4.

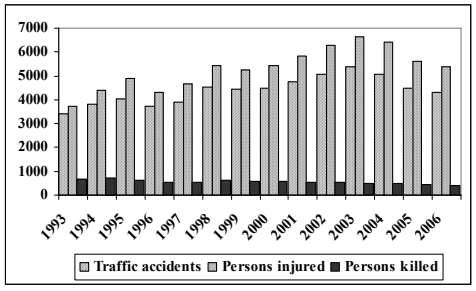

Figure 4: The diagram of traffic accidents on roads of Latvia [7, 8]

A new approach of investigation of a large scale socio-economic problem is shown in [9]. Let us consider what scan statistics is on the basis of examples. Scan statistics is used as an attempt to define clusters of events, using the saved up information distributed in space and time. We assume the scientific definition of variability (development) of an object (a changing object) has been non-uniformly distributed in time and space of the information field.

Thus, an important achievement in research is to reveal the structure of corresponding information fields. Scan statistics allows defining clusters of factors which describe heterogeneity. For better understanding of the idea of scan statistics, let us illustrate an example which is typical for statistical research of events which are distributed in time and space [4].

The distribution of road accident black spots in a small fragment of one of the regions of Latvia is presented in Figure 5.

Figure 5: Illustration of road accident black spots in a small fragment of one of the regions of Latvia [7, 8]

APPLICATION OF SCAN STATISTICS METHOD

Let us have N events, distributed at time interval (0, T). Denote S_w as maximal number of events at a time interval with length w (the window of fixed length w of time). The maximum cluster S_w is called the scan statistics from the viewpoint that one scans the time period (0, T) with a window of size w and observes a large number of points (see Figure 6). W_k is the shortest period of time containing a fixed number of k events. The interval W_{r+1} is called the minimum r^{th} order gap, or r-scan statistics. The distributions of the statistics S_w and W_k are related. If the shortest window that contains k points is longer than w, then there is no window of length w that contains k or more points:

$$P(W_k > w) = P(S_w < k). \quad (1)$$

Let us illustrate the distribution of number of accidents on the roads by an example of some region at the time period from 2002 till 2007. In reviewing the data note that there is a 1 year period (from 1 March 2004 through 1 March 2005) when eight accidents were registered (see Figure 6).

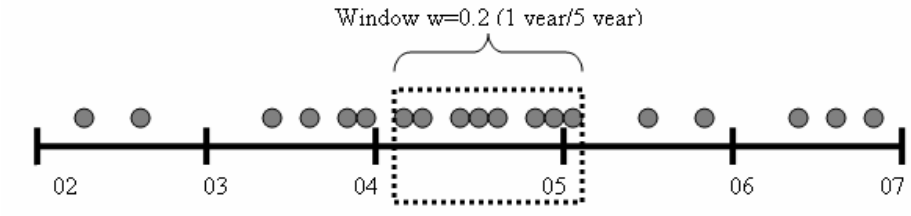

Figure 6: Scanning the unit time interval with the window of length w=0.2. Black points represent times of occurrence of N=19 events, $S_{0.2}$=8. The centers of the occurred "points" C_i have coordinates $t_1, t_2, \ldots, t_N$

In Figure 6 we see concentration of 8 points at the time interval from 1 March 2004 to 1 March 2005. There is a question whether it is possible to explain such concentration proceeding from a null hypothesis. If it is impossible to explain the given concentration of points (accidents) by means of a null hypothesis it is necessary to recognize that the given concentration of points (accidents) has a special character. It means that process of occurrence of failures in the given region is influenced by additional factors. The given conclusion is an objective signal for decision-making in the sphere of traffic management in Latvia. We might explain this as follows: each of the 19 cases could either fall in the period (from 1 March 2004 to 1 March 2005) or not, independently of the other cases. The probability b (k, N, w) found by computing the binomial probability for N = 19, p = 1/5:

$$b(k, N, w) = \binom{N}{k} w^k (1-w)^{N-k} = \quad (2)$$
$$b(k,19,0.2) = \binom{19}{k} 0.2^k (1-0.2)^{19-k}$$

gives the following probabilities for k = 8 and k > 8 (see Formula 2). Probabilities in Table 2 indicate that the cluster under investigation is somewhat unusual but this calculation does not answer the question posed.

Table 2: Probabilities for k = 8 and k > 8

	k = 8	k ≥ 8
p	**0.01662**	**0.02328**

Managers (researchers) want to know how unusual it is to get any 1 year period (not a specific year) containing eight cases. It is possible to solve this problem by dividing a period of five years into five disjoint 1 year periods and use the distribution of the maximum number of cases falling in any 1 year period. However, this does not answer the researchers question since the specific year observed overlaps two disjoint years (see Figure 7). Researchers are not limited by calendar years. The researchers are in effect scanning the 5 year period with a window of length of 1 year focusing on the maximum number of points in the scanning window (see Figure 7).

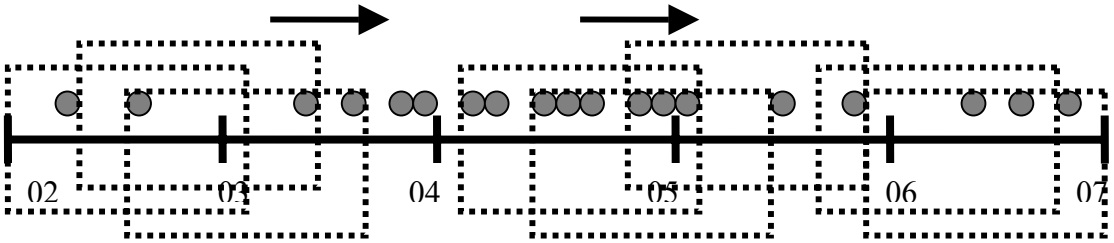

Figure 7: Illustration of the scanning window of fixed length w=0.2

It is easy to understand that there is an infinite number of sliding windows during the time interval (2002- 2007). To solve this problem in a constructive way we must assume some finite set of sliding windows (Figure 8).

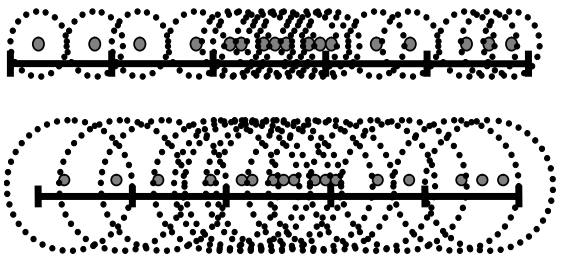

Figure 8: Illustration of the scanning window of two fixed lengths w=0.1 and w=0.2. The centers of scanning windows are points with time coordinates t_1, t_2, …, t_N

Let us denote P(k, 19, 0.2) the distribution of the maximum number of cases in a year under the null hypothesis model for the unrestricted continuum of 1 year periods. The null hypothesis model for 19 cases C_i (i=1, 2,…, 19) were distributed independently by the binomial distribution function) and completely at random over the 5 year period.

The answer to the researchers question is provided by a scan statistics probability. Here the cluster is k=8, the total number of cases over the whole 5 year period is N=19 and the window size w is one year out of five, or w=0.2. The answer is provided by the probability P (8, 19, 0.2). Using formula (2.3) from the [11] probability P (k, N, w) can approximately be calculated as:

$$P(k,N,w) \approx (N-k+1)b(k-1,N,w) - (N-k-1)b(k,N,w) + 2G_b(k+1,N,w), \quad (3)$$

where

$$b(k,N,w) = \binom{N}{k} w^k (1-w)^{N-k}$$

$$G_b(k,N,w) = \sum_{i=k}^{N} b(i,N,w)$$

Probabilities P (k, N, w), calculated with the help of formula (3), are presented in Table 3 (k = 7, 8,…,13). In our example: P(8, 19, 0.2) ≈ 12b(7, 19, 0.2) – 10b(8, 19, 0.2) + 2Gb(9, 19, 0.2) = 12·0.0443 - 10·0.0166 + 2·0.00666 = 0.379. It shows that the cluster observed is not unusual. How accurate is this

value (0.379) for this example? Very precise! The exact tabled result from [10] is 0.376. Exact formulas for the two special cases, P(2, N, w) and P(N, N, w) appear in various probability texts and had been known for many years. An exact formula for P(N, N, w) the cumulative distribution function of W_N the sample range of the N points is derived from Burnside [2].

Table 3: Values of functions b(k,N,w), G_b(k,N,w), P(k,N,w)

k	b(k, N, w)	G_b(k,N,w)	P(k, N, w)
7	0.04432	0.06760	0.8000264
8	0.01662	0.02328	0.3789698
9	0.00508	0.00666	0.1402787
10	0.00127	0.00158	**0.0412472**
11	0.00026	0.00031	**0.0097084**
12	0.00004	0.00005	**0.0018309**
13	0.00001	0.00001	**0.0002752**

For the function P(N,N,w) in case k=N, there is an exact formula:

$$P(N,N,w) = Nw^{N-1} - (N-1)w^N . \qquad (4)$$

In our case k = N = 19, w = 0.2 from (3), we derive the value for P(N, N, w) = P(19, 19, 0.2)-see Table 4.

Table 4: Values of functions b(k,N,w), G_b(k,N,w), P(k,N,w) in case k=18, 19

k	b(k, N, w)	G_b(k,N,w)	P(k, N, w)
18	0.00000	0.00000	0.0000000002870
19	0.00000	0.00000	**0.0000000000040**

An exact formula for P (2, N, w) the cumulative distribution function of W2, the smallest distance between any of the N points, is derived from Parzen [6] by a direct integration approach:

$$\begin{aligned} P(2,N,w) &= 1-[1-(N-1)w]N, \\ &0 < w < 1/(N-1) \\ P(2,N,w) &= 1, \quad for \quad 1/(N-1) < w < 1 . \end{aligned} \qquad (5)$$

Analysing the statistics of accidents on roads of Latvia it is seldom possible to get an exact analytical solution for the distribution of road accidents. In this case there exists only one possibility and it is to use computer modelling. In the paper we have considered to use the Monte-Carlo method of scan statistics for calculation of p-value and testing null hypothesis H_0 (no clusters). The authors are using the approach by Wallenstein and Naus [10] in assuming the null hypothesis model which can be used for investigations of similar problems. Let S[x, x + w) = $S_{x,w}$ denote the number of events in [x, x + w). The scan statistics S_w is defined as [11]:

$$S_w = \sup_{0 \le x \le 1-w} S[x, x+w) . \qquad (6)$$

This is often suggested as statistics (with an appropriate window length w) for testing the presence of clustering. Indeed it arises from the generalized likehood ratio test of uniformity (H_0) against the alternatives (H_1):

$$\begin{aligned} f(x) &= 1/\{1+(\mu-1)\cdot w\}, \quad 0 \le x < T, \\ &= \mu/\{1+(\mu-1)\cdot w\}, \quad T \le x < T+w, \\ &= 1/\{1+(\mu-1)\cdot w\}, \quad T+w \le x < 1 \end{aligned} \qquad (7)$$

,

where f(x) is the density function, $\mu > 1$ and T are unknown but w is known.

For our case the computer programs only look at a constant background rate of events and for the scenarios of grouped data. The authors calculate the scan statistics S_w for continuous data where it has been assumed that N = n events are occurring on a time-line (0, T). The authors generate uniform samples from this time-interval and construct an empirical distribution of the $Pr(S_w > k)$ where k is the maximum number of events in a subinterval of width w (scanning window). There is an infinite number of sliding windows at the time interval (0, T). However, our approach was to consider only the N = n points in a sample derived from this time-interval, and then: (Step 1) sliding windows on the ordered sample data values ($t_1, \dots, t_n$) using the windows $[t_i, t_i + w)$ for i = 1, … , n has been computed. The left square bracket "[" indicates inclusion of the lower point; the open right parenthesis ")" indicates exclusion of the upper point, which is a typical convention in mathematical analysis. In Step 2 the authors looked at the reverse chain of the sliding windows $[t_i , t_i - w)$ for i = n, n-1, … , 1. By generating a large number of samples from the uniform distribution we derived an empirical distribution of $Pr(S_w > k)$. We are interested in finding the value of "k" which shows a small p-value, typically 0.05 or smaller. The modelled p-value can be used for testing null hypothesis the samples of which are uniformly distributed against a clustering alternative. We assume that the given scan statistics (see Formula 6) is a fixed number N of events that have occurred at random in a fixed time period. This conditional (on N) case is called retrospective because in typical applications the researcher scans and makes inferences about N events that have occurred. In practice it is important to investigate the situation when a number of events in the time period are not viewed as a fixed number N that has already occurred, but rather as a random variable with a known probability distribution. We will consider a more popular discrete distribution for Poisson process – Poisson distribution function with an average of λ events per unit time. The typical application is prospective. The researchers seek to use scan statistics to monitor future road accident data or to design a Latvian road management system with the purpose to minimize the number of road accidents.

The Poisson process has been used for modelling real systems dealing with the occurrence of events in time or space. First useful applications of spatial scan

statistics are shown in [5]. Spatial scan statistics is a powerful method for spatial cluster detection. With spatial scan statistics it is possible to search over a given set of spatial regions, find those regions which are most likely to be clusters and correctly adjust for multiple hypothesis testing. Figure 9 illustrates a suspicion cluster – region in S with a high level of intensity q_{in} = 0.02 of accidents [5]. Scan statistics gives answer to the question – is this cluster real or is it a "visual illusion"?

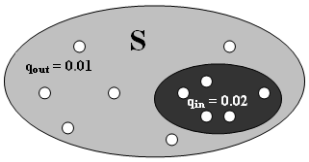

Figure 9: Frequency model of cluster - critical region

The simplest frequency model for this situation (Figure 9) can be written as:
Null hypothesis H_0 (no clusters) $q_{in} = q_{all}$ everywhere (use maximum likelihood estimate of q_{all});
Alternative hypothesis H_1 (cluster in region S), $q_{in} = q_{out}$ elsewhere (use maximum likelihood estimates of q_{in} and q_{out}, subject to $q_{in} > q_{out}$).
In one dimensional case Poisson scan statistics can be calculated as the computation of a scan statistics S_w for grouped data where we assumed that N = n events occur over T disjoint time intervals. Poisson samples (x_1, … , x_T) with the intensity rate λ= N/T were generated. Then we summed the values x_j for all j in each interval [i-1, i-1 + w), for i = 1,…,T and found the maximum of these sums. The intervals here are finite since there are only T intervals if size w = 1, etc. For w = 2 we computed sliding windows of size 2 from (2, 4 , …, T) and then (1, 3, …, T) in steps of 2. Similarly scan statistics for other values of increasing w has been calculated. Generating a large number of samples from the Poisson distribution with rate "λ" an empirical distribution for $Pr(S_w \geq k)$ has been derived. It is important to see which value of "k" showed a small p-value, typically 0.05 or smaller. Modelled p-value can be used for testing null hypothesis. That means that accident events are distributed from the Poisson distribution with rate "λ" against a clustering alternative. A seed (initial number which random number generator uses to start random number generation) is automatically inserted as the current time (in hours, minutes, and seconds).
The intensity of the accident development in one of the regions of Latvia is shown in Figures 10, 11. The relocation of the maximum intensity factor from one sector to another sector of the region was investigated.
Preliminary results of investigating accidents on the roads of Latvia show the following specific features of their distribution:
- lack of traffic saturation in rural areas does not allow to correctly identify clusters of road accidents in these regions;
- significant clusters occur in big cities of Latvia showing possibilities of improving road management system in Latvia.

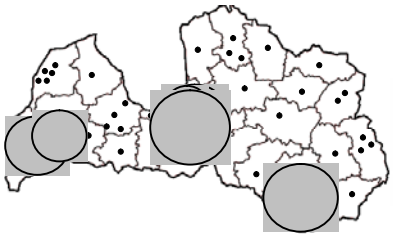

Figure 10: Spatial-time scan statistics with a circle window

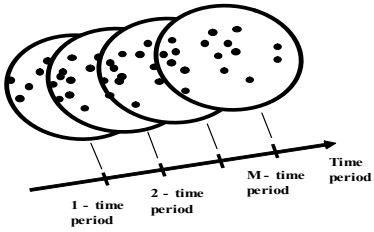

Figure 11: Scan statistics in time

Every year it is necessary to analyse the level of intensity of road accidents applying the method of scan statistics. The paper illustrates the scanning process with a circle window with fixed radius. It is necessary to scan accidents on roads of Latvia with windows of different circles with different radii. The illustration of the Monte-Carlo algorithm for cluster detection is presented in Figure 12.

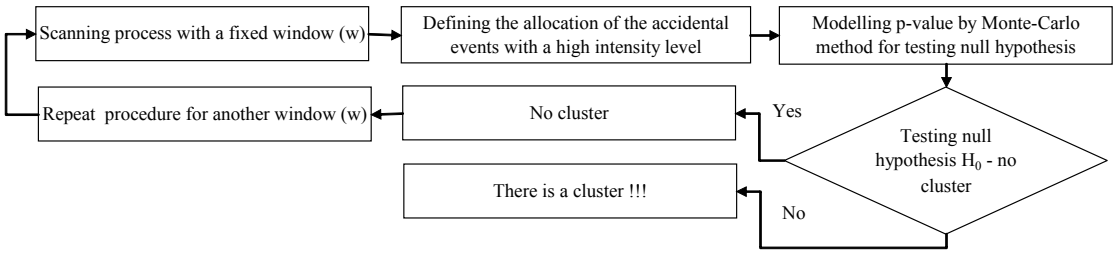

Figure 12: Illustration of the Monte-Carlo algorithm for cluster detection

Alongside with the Monte-Carlo method for calculation of p-value it is possible to apply the direct method of statistics modelling taking into consideration the empirical rules of road accident

distribution on roads of Latvia. In this case both parametric and non-parametric methods of describing road accidents may be used.

CONCLUSION

The opportunity of using scan statistics methodology for research of road accidents on highways of Latvia is presented in this paper. The stochastic modelling of p-value by the method of Monte-Carlo for identifying the black spots of road accidents of Latvia is used.
The application of scan statistics enabled:
- to make analysis of road accidents in some big towns and regions of Latvia;
- to detect clusters with utmost intensity of road accidents applying scan windows of different sizes;
- to check significance of detected clusters with highest frequency of road accidents on the basis of null hypothesis equal to value of 0.05;
- to analyse the dynamics of changes of clusters detected, taking into consideration the time factor (in this case the research was limited by lack of information). Notwithstanding the limitation mentioned above, the possibility of analysing the existing clusters of road accidents in space and time, as well as those emerging in the future, is addressed by the authors of this paper. The findings of the research made about road accidents on Latvian highways highlight the following specific features of their distribution:
- insufficient transport vehicle saturation in rural areas does not allow to identify clusters of road accidents in these regions;
- significant clusters are detected in big cities of Latvia that provide an opportunity to improve the road system management in Latvia.

REFERENCES

1. Berman, M. and. Eagleson, G. K. Dec., 1985. "A Useful Upper Bound for the Tail Probabilities of the Scan Statistics." Journal of the American Statistical Association, Vol. 80, No. 392, doi: 10.2307/2288548, 886-889.

2. Burnside, W. 1928. "Theory of Probability". Cambridge, University Press, 106.

3. D'Antuono, M.F. 2006. ScanoR – "Computations of Scan statistics using R": A language and environment for statistical computing. Department of Agriculture and Food, Western Australia. http://www.agric.wa.gov.au (Email:mdantuono@agric.wa.gov.au).

4. Glaz, J.; Naus, J. and Wallenstein, S. 2001. "Scan Statistics". Springer Series in Statistics, 367 -370. http://books.google.lv/books?id=CHUwtWl6zOYC&pg=PR7&vq=scan+statistics&sig=r-YsGg5HK7TEVh6ZnsWypRQVPqI#PPR13,M1,

5. Kulldorff, M. 1997. "A spatial scan statistics. Communications in Statistics". - Theory and Methods, Volume 26, Issue 6, 1481 – 1496.

6. Parzen, E. Modern Probability "Theory and Its Applications". John Wiley & Sons Inc. December 1960, 464.

7. Road traffic accidents. http://www.csb.gov.lv/csp/content/?lng=en&cat=354

8. Road Traffic Accidents. http://www.sam.gov.lv/satmin/content/?cat=148

9. Tuia, D.; Kaiser, C.; Da Cunha; A., Kanevski, M. 2007. "Socio-economic cluster detection with spatial scan statistics". Case study: services at intra-urban scale. Geocomputation, National University of Ireland, Maynooth, 3-5 September 2007. http://www.clusterville.org/?CaseStudies/ScanStat.

10. Wallenstein S. and Naus J. 2003. "Statistics for temporal surveillance of bioterrorism". In: Syndrome Surveillance: Reports from a National Conference, Morbidity and Mortality Weekly Report 2004; 53, 74-78.

11. Wallenstein, S. 2005 "Scan Statistics". http://c3.biomath.mssm.edu/wscan.html

12. Ward, M.P. and Carpenter, T.E. 2003 "Methods for Determining Temporal Clusters in Surveillance and Survey Programs". In: Animal disease Surveillance and Survey Systems. Methods and Applications. Ed. Salman, M.D. Iowa State Press, Ames Iowa, 87-99.

AUTHORS BIOGRAPHIES

VITALIJS JURENOKS was born in Riga, Latvia. In 1976 he graduated from the Faculty of Engineering Economics of Riga Technical University, and for ten years has worked in an industrial enterprise in Riga. Since 1986 he has been lecturing at the Riga Technical University, and in 1987 was awarded the doctoral degree in economics (Dr.oec.). The main field of research pursued is planning, simulation and optimization of economic processes and systems.
E-mail: vitalijs.jurenoks@rtu.lv.

VLADIMIRS JANSONS was born in Daugavpils, Latvia and is a graduate of the University of Latvia, where he studied mathematics and obtained his degree in 1970. For eight years he has worked in the Computing Centre of the University of Latvia. Since 1978 he has been lecturing at the Riga Technical University, where in 1983 he was awarded the doctoral degree in mathematics. The main field of research pursued is simulation and optimization of economic systems.
E-mail: vladjans@latnet.lv.

KONSTANTINS DIDENKO was born in Jelgava, Latvia. In 1969 he graduated from the Faculty of Engineering Economics of Riga Technical University. Since 1969 he has been lecturing at the Riga Technical University where in 1985 he obtained the doctoral degree in economics. In 2006 he was elected a corresponding member of the Latvian Academy of Sciences. The main field of research pursued is planning and optimization of economic processes and systems. E-mail: ief@rtu.lv