

## SCAN STATISTICS USING FOR ECONOMICAL RESEARCH

Vladimirs Jansons, Vitalijs Jurenoks, Konstantins Didenko  
Riga Technical University - Latvia

21 century is a century in which the basic defining factor of technological development of a society is rapid development of information technologies. Information technologies quickly get literally into all spheres of human activity. The degree of development of information systems in social and economic area allows solving the problem on revealing factors which influence development social-economic systems as a whole. These allow an opportunity to use methods of statistics (scan statistics methods) more widely, using greater files of the information on dynamics of development of investigated object in space and time (scanning in space and time) [1]-[5].

For example, a growing requirement of marketing surveillance activities is the capacity to quickly and accurately determine whether the extra-demand is occurring. The use of centralized data recording and databases generates large quantities of data, with consequent limitations on human resources to search and analyze these data for clustering. Automated processes to evaluate clusters in time (temporal clustering) are needed to enhance the efficiency of surveillance. Detection of temporal clustering by the use of the Scan test offers some advance towards rapid detection of extra-demand, and both spreadsheet methods [2], and internet-based methods [3] have been made available.

**Key words:** Scan statistics methods, Economical research

21 century is a century in which the basic defining factor of technological development of a society is rapid development of information technologies. Information technologies quickly get literally into all spheres of human activity. The degree of development of information systems in social and economic area allows solving the problem on revealing factors which influence development social-economic systems as a whole. These allow an opportunity to use methods of statistics (scan statistics methods) more widely, using greater files of the information on dynamics of development of investigated object in space and time (scanning in space and time) [1]-[5].

For example, a growing requirement of marketing surveillance activities is the capacity to quickly and accurately determine whether the extra-demand is occurring. The use of centralized data recording and databases generates large quantities of data, with consequent limitations on human resources to search and analyze these data for clustering. Automated processes to evaluate clusters in time (temporal clustering) are needed to enhance the efficiency of surveillance. Detection of temporal clustering by the use of the Scan test offers some advance towards rapid detection of extra-demand, and both spreadsheet methods [2], and internet-based methods [3] have been made available.

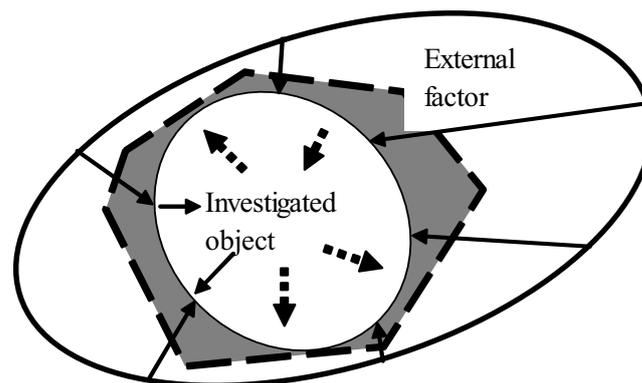


Figure 1. The illustration of area of using of traditional statistics methods

In most cases possibility of traditional statistical conceptions and methods for investigation of real socio-economical object is bounded (grey zone). Traditional statistical methods are more appropriate for investigation of influences of internal factors – for localized (in grey zone) objects investigation. Scan

statistics allows to investigate the socio-economical problems having extremely complex, i.e. sinenergic structure of interrelations (structure of the open systems). The analytical description of such systems probably only in the simplified kind, enabling to consider the likelihood scenarios of development of in-

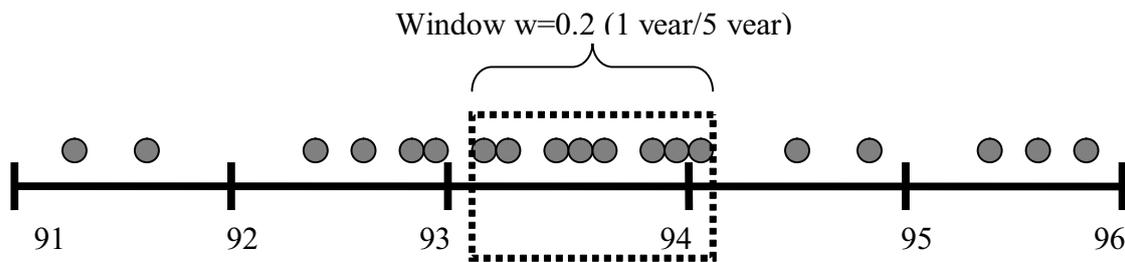
vestigated object, but it is rare when all object as a whole. Real investigation of large scale socio-economical problem is in the project "Urbanisation Regime and Environmental Impact: Analysis and Modelling of Urban Patterns, Clustering and Metamorphoses" [6]. The research will focus on three major Swiss agglomerations: Lausanne, Bern and Zurich. Lausanne will be the main study region. The spatial scan statistics have been used to detect and extract spatiotemporal clusters of service within the city of Lausanne. As a case study, the distribution of restaurant and hotel business has been considered so far. Results show a dependency between high- and low-rates regions: the low-rates regions develop around the high-rate regions, in areas linked by public transports. The use of population adjustments allows understanding the reason of the clustering in a certain area. Results of this transdisciplinary research will increase the understanding of urban phenomena, improve urban analysis and modelling processes. The project will improve data collection and management processes. Let us consider the basic ideas scan statistics on examples. This statistics appears naturally at 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 object (a changing object) has non-uniformly distributed in time and space an information field. Thus, important achievement at research of objects, revealing structure of corresponding infor-

mation fields is. Scan statistics allows defining clusters of factors describing heterogeneity. For better understanding of the idea of the scan statistics, let us illustrate example which is typical for epidemiology medicine [1].

Let us have  $N$  events, distributed in time interval  $(0, T)$ . Denote  $S_w$  maximal number of the events, which is in the time interval with length  $w$  (window of fixed length  $w$  of time). The maximum cluster,  $S_w$ , is called the scan statistic, 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 2).  $W_k$  is the shortest period of time containing a fixed number of  $k$  of events. The interval  $W_{r+1}$  is called the minimum  $r$ -th-order gap, or  $r$ -scan statistic. 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:

$$P(W_k > w) = P(S_w < k) \tag{1}$$

Over a five year period, 1991 to 1995, there were 19 cases of particular type of cancer reported in a city. In reviewing the data, the epidemiologist notes that there is a 1 year period (from April 4, 1993, through April 13, 1994), that contains eight cases. This is illustrated in Figure 2. The researcher asks the following question: "Given 19 cases over 5 years, how unusual is it to have a 1 year period containing as many as eight cases?"



**Figure 2. Scanning the unit time interval with 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  $t_1, t_2, \dots, t_N$ .**

We might try to answer the researcher's question as follows. Each of the 19 cases could either fall in the period (April 14, 1993 to April 13, 1994) or not fall in it, independently of the other cases. The probability  $b(k, N, w)$  of as many as eight cases fall in this interval is found by computing the binomial probability for  $N = 19$ ,  $p = 1/5$ :

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

Formula (2) gives the following probabilities for  $k = 8$  and  $k \geq 8$ :

**Table 1**

	$k = 8$	$k \geq 8$
$p$	<b>0.01662</b>	<b>0.02328</b>

Probabilities in Table 1 indicate that the observed cluster is somewhat unusual. But this calculation does not answer the question posed. The researcher wants to know how unusual it is to get any 1 year period (not a specific year) containing as many cases as eight cases. We might try to get around this problem by dividing the five years up into five disjoint 1 year

periods, and using the distribution of the maximum number of cases falling in any 1 year. However, this too does not answer the researcher's question; since the specific year observed overlaps two disjoint years (Figure 2). The researcher is not limiting the question to calendar years. The researcher is in effect scanning the 5 year period with a window of length 1 year, and focusing on the scan statistic  $S_w$ , the

maximum number of points in the scanning window (Figure 3).

It is easy to understand, that there are an infinite number of sliding windows on a time-interval (91, 96). But for constructive solving of the problem we must to assume some finite set of sliding window (Figure 4):

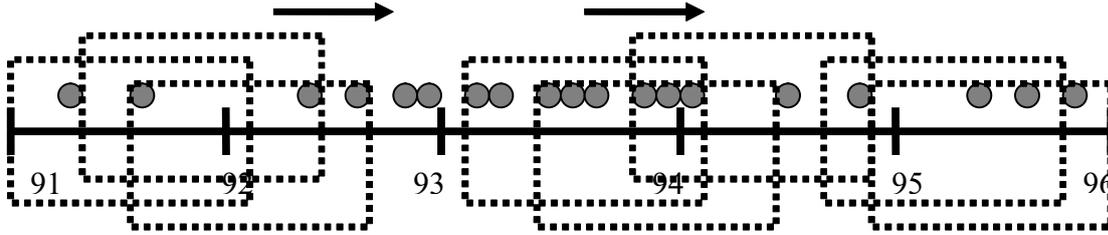


Figure 3. Illustration of the scanning window of fixed length  $w = 0.2$ .

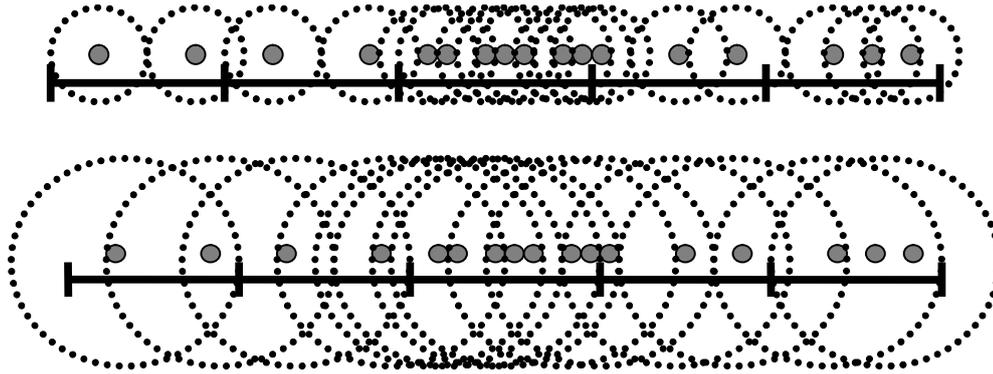


Figure 4. Illustration of the scanning window of two fixed lengths  $w = 0.1$  and  $w = 0.2$ . The centers of scanning windows are points with time coordinates  $t_1, t_2, \dots, t_N$ .

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 model used was that the 19 cases  $C_i$  ( $i = 1, 2, \dots, 19$ ) were distributed independently (with the binomial distribution function) and completely at random over the 5 years period.

The answer to the researcher's question is given 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 given by the probability  $P(8, 19, 0.2)$ . Using formula (2.3) from [1] probability  $P(k, N, w)$  approximately can 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  $b(k, N, w) = \binom{N}{k} w^k (1 - w)^{N-k}$  and  $G_b(k, N, w) = \sum_{i=k}^N b(i, N, w)$ .

$$P(8, 19, 0.2) \approx 12b(7, 19, 0.2) - 10b(8, 19, 0.2) + 2G_b(9, 19, 0.2) = 12 \cdot 0.0443 - 10 \cdot 0.0166 + 2 \cdot 0.00666 = 0.379.$$

Probabilities  $P(k, N, w)$ , calculated with formula (3), are in the Table 2 ( $k = 7, 8, \dots, 13$ ). In our example:

This tells our researcher that observed cluster is not unusual. How accurate is (1) for this example? Very! The exact tabled result from [10] is 0.376.

**Table 2**

k	b(k, N, w)	G <sub>b</sub> (k,N,w)	P(k, N, w)	P(k, N, w)
7	0.04432	0.06760	0.8000264	0.8000
8	0.01662	0.02328	0.3789698	0.3790
9	0.00508	0.00666	0.1402787	0.1403
<b>10</b>	0.00127	0.00158	<b>0.0412472</b>	0.0412
<b>11</b>	0.00026	0.00031	<b>0.0097084</b>	0.0097
<b>12</b>	0.00004	0.00005	<b>0.0018309</b>	0.0018
<b>13</b>	0.00001	0.00001	<b>0.0002752</b>	0.0003

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 in Burnside [7]:

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

In our case for  $k = N = 19, w = 0.2$  we have from (3) value for  $P(N, N, w) = P(19, 19, 0.2)$  in Table 3:

**Table 3**

k	b(k, N, w)	G <sub>b</sub> (k,N,w)	P(k, N, w)
18	0.00000	0.00000	0.0000000002870
<b>19</b>	0.00000	0.00000	<b>0.0000000000040</b>

An exact formula for  $P(2, N, w)$  the cumulative distribution function of  $W_2$ , the smallest distance between any of the  $N$  points, is derived in Parzen [8], by a direct integration approach:

$$P(2, N, w) = 1 - [1 - (N-1)w]^N, \quad 0 \leq w \leq 1/(N-1) \quad (5)$$

$$P(2, N, w) = 1 \quad \text{for } 1/(N-1) \leq w \leq 1.$$

Of course solving real problem, seldom when it is possible to get analytical solution and we have only one possibility - to use computer modeling of the investigated object. In paper we consider the method of the Monte-Carlo in scan statistics. Authors are using the approach by Wallenstein and Naus [2] in assuming the null probability model may arise in a number of situations. Now we more precisely define scan statistics. Let  $S[x, x + w) = S_{x,w}$  denote the number of events in  $[x, x + w)$ . The scan statistics is defined as [11]:

$$S_w = \sup_{0 \leq x \leq 1-w} S[x, x + w) \quad (6)$$

This is often suggested as a statistic (with an

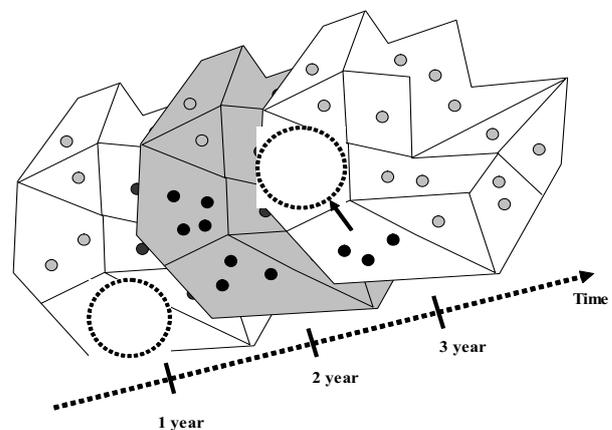
appropriate window length  $w$ ) for testing the presence of clustering. Indeed it arises from the generalized likelihood ratio test of uniformity ( $H_0$ ) against the alternatives ( $H_1$ ):

$$f(x) = 1 / \{1 + (\mu - 1) \cdot w\}, \quad 0 \leq x < T,$$

$$= \mu / \{1 + (\mu - 1) \cdot w\}, \quad T \leq x < T + w, \quad (7)$$

$$= 1 / \{1 + (\mu - 1) \cdot w\}, \quad T + w \leq x < 1$$

where  $f(x)$  density function,  $m > 1$  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 two scenarios of continuous and grouped data. We compute a scan statistic  $S_w$  for continuous data where we assumed that there were  $N = n$  events occurring on a time-line  $(0, T)$ . We 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 sub-interval of width,  $w$  (scanning window). We know that there are an infinite number of sliding windows on a 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) we computed sliding windows on the ordered sample data values  $(t_1, \dots, t_n)$ , using the windows  $[t_i, t_i + w)$  for  $i = 1, \dots, n$ . 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. We also looked (Step 2) at the reverse chain of sliding windows  $[t_j, t_j - w)$  for  $j = n, n-1, \dots, 1$ . By generating a large number of samples from the uniform distribution we derived an empirical distribution of  $\Pr(S_w = k)$ . Our interest was to see which value of “ $k$ ” showed a small p-value, typically 0.05 or smaller. We can use modeled p-value for testing null hypothesis that samples are uniformly distributed against a clustering alternative.



**Figure 5. Spatial-time scan statistic**

In the Figure 5 factor (for example, intensity of the demand) development in some region. After 3 years we see the relocation of the maximum intensity factor from one sector to the other sector of the region.

### Conclusion

The scan statistic methodology at this stage looks at scenarios and probability models for the null distribution in the time series of events, assuming a constant background rate of random events according to the null distribution. It is important to distinguish the appropriate model for the relevant sampling process. Two major concerns with existing methods or approximations of a scan statistic are, knowing what really is correct, and what assumptions are made about statistical distributions used in the approximation. Scan statistic has wide area of applications: biology surveillance, carbon management, marketing, community infrastructure, crop surveillance, disaster management, disease surveillance, ecosystem health, environmental justice, sensor networks, robotic networks, environmental management, environmental policy, homeland security, invasive species, poverty policy, public health and environment, social networks.

### Reference

1. Joseph Glaz, Joseph Naus, Sylvan Wallenstein. Scan Statistics. Springer Series in Statistics. 2001. 367 p.  
<http://books.google.lv/books?id=CHUwtWI6zOYC&pg=PR7&vq=scan+statistics&sig=r-YsGg5HK7TEVh6ZnsWypRQVPql#PPR13,M1>
2. Wallenstein S. and Naus J. (2003) Statistics for temporal surveillance of bioterrorism. In: Syndrome Surveillance: Reports from a National Conference, 2003. Morbidity and Mortality Weekly Report 2004; 53 (Suppl), 74-78.
3. Wallenstein, S. (2005) Scan Statistic. <http://c3.biomath.mssm.edu/wscan.html>
4. D'Antuono, M.F. (2006) ScanoR – Computations of Scan statistics using R<sup>®</sup>: 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).
5. 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. Pp 87-99.
6. 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 2007, National University of Ireland, Maynooth, 3-5 September 2007. <http://www.clusterville.org/?CaseStudies/ScanStat>.
7. W. Burnside, Theory of Probability. Cambridge, University Press 1928. 106 p.
8. Parzen, E. Modern Probability Theory and Its Applications. John Wiley & Sons Inc. December 1960. 464 p.
9. Glaz, J. Approximations for the Multivariate Normal Distribution with Applications in Finance and Economics. In: Applied Stochastic Models and Data Analysis. G. Govaert, J. Janssen and N. Limnios, eds., Volume 1, 37-43, 2001, Universite de Technologie de Compiègne, Compiègne, France.
10. Sylvan Wallenstein, Norman Neff. An approximation for the distribution of the scan statistic. Statistics in Medicine. Volume 6, Issue 2, Published Online: 12 Oct 2006 Copyright © 1987 John Wiley & Sons, Ltd. Pages 197 – 207.
11. M. Berman, G. K. Eagleson. A Useful Upper Bound for the Tail Probabilities of the Scan Statistic When the Sample Size is Large Journal of the American Statistical Association, Vol. 80, No. 392 (Dec., 1985), pp. 886-889 doi:10.2307/2288548
12. Martin Kulldorff. A spatial scan statistic. Communications in Statistics - Theory and Methods, Volume 26, Issue 6 1997, pages 1481 – 1496.