Mikhail ABRAMOVICH,
Mikhail MITSKEVICH

Belarusian State University,
Nezavisimosty pr., 4 Minsk, Republic of Belarus

Statistical methods and algorithms for spatio-temporal cluster analysis

Abstract. The global clusterization test and scan statistic method for studying geographical distribution of the objects are considered. The algorithm of windows set construction for the flexible spatial was developed. The robust version of spatial scan statistic method is proposed. The children carcinoma of the Belarus was analyzed using scan statistic method.

Keywords: Cluster analysis, spatio-temporal, scan statistic, flexible, robust, algorithm, cluster construction, thyroid carcinoma, detection.

1. Introduction

Modern geoinformation systems (GIS) provide a solution of a wide circle of problems that are related with using of spatio-temporal data. Large group of methods, algorithms and technologies from GIS is based on approaches of spatio-temporal data clustering. Clustering is the task of grouping a set of objects in such way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters). Methods of spatio-temporal clustering give a profit in a processing of geoinformation data. Geographical and

temporal properties are a key aspect of many data analysis problems in business, medicine and science.

The methods of spatio-temporal cluster analysis can be classified as local and global [1-3]. The global tests determine a possibility of a presence of a cluster structure in the data under consideration [2]. Global tests are usually performed on the base of a result statistic that provide examination of null hypothesis.

By using local methods of clustering it is possible to find the places and sizes of clusters [1,3]. Local methods of clustering are aimed for searching the data and uncovering the size and location of all possible clusters. In our approach the so-called «spherical windows» of the different sizes are used for construction of spatial clusters. They determine the potential territory of the cluster, so it seems to be impossible to identify the cluster of an arbitrary form by this approach. For these cases the result clusters often include the territories with the low incidence risk. That is why the algorithm of the flexibly shaped cluster construction will be considered.

At presence of outliers local method often determines the cluster that includes only this outlier. In these cases robust versions of statistical procedures should be used to perform the cluster analysis correctly. Therefore, we consider the robust version of the spatial scan statistic.

The methods of spatio-temporal cluster analysis were used for detecting thyroid carcinoma clusters among children and adolescence on Belarus territory in 1989-2005.

2. Global clusterization test

Let us suppose that the whole geographical territory under study is partitioned into m districts, and for every district $i = 1,...,m$ c_i is the number of cases as well as n_i is the risk population. Denote by $C = \sum_{i=1}^{m} c_i$ total number of cases and by $N = \sum_{i=1}^{m} n_i$ total risk population. The location of each district is specified by the pair of the geographic coordinates of its center.

Under the null hypothesis of spatial randomness, the random number of cases in the i th district can be modeled as independent Poisson random variables with an expected number of cases $\mu_i, i = 1,...,m$

$$H_0 : c_i \sim Pois(\mu_i) \ .$$

The expected number of cases i-th district is proportional to the risk population $\mu_i = \lambda n_i$. Under the null hypothesis, the maximum likelihood estimate of λ is given $\hat{\lambda} = \frac{C}{N}$. Assuming the significant expected number of cases ($\mu_i \geq 10$) its distiribution can be approximated with normal distribution with mean and variance μ_i.

And let the local statistic be of form

$$y_i = \frac{c_i - \mu_i}{\sqrt{\mu_i}}.$$

This statistic will have, approximately, a standard normal distribution under the null hypothesis. Then statistic y_i^2 can be considered as a random variable with chi-square distribution with 1 degree of freedom and distribution function $F_{\chi_1^2}(x)$.

Finally, a test for the detection of clustering is based upon maximum y_i^2 observed across all districts:

$$M = \max_i y_i^2. \qquad (1)$$

Under the null hypothesis M statistic has distribution

$$F_M(x) = P\{M \leq x\} = \prod_{i=1}^{m} P\{y_i^2 \leq x\} = F_{\chi_1^2}^m(x).$$

Then the hypothesis test for global clustering is constructed, so choose hypothesis $H_0,\ if\ \ P > \alpha$ and alternative $H_1,\ if\ \ P \leq \alpha$. There P-value is calculated as $P = 1 - F_{\chi_1^2}^m(\cdot)$ and α is a significance level, $0 < \alpha < 1$.

3. Cluster construction using the spatial scan statistic

Among commonly used methods of the local cluster analysis there is the method of the spatial scan statistic [1]. The spatial scan statistic detects spatial clusters using a scanning window that varies its center and radius in order to scan the region under study. When a circular window is used, it may be centered at each district center, and its radius varies from zero to a certain pre-specified maximum possible number of districts K. The method is repeated for each one of the centers.

Let $z_{ik}, k = 1,...,K$ denote a circle which contains *i*-th district and its (*k*-1) neighbours. All circles to be scanned with circular spatial scan statistic are included in set

$Z = \{z_{ik} \mid 1 \le i \le m, 1 \le k \le K\}$.

For each circle *p* is the probability of being a case for population at risk, whereas *q* is the same probability for population at risk outside the circle.

The null hypothesis is $H_0 : p = q.$ The alternative hypothesis is $H_1 : p > q, z \in Z$.

The spatial scan statistics imposes a circular window on the map and lets the centre of the circle move over the area so that at different positions the window includes different sets of neighbours. If the window contains the center of the area, then that whole area is included in the window.

Expected number of cases can be calculated as:

$\mu_i = \frac{C}{N} n_i$.

The maximum likelihood function is:

$$L(z) = \sup_{p>q} \frac{e^{-pn_z - q(N-n_z)}}{C!} p^{c_z} q^{C-c_z} \prod_i \mu_i =$$

$$= \begin{cases} \dfrac{e^{-C}}{C!}\left(\dfrac{c_z}{n_z}\right)^{c_z}\left(\dfrac{C-c_z}{N-n_z}\right)^{C-c_z} \prod_i \mu_i, & if \ \dfrac{c_z}{n_z} > \dfrac{C-c_z}{N-n_z}, \\ \dfrac{e^{-C}}{C!}\left(\dfrac{C}{N}\right)^{C} \prod_i \mu_i, & if \ \dfrac{c_z}{n_z} \le \dfrac{C-c_z}{N-n_z} \end{cases}$$

where

$c_z = \sum_{i \in z} c_i$ is the observed number of cases in the window *z*;

$\mu_z = \sum_{i \in z} \mu_i$ is the expected value number of cases in the window *z*.

Under the null hypothesis the maximum likelihood function is constant:

$$L_0 = \sup_{p=q} L(z) = \sup_p \frac{e^{-pN} p^C}{C!} \prod_i \mu_i = \frac{e^{-C}}{C!}\left(\frac{C}{N}\right)^C \prod_i \mu_i .$$

Spatial scan statistic *S* is the maximum likelihood ratio over all possible circles *z*:

$$S = \frac{\max\limits_{Z}\{L(z)\}}{L_0} = \max\limits_{Z}\left\{\frac{L(z)}{L_0}\right\},$$

The spatial scan statistic for Poisson model is constructed in the form [2]:

$$S = \sup_{z \in Z}\left(\left(\frac{c_z}{\mu_z}\right)^{c_z}\left(\frac{C-c_z}{C-\mu_z}\right)^{C-c_z}\right)^{I\left(\frac{c_z}{\mu_z}>\frac{C-c_z}{C-\mu_z}\right)}, \qquad (2)$$

where $I(.)$ is the indicator function.

The inequality $\frac{c_z}{\mu_z} > \frac{C-c_z}{C-\mu_z}$ in (2) means, that the number of cases inside the window z in comparison to the average is greater than outside the window. The window $z^* \in Z$ that gives the maximal value to the statistic (1) determines the cluster searched for with the highest probability value.

The procedure of the statistical significance testing is organized with the use of the Monte Carlo method. Below the procedure is presented.

1) Calculate test statistics based on given data.
2) For the given integer number generate random datasets $\{c_i \mid c_i \sim Poisson(\mu_i)\}$ under null hypothesis.
3) Calculate test statistics for every simulation.
4) Sort the resulting real and simulated data statistic values and mark the rank of the statistic value which was calculated on the real dataset. The null hypothesis is rejected with specified significance level α, if that rank belong to the top α portion of the whole set.

The algorithm of the spatial scan statistic can be modified to analyze spatio-temporal data. In this case, the time is introduced as the third measurement (coordinate), and the circular windows for calculation of the spatial scan statistic are replaced by cylinders. The base of these cylinders corresponds to some area, as in the spatial case, and the height means the spread of the potential cluster in time.

The formula for the statistic (2) calculation remains the same, but the set of scanning windows is changed. The windows turn to cylinders, as the third (time) coordinate is added.

Then spatio-temporal window $z_{ik}[a,b]$ means the window of the cylinder form, including the district *i* and its (*k*–1) nearest neighbours for each time interval T_P from the set $T_a, T_{a+1}, ..., T_b$.

The set *Z* in the formula (1) is replaced by the set Z_T:

$$Z_T = \{z_{ik}[a,b] \mid 1 \le i \le m, 1 \le k \le K; a,b = T_1, T_2, ..., T_P, a \le b\}.$$

4. Cluster construction using the flexible spatial scan statistic

An important parameter for the use of the spatial scan statistic method is the form of the scanning window. The use of round windows makes not possible to find clusters of the special forms, e.g., long areas along the rivers, parts of the polluted territories. We will consider a method based on a statistic which allows detecting clusters with different flexible shapes [3]. These clusters will be limited with relatively small number of nearest neighbours of each district.

Let $z_{ik(j)}, j = 1, ..., j_{ik}$ denote the *j*-th window which is a set of *k* connected districts starting from district *i*, where j_{ik} is the number of *j* values meeting condition $z_{ik(j)} \subseteq z_{ik}, k = 1, ..., K$. As the result all windows to be scanned are included set

$$U = \{z_{ik(j)} \mid 1 \le i \le m, 1 \le k \le K, 1 \le j \le j_{ik}\}.$$

For each district *i* the circular spatial scan statistic scans *K* concentric circles while the flexible scan statistic scans *K* concentric circles plus all sets of connected districts (including district *i*) with centers lying within the *K*-th largest concentric circle.

While searching the clusters of arbitrary form, it is important to select an algorithm to construct a set of districts of random shape used in the scan statistic calculation. It is necessary to understand, that observable territory can be partitioned into thousands of districts. Therefore, such algorithm must be sufficiently fast and effective. Note, that some restrictions on cluster form must exist to prevent the detection of clusters of unrealistic and unlikely shape.

Algorithm of windows set construction for the flexible spatial scan statistic consists of the following steps.

Step 1. Define $m \times m$ adjacency matrix $A = (a_{ij})$; $a_{ij} = 1$, if districts *i* and *j* are connected, else $a_{ij} = 0$.

Step 2. Define collection of sets $V = \{\varnothing\}$, $i_0 = 0$.

Step 3. Define $i_0 = i_0 + 1$, and select i_0 ($i_0 = 1,..., m$) as a starting district. Define set W_{i_0} consisting of indices of district i_0 and its (K–1) nearest neighbours: $W_{i_0} = \{i_0, i_1, i_2, ..., i_{K-1}\}$.

Step 4. Define current set $X = \{i_0\}$, that specified one window of flexible form.

Step 5. Add set X to V if $X \notin V$.

Step 6. Add every district $j, j \in W_{i_0}, j \notin X$ that has connected district in set X to this set X and recursively repeat steps 5 and 6 if there is at least one such district j.

Step 7. Repeat steps 3-6 until getting desired set V. We can build a set of windows U by collection V of indices of districts for each window.

Computational complexity of the described algorithms is $O\{m(K-1)!\}$

5. Robust version of the spatial scan statistic construction

If the probability model describes observations with outliers, we consider the so-called robust statistical methods of the spatial scan statistic. Let $\overline{c_{z_{ik}}} = \frac{c_{z_{ik}}}{k}$ be the sample mean of cases in window z_{ik} and let $\overline{C} = \frac{C}{m}$ be the sample mean of all cases. Then expression (2) may be written in the form:

$$S = \sup_{z_{ik} \in Z} \left(\left(\frac{\overline{c_{z_{ik}}} k}{\mu_{z_{ik}}} \right)^{\overline{c_{z_{ik}}} k} \left(\frac{m\overline{C} - \overline{c_{z_{ik}}} k}{m\overline{C} - \mu_z} \right)^{m\overline{C} - \overline{c_{z_{ik}}} k} \right)^{I\left(\frac{\overline{c_{z_{ik}}} k}{\mu_z} > \frac{m\overline{C} - \overline{c_{z_{ik}}} k}{m\overline{C} - \mu_z} \right)} \qquad (3)$$

If data has at least one outlier, then statistic (3) often determine the cluster that include only an outlier. At the presence of outliers we need to determine the lower bound of number of observation in the cluster for reducing outlier influence.

Due to outliers $\overline{c}_z$ is a biased estimator for the location parameter. If we a sufficient number of observation in a cluster, the robust estimator of the mean can be used in (3) instead of $\overline{c}_z$.

The robust estimators of the mean proposed by Hampel, Andrew's, Huber and Winsor's mean [4] were used.

Sensitivity of a spatial scan statistic to outliers in the cluster was analyzed by using robust estimators of the mean. Robust estimators decrease the influence of large outlier values to the spatial scan statistic[5].

6. Detecting thyroid carcinoma clusters

The methods of spatio-temporal cluster analysis were used for study clusterization of thyroid carcinoma among children and adolescence of Belarus in 1989-2005. Population at risk and number of cases was available for every year and region of the country. 98.699 km (is equivalent to 1° of eastern longitude for Belarus) and 111.272 km (is equivalent to 1° of northern latitude for Belarus) constants were used for calculating distances in kilometers. The whole territory of the country is divided into 119 regions, the population and the incidence data for every region were used in the analysis.

Test based on statistic (1) was used for investigation of global clusterization. A global clusterization was detected at the level of significance α=0.05 on the Belarus territory in 1991-2002 and on the territory of Brest Region in 1993-2001.

The thyroid carcinoma diagnostical dataset was analyzed using the method of the spatial scan statistic with maximum cluster size set to K=20. Poisson model of disease was considered. The data for each year were analyzed separately and the dependence on data for other years was omitted. The significance of the calculated statistic values (2) was evaluated using 999 Monte Carlo simulation under null hypothesis.

Spatial-temporal cluster analysis was carried out for 3 time periods: 1990-1995, 1996-2001, 2002-2005. For the first time period a statistically significant cluster was detected in 1991-1995, for the second time period in 1996-2000, and for the third time period in 2002-2003.

The results are presented in table 1. Each cluster is represented by its center, number of districts in the cluster, p-value and number of cases in the cluster.

Table 1. Clusters detection by spatial scan statistic using Poisson model

Clusters	Center	Number of districts in cluster	*P*-value	Number of cases in cluster
First time period in 1990-1995				
Cluster of 1991-1995 years	Brahin	15	0.000	191
Second time period in 1996-2001				
Cluster of 1996-2000 years	Lojeŭ	9	0.000	145
Third time period in 2002-2005				
Cluster of 2002-2003 years	Pinsk	10	0.000	32

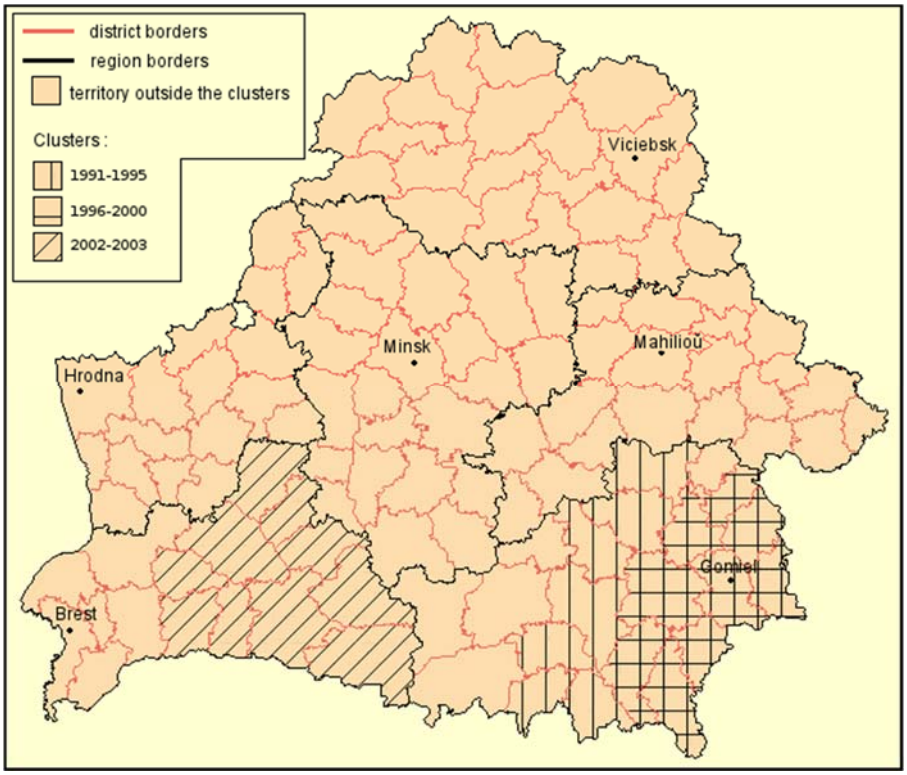

Figure 1. Thyroid carcinoma clusters detected using the method of spatial-temporal scan statistic

Clusters detected using the spatial-temporal scan statistic are presented in figure 1. Spatial-temporal cluster analysis confirmed that there was a significant increase in the number of thyroid carcinoma cases among chidren throughout the territory of Gomel region in the 1990s and Brest region in the 2000s.

References

1. Kulldorff M., Song C., Gregorio D., Samociuk H., DeChello L.: Cancer map patterns: are they random or not? American Journal of Preventive Medicine, № 30(2), pp. 37-49, 2007.

2. Rogerson P.: A set of associated statistical tests for detection of spatial clustering. Ecological and Environmental Statistics. Vol. 12, pp. 275–288, 2005.

3. Tango T., K. Takahashi K.: A flexibly shaped scan statistic for detecting clusters. International journal of Health Geographics. Vol. 4, pp. 115– 125, 2005.

4. Huber P.J.: Robust Statistics. New York: Wiley, 1981.

5. Abramovich M.S., Mitskevich M.M.: Robust spatio-temporal cluster analysis of disease. Proceedings of the 10th International Conference “Computer Data Analysis and Modeling”. Vol. 2. Minsk: “Publishing center of BSU", pp. 95-98, 2013