ROBUST SPATIO-TEMPORAL CLUSTER ANALYSIS OF DISEASE

M.S. Abramovich, M.M. Mitskevich
Research Institute for Applied Problems of Mathematics and Informatics
Minsk, BELARUS
e-mail: Abramovichms@bsu.by

Abstract

The robust version of the spatial scan statistic method for studying geographical distribution of disease is proposed. The thyroid carcinoma diagnostical dataset was analyzed using the method of the robust spatial scan statistic.

1 Introduction

The spatio-temporal cluster analysis can provide an important information concerning the geographic variability of the incidence cases. The algorithms of spatio-temporal cluster analysis can be classified as local and global [1]. The global tests give a possibility to determine whether a cluster structure is presented in total on the territory under consideration. The methods of the local clusterization enable to identify the locations and the sizes of one or several clusters on the territory being analyzed and to test their statistical significance. The most commonly used and effective method for identifying the cluster structure of disease is the method of the spatial scan statistic [2].

The results of analysis for thyroid carcinoma spreading within children and adolescents population at the age up to 14 years old of the Republic of Belarus demonstrate that the data on the thyroid carcinoma incidence contain outliers. In these cases to perform the statistical data analysis (including the cluster analysis) adequately and correctly, it can be recommended to use robust versions of statistical procedures. If the probability model is distorted by outliers in observations, we consider the so-called robust statistical methods [3] for the spatio-temporal cluster analysis.

2 Cluster construction using the spatial scan statistic

Let us suppose that the whole territory under study is divided into m regions, and for every region $c_i,\ i = 1, \ldots, m$, is the number of cases, $n_i, i = 1, \ldots, m$, is the population at risk. Denote $C = \sum_{i=1}^{m} c_i$ as the total number of cases, $N = \sum_{i=1}^{m} n_i$ as the total population at risk.

Suppose that under the null hypothesis H_0 of no clustering, the c_i are independent Poisson random variables such that

$$E(c_i) = \mu_i, c_i \sim Pois(\mu_i), i = 1, \ldots, m,$$

where $Pois(\cdot)$ denotes the Poisson distribution, μ_i is the expected number of cases in the region i.

Each region position is set by the pair of geographical coordinates of its centroid.

Consider the procedure of detecting clusters. Let us define a window as a set of regions and construct it in the following way. Impose a circle on every region with region's centroid being a center of the circle. The radius of the circle varies from 0 to the value at which the circle contains the maximum allowed number of regions K. We end up with an infinite number of circular windows each of them being a potential cluster. Assume that a region is included in the window if this region's centroid lies inside the circle. The number of such windows is finite since the number of all regions m is also being finite.

Let $Z_{ik}, k = 1, ..., K$, denotes a window which contains $(k-1)$ nearest neighbors of i-th region. All windows to be scanned with circular spatial scan statistic are included into the set:

$$Z_c = \{Z_{ik} | 1 \leq i \leq m, 1 \leq k \leq K\}.$$

The spatial scan statistic is based on the likelihood ratio and takes the following form [2]:

$$S = \sup_{Z \in Z_c} \left(\frac{c_Z}{\mu_Z}\right)^{c_Z} \left(\frac{C - c_Z}{C - \mu_Z}\right)^{C - c_Z} I\left(\frac{c_Z}{\mu_Z} > \frac{C - c_Z}{C - \mu_Z}\right), \quad (1)$$

where $c_Z = \sum_{i \in Z} c_i$ is the observed number of incidence cases in the window Z, $\mu_Z = \sum_{i \in Z} \mu_i$ is the expected number of incidence cases in the window Z, $I(\cdot)$ is the indicator function.

The inequality $\frac{c_Z}{\mu_Z} > \frac{N - c_Z}{N - \mu_Z}$ in (1) means that the number of cases inside the window in comparison to the average is greater than outside the window. The window $Z^* \in Z_c$ that gives the maximal value to the statistic (1) is 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.

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 used 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 set Z_c in the formula (1) is replaced by the set

$$Z_{st} = \left\{ Z_{ik[a,b]} |\ 1 \leq i \leq m,\ 1 \leq k \leq K;\ a,\ b = \overline{t_1,\ t_p},\ a \leq b \right\}, \quad (2)$$

where $Z_{ik.[a,b]}$ means the window of the cylinder form including the region i and its $(k-1)$ nearest neighbors for each time interval t_p from the set $\{t_a,\ t_{a+1},\ \ldots,\ t_b\}$.

Here $t_1, \ldots, t_p$ mean the ordered set of the subsequent time intervals, the diagnostic data for them are known. The maximal number of time intervals included in the cluster may be bounded by some value t, $1 \leq t \leq p$ (in analogy with the maximal spatial cluster size K.

3 Robust version of the spatial scan statistic construction

If the probability model describes observations with outliers, we consider the so-called robust statistical methods [3] of the spatial cluster analysis.

Let $\bar{c}_Z$ be the sample mean of cases in window Z, and $|Z|$ be the number of cases in window Z. Analogously, let $\bar{C}$ be the sample mean of all cases, and $|C|$ be the total number of cases.

As $c_Z = \bar{c}_Z |Z|$ and $C = \bar{C} |C|$, expression (1) may be written in the form:

$$S = \sup_{Z \in Z_c} \left(\frac{\bar{c}_Z |Z|}{\mu_Z} \right)^{\bar{c}_Z |Z|} \left(\frac{\bar{C} |C| - \bar{c}_Z |Z|}{\bar{C} |C| - \mu_Z} \right)^{\bar{C}|C| - \bar{c}_Z |Z|} I \left(\frac{\bar{c}_Z |Z|}{\mu_Z} > \frac{\bar{C} |C| - \bar{c}_Z |Z|}{\bar{C} |C| - \mu_Z} \right). \quad (3)$$

If data has at least one outlier, then statistic (1) often determine the cluster that include only this outlier. If the goal of our research is to find a cluster spread in space or time, then we need to determine the lower bound of number of observation in the cluster for reducing outlier influence.

As under outliers $\bar{c}_Z$ is a biased estimator for the location parameter, due to a sufficient number of observation in a cluster, the robust estimator of the mean was used in (3) instead of $\bar{c}_Z$. The robust estimators of the mean proposed by Hampel, Andrew's, Huber and Winsor's mean [3] were used.

A spatial scan statistic sensitivity to outliers in the cluster was analyzed by using robust estimators of the mean. This statistic grows extremely with the outlier value increased as presented in figure 1.

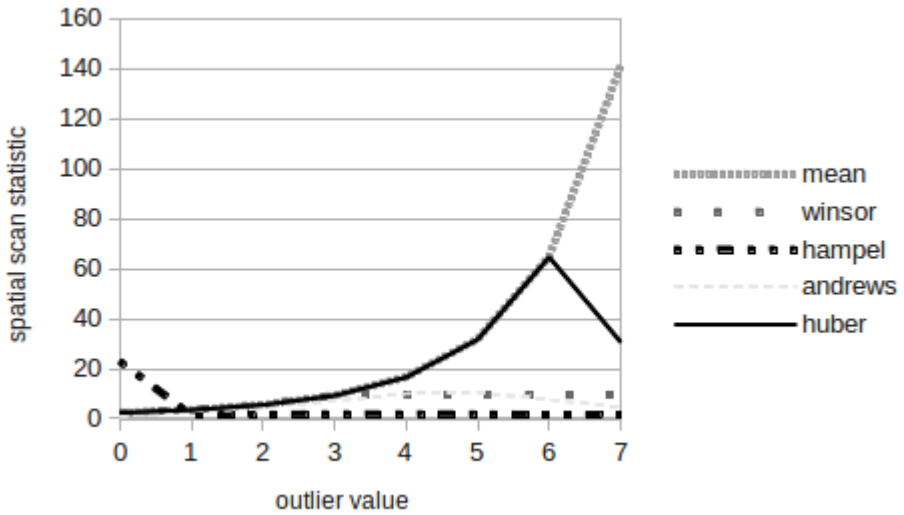

Figure 1: Sensitivity of the spatial scan statistic to outlier

4 Results of robust cluster analysis

A cluster detection study was made for thyroid carcinoma diagnostic data among children up to 14 years old from 1989 to 2009 in the Republic of Belarus. 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.

The thyroid carcinoma diagnostical dataset was analyzed using the method of the robust spatial scan statistic with maximum cluster size set to K =30. The Huber estimator was used for mean estimation. The data for each year were analyzed separately, the dependence on data for other years was omitted. The 6 statistically significant clusters were detected among thyroid carcinoma cases. was evaluated using 999 Monte Carlo simulation runs under the null hypothesis.

The results are given in Table 1. Every cluster is represented by its centroid, the number of contained regions, the number of cases inside the cluster, the expected number of cases under the null hypothesis and the p-value.

Table 1: Results of the robust procedure of spatio-temporal cluster analysis

Year	Centroid	Regions	Cases	Expected	P-value
1991	**Cherikov**	16	20	7.71	**0.001**
1992	**Khotimsk**	28	22	13.66	**0.002**
1994	**Cherikov**	16	23	10.98	**0.003**
1995	**Cherikov**	16	18	11.37	**0.004**
1996	**Cherikov**	25	29	16.09	**0.001**
1997	**Cherikov**	16	18	8.36	**0.001**

Spaio-temporal cluster analysis confirmed that there was a significant increase in the number of thyroid carcinoma cases among children aged from 0 to 14 throughout the territory of Gomel district in the 1990s.

These investigations were supported by ISTC (Project B-1910).

References

[1] Rogerson P. (2005). A set of associated statistical tests for the detection of spatial clustering. *Ecological and Environmental Statistics*. Vol. **12**, pp. 275–288.

[2] Tango T. (2008). A spatial scan statistic with a restricted likelihood ratio. *Japanese Journal of Biometrics*. Vol. **29**, No **2**, pp. 75–95.

[3] Huber P. J. (1981) *Robust Statistics*. Wiley, New York.