SPATIAL CLUSTERING IN RARE DISEASES AND ITS USE FOR CHILDHOOD LEUKEMIA

M.S. Abramovich, S.V. Anishchanka, A.V. Anishchanka, N.N. Savva
Research Institute for Applied Problems of Mathematics and Informatics
Belarusian Research Center for Pediatric Oncology and Hematology
Minsk, BELARUS
e-mail: Abramovichms@bsu.by

Abstract

We consider the spatial circular and flexible scan statistic methods for studying geographical distribution of a rare disease. The methods was applied for childhood leukemia cases of the Republic of Belarus.

1 Introduction

While researching a geographical spread of rare diseases, a critical question arises concerning the randomness of case distribution throughout the territory under examination. The goal is to find out whether this spread is even or there are clusters with a number of cases that is significantly larger than the expected value. In general, all of the methods of cluster analysis can be divided into two groups: local and global [1, 2, 3]. Global clustering tests are evaluating the presence of clustering throughout the territory. Local clustering tests make it possible to detect local clusters, their position and size. It is a common way in local clustering tests to use a circular scan windows of variable size to discover a territory of potential clusters [2]. That is why it becomes difficult to detect clusters with arbitrary shapes such as a territory along a river. Use of a circular scan statistic in this situation may end up detecting clusters that contain surrounding regions without raised disease incidence. In this paper we consider a method for detecting local clusters with flexible shapes. This method was used to detect clustering in childhood leukemia in the Republic of Belarus.

2 Flexible scan statistic for local cluster detection

Consider the situation where an entire study area is divided into m regions. Random variable N_i denotes the number of cases in the region i, n_i is the observed value of N_i, $i = 1, ..., m$. Under the null hypothesis of no clustering, the N_i are independent Poisson random variables such that

$$H_0 : E(N_i) = \xi_i, N_i \sim Pois(\xi_i), i = 1, ..., m, \quad (1)$$

where $Pois(\cdot)$ denotes Poisson distribution, ξ_i are the null expected number of cases in the region i. The geographical position of each region will be specified by the coordinates of its administrative population centroid. The circular scan statistic imposes a circular window z on each centroid [2]. For all of this centroids, the radius of

the circle varies from zero to a pre-set maximum value d or a pre-set maximum number of regions K to be included in the cluster. If such window contains the centroid of some region, then this whole region is included in the window.

Let $z_{ik}, k = 1, ..., K$ denote the window that is composed by the $(k-1)$-nearest neighbours to region i. All the windows to be scanned by the circular spatial scan statistic are included in the set

$$Z_1 = \{z_{ik} | 1 \leq i \leq m, 1 \leq k \leq K\}. \tag{2}$$

The case is slightly different when we use a flexible scan statistic. This statistic imposes an irregularly shaped window z on each region by connecting its adjacent regions. For any given region i a set of irregularly shaped windows is created consisting of k connected regions including i; k changes from 1 to the pre-set maximum K. To avoid detecting a cluster of unlikely peculiar shape, the connected regions are taken only as the subsets of the set of region i and its $(K-1)$-nearest neighbours. In the end we get a very large number of different but overlapping flexibly shaped windows, each of them being a potential cluster [4]. Let $z_{ik(j)}, j = 1, ..., j_{ik}$ denote the j-th window which is a set of k connected regions starting from the region i, where j_{ik} is the number of j satisfying $z_{ik(j)} \subseteq z_{ik}, k = 1, ..., K$. All the windows to be scanned are included in the set

$$Z_2 = \{z_{ik(j)}, | 1 \leq i \leq m, 1 \leq k \leq K, 1 \leq j \leq j_{ik}\}. \tag{3}$$

For any given region i, the circular spatial scan statistic [2] consider K concentric circles, whereas the flexible scan statistic consider K concentric circles and all the sets of connected regions (including the single region i) whose centroids are located within the K-th largest concentric circle.

We are considering the following hypothesis of clustering in the study territory:

$$H_0 : E(N(z)) = \xi(z), \forall z, \tag{4}$$

$$H_1 : \exists z, E(N(z)) > \xi(z), \tag{5}$$

where $N(\cdot)$ and $\xi(\cdot)$ denote the random number of cases and the null expected number of cases within the specified window z, respectively. Under the alternative hypothesis H_1 there are at least one window z where the risk is elevated as compared to the whole territory outside this window. Under the Poisson assumption, the test statistic, which was constructed with the likelihood ratio test, is given by

$$\sup_{z \in Z_l} \left(\frac{n(z)}{\xi(z)} \right)^{n(z)} \left(\frac{n(z^c)}{\xi(z^c)} \right)^{n(z^c)} I\left(\frac{n(z)}{\xi(z)} > \frac{n(z^c)}{\xi(z^c)} \right) \quad , \tag{6}$$

where z^c indicates all the regions outside the window z, and $n(\cdot)$ denotes the observed number of cases within the specified window and $I(\cdot)$ is the indicator function, $l = 1, 2$. The window z^* that attains the maximum likelihood is defined as the most likely cluster. To find the critical value for the test statistic (6) under the null hypothesis, the Monte Carlo hypothesis testing method can be used [2].

The spatial scan statistic can be modified for analyzing space-time data. We are replacing the circular windows with the cylindrical ones with circular bases, which are defined exactly like in the spatial case, and with heights corresponding to a time period. Then, all the windows to be scanned are included in the set

$$Z_3 = \{Z_{ik,[a;b]} | 1 \leq i \leq m, 1 \leq k \leq K; a, b = \overline{T_1, T_P}, a \leq b\}, \quad (7)$$

where $Z_{ik,[a;b]}$ indicates a cylindrical window, which includes the region i and its $(k-1)$-nearest neighbours for each time period T_p from the set $\{T_a, T_{a+1}, ..., T_b\}$. $T_1, T_2, ..., T_P$ indicate the set of available time periods in consecutive order. The maximum number of time periods in one cluster is limited by the pre-set value T, $1 \leq T \leq P$.

3 Clustering of childhood leukemia cases

Methods from section 2 were applied for leukemia cases data of the Republic of Belarus for population under the age of 19 during the period from 1986 to 2005. The whole territory of the country is divided into 118 regions, population and case data for every region was used in analysis. Every administrative centroid was presented by its geographical coordinates. Maximum cluster size was unlimited for circular spatial scan statistic and set to 11 regions for flexible scan statistic. The significance of the calculated statistic values (6) was evaluated using 999 Monte Carlo simulations under the null hypothesis. The significance level was set to $\alpha = 0.05$.

Here are the most significant clusters that were detected (figure 1):

- Year 1993: a significant flexibly shaped cluster with lengthy shape which led the scan statistic to find very large significant clusters that were absorbing the flexibly shaped cluster. The cluster contains Stolbtsy, Nesvizh, Korelichi, Volozhin, Molodechno and Kletsk regions.

- Year 1994: the most significant flexibly shaped cluster with very high statistic value as compared to the other years. The cluster contains Osipovichi, Glusk, Berezino, Pukhovichi, Oktiabr'skii and Starye Dorogi regions.

- Year 1995: not significant enough clusters were found with scan statistic which absorbed a significant flexibly shaped cluster consisted of Zhitkovichi, Liuban, Soligorsk, Glusk, Kopyl' and Lel'chitsy regions.

- Year 1997: a significant cluster that contains only one region of Belynichi.

- Year 2002: a significant flexibly shaped cluster including Lepel', Borisov, Krupki, Dokshitsy and Chashniki regions.

The maximum spatial cluster size for spatial-time scan statistic was set to 60 regions and the maximum size of a time period in a cluster was set to 5 years. The dataset was tested for two periods of time: 1986-2005 and 1990-2005.

In the case of the 1986-2005 period a space-time clustering was detected over 1987-1989 in 51 regions of the southeast Belarus with the p-value of 0.017. In the meantime,

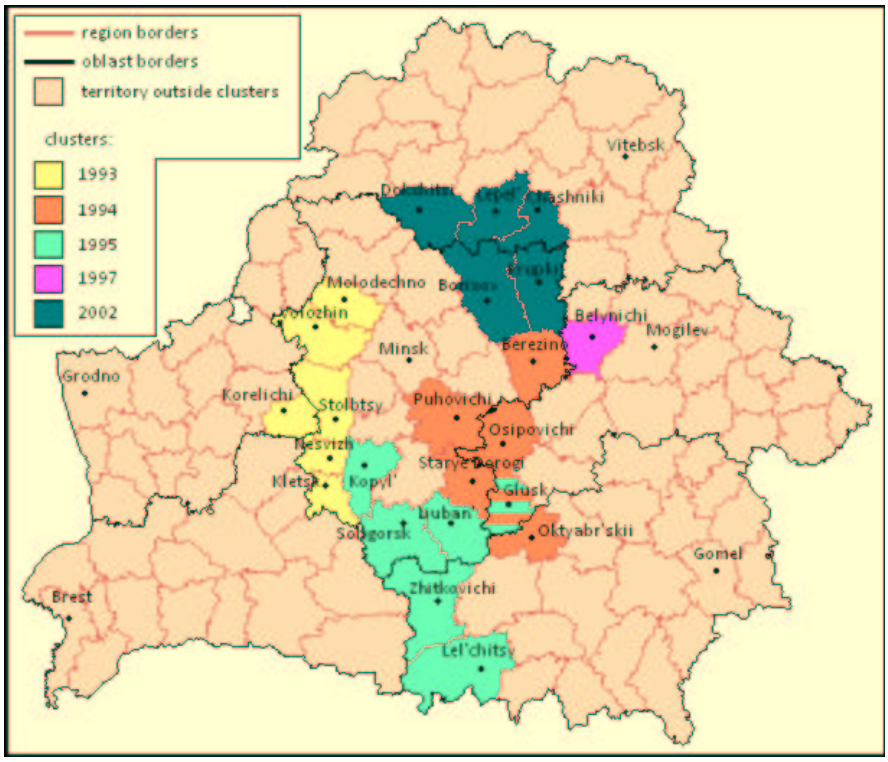

Figure 1: Clusters that were detected with flexible scan statistic

a lot of secondary significant clusters were found over 1987-1991 with large spatial sizes (42 to 60 regions).

In the case of the 1990-2005 period only one significant cluster was detected with the p-value equal to 0.035, which included years 1994-1995 and Starye Dorogi, Liuban', Glusk, Osipovichi, Puhovichi, Slutsk, Oktyabr'skii, Cherven' and Soligorsk regions.

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] Kulldorff M. (1997). A spatial scan statistic. *Common Stat Theory Methods*. Vol. **26**, pp. 81–96.

[3] Tango T. (2000) A test for spatial disease clustering adjusted for multiple testing. *Stat Med*. Vol. **19**, pp. 191–204.

[4] Tango T., Takahashi K (2005) A flexibly shaped scan statistic for detecting clusters. *International journal of Health Geographics*. Vol. **4**, pp. 115–125.