Evaluation of Hotspot Detection Method based on Echelon Structure

Fumio Ishioka[1,3], and Koji Kurihara[2]
[1] School of Law, Okayama University, Okayama, JAPAN
[2] Graduate School of Environmental and Life Science Okayama University, Okayama, JAPAN
[3] Corresponding author: Fumio Ishioka, e-mail: fishioka@law.okayama-u.ac.jp

Abstracts

The study for finding a hotspot, such as disease clustering or hazard map is one of important issue. A spatial scan statistics based on the likelihood ratio associated with the number of events inside and outside a circular scanning window has been widely used as a hotspot detection method. However, it is noted that a non-circular shaped hotspot, such as the shape made by a river or a road cannot be detected. We have proposed a technique using an echelon analysis as a non-circular shaped hotspot detection. The echelon analysis is a useful technique for systematically and objectively investigating the phase-structure of spatial regional data. In this paper, we evaluate the validity of echelon's hotspot detection by comparing with the result of all possible scanning method which certainly detect the hotspot with the highest likelihood ratio.

Keywords: echelon analysis, spatial scan statistic, hotspot

1. Introduction

The detection of problems such as the generation status of infective diseases or hazard maps of natural disasters is a very basic and important issue. GIS provides powerful tools useful for studying the various kinds of spatial data, but it is very difficult to determine the location of a significant spatial cluster: so-called hotspot. For example, a statistical map with shading, such as a choropleth map is used to show how quantitative information varies geographically. However, we can only find contiguous clusters in this map because of the poor accuracy of visual decoding.

To detect the hotspot, several studies corresponding to various types of spatial data have so far been proposed. Among them, spatial scan statistics (Kulldorff (1997)) has been widely used as a hotspot detection method. The spatial scan statistic is based on the likelihood ratio associated with the number of events inside and outside a circular scanning window. However, it is noted that a non-circular shaped cluster, such as the shape made by a river or a road cannot be detected. To solve this problem, several non-circular scanning techniques have been proposed (Patil and Taillie (2004); Duczmal and Assunção (2004); Tango and Takahashi (2005)). However, in these previously reported methods, a detected hotspot sometimes has an unlikely and uncommon shape which requires long computation times in cases where there is a large amount of regional data. In addition to these methods, we have proposed a technique using an echelon analysis as a non-circular shaped hotspot detection.

Echelon analysis (Myers et al. (1997); Kurihara (2004); Kurihara and Ishioka (2007)) is a useful technique for systematically and objectively investigating the phase-structure of spatial regional data. The echelons are derived from changes in topological connectivity. To use the echelon analysis for scanning methods, we can effectively obtain the significant spatial clusters. In this paper, we detect a hotspot by using an echelon scanning method for spatial regional data, and compare it with those detected by a previous study's method. In addition, we propose an all possible scanning method which can absolutely detect a hotspot with the highest likelihood ratio. We demonstrate the further validity of echelon scan by comparison with all possible scan for simulated data.

2. Spatial scan statistic

Spatial scan statistic (Kulldorff (1997)) has been widely used in areas such as epidemiology, criminology, and economics. Now, we discuss the test statistic based on the Poisson model. Assuming that the hotspot-candidate area Z is within the entire area G, the probability of events for the population inside of Z is denoted by p_z, and the one outside Z is denoted by $p_z{}^c$. Then, the null hypothesis is $H_0 : p_z = p_z{}^c$, and the alternative hypothesis is $H_1 : p_z > p_z{}^c$, respectively. When we consider the model based on the Poisson distribution, the maximum likelihood ratio $\lambda(Z)$ is given by

$$\lambda(Z) = \begin{cases} \dfrac{\left(\dfrac{c(Z)}{n(Z)}\right)^{c(Z)} \left(\dfrac{c(Z^c)}{n(Z^c)}\right)^{c(Z^c)}}{\left(\dfrac{c(G)}{n(G)}\right)^{c(G)}} & \text{if } \dfrac{c(Z)}{n(Z)} > \dfrac{c(Z^c)}{n(Z^c)} \\ 1 & \text{otherwise} \end{cases}$$

where $c()$ and $n()$ denote the number of cases and the population size for each region, respectively. To evaluate the significance of the test statistic, Monte Carlo hypothesis testing is commonly used.

Kulldorff (1997) proposed the method that uses a circular window to explore the hotspot candidate Z. However, it is difficult to detect the cluster of non-circular shape cluster such as the shape of a river or a road. To solve this problem, several non-circular scan techniques has been proposed (Patil and Taillie (2004); Duczmal and Assunção (2004); Tango and Takahashi (2005)).

3. Scanning methods

3.1 Echelon scanning method

We have proposed a technique using an echelon analysis as a non-circular shaped hotspot detection. Echelon analysis provides an objective description of between regions by a spatial structure depending on vertical intervals in each region. The echelons are derived from changes in the topological connectivity with a decreasing surface level. For example, Figure 1 (left) shows nine regions (A-I) and their values h. This spatial data is divided into a similar structured area as in Figure 1 (center). These parts are called echelons. The 1st, 2nd, 3rd, and 4th echelons are peaks, the 5th echelon is the foundation of 2nd and 3rd echelons, the 6th echelon is the foundation of the 1st and 5th echelons, and the 7th echelon is the basal foundation. Each region belongs to every echelon. For example, the 1st peak consists of the region {A}, and the 3rd peak consists of the regions {H,E}. Finally, a spatial structure of this regional data is given by the echelon dendrogram shown in Figure 1 (right).

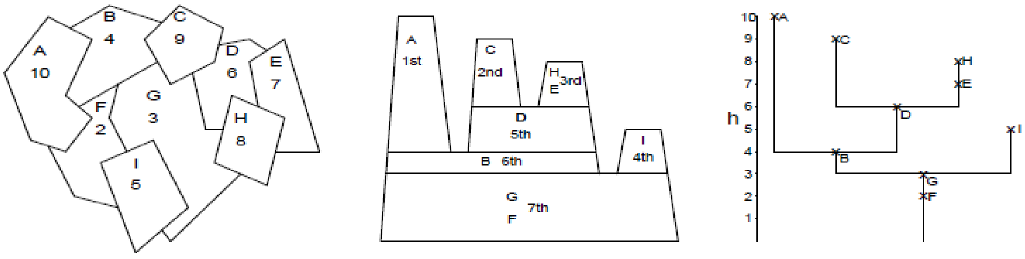


Fig.1 Regional data (left), a division of the same echelon (center), and dendrogram (right)

The procedure of hotspot detection is performed as follows.

1. Scan the region and add it to Z, from the upper to the bottom echelon based on the hierarchical structure of the dendrogram.
2. Detect the cluster region which has the maximum log-likelihood ratio.
3. Estimate the p-value using Monte Carlo simulation under the distribution of the null hypothesis.

The echelon scanning method is preferentially explored from the peak of the structure of the dendrogram, so it can reduce the size of the scanned window Z and computation time. Accordingly, it can apply to a large amount of regional data.

3.2 All possible scanning method

We need a exploring for all regional patterns which is adjacent to each other in spatial regional data to detect a true hotspot with the maximum log-likelihood ratio. However, it usually has a huge amount of combinations and it is utopian to explore the all of them. On the other hand, if we treat spatial data with small areas, we can calculate all combination patterns, thus we require consideration about the trend of the true hotspot. In this section, we propose an algorithm for exploring all regional patterns scanned as the window Z.

Now, we assume a spatial data divided into M regions. Here $\boldsymbol{Z}_m$ (m=1,2, …, M) denotes a group of Z consisting of m regions connected each other. In addition, K_m denotes the total number of Z included in each $\boldsymbol{Z}_m$. Now it is clear that K_1=M because the window Z consisting of one region generates M regions. Next, we find regions $j \in NB(i_k)$ which is adjacent to a region $i_k \in \boldsymbol{Z}_1 (k = 1,2, \ldots, K_1)$. We form a group $\boldsymbol{Z}_2$ and store a set $\{i, j\}$ as Z into $\boldsymbol{Z}_2$. At this time the total set of Z included in $\boldsymbol{Z}_2$ is obtained as $\{(i_k, j) | 1 \le k \le K_1, j \in NB(i_k)\}$. Then, we delete all Z with same pattern shapes in $\boldsymbol{Z}_2$ except one. As a result, we obtain all Z consisting of two connected regions.

Next, in order to obtain the Z consisting of three connected regions, we also find regions $j \in NB(i_k)$ which is adjacent to a region $i_k \in \boldsymbol{Z}_2 (k = 1,2, \ldots, K_1)$, and form a group $\boldsymbol{Z}_3$. In this way, the group $\boldsymbol{Z}_m$ consisting of Z with m regions can obtain using $\boldsymbol{Z}_{m-1}$. Finally, we can obtain all non-overlapping Z which is adjacent to each other by calculating the $\boldsymbol{Z}_m$ until m=M.

4. Evaluation of echelon's hotspot detection method using simulated data

In this section, we discuss about properties of detected hotspot for each scanning methods. We use a small mesh data with 6x4 regions in order to be able to apply the all possible scan. Here, we define a neighbor information for a region $x_{i,j}$ as follows.

$$NB(x_{i,j}) = \{\{a,b\} | a = i,\ j-1 \le b \le j+1\} \cup \{\{a,b\} | i-1 \le a \le i+1,\ b = j\}$$
$$\cap \{\{a,b\} | 1 \le a \le 6,\ 1 \le b \le 4\} - \{(i,j)\}$$

This means a region is adjacent to up to four regions; that is up, down, left and right region. Now we randomly generates a set of Poisson random number under the fixed population (=1,000) in each region, and considered two types of true hotspots, that is {C1, B2, C2} and {A6, B6, C6, D6}, where their value was set to about 3.0. (see Figure 2)

	A	B	C	D
1	2	5	15	9
2	7	21	18	4
3	4	3	6	5
4	6	5	4	2
5	3	1	9	4
6	18	27	21	24

Fig.2 The mesh data with 6x4 regions

	A	B	C	D
1	2	5	15	9
2	7	21	18	4
3	4	3	6	5
4	6	5	4	2
5	3	1	9	4
6	18	27	21	24

Fig.3 The hotspot using circular scan

4.1 Kulldorff's circular scanning method

To begin with, we detected a hotspot using the Kulldorff's circular scanning method. The circular scanned hotspots are obtained using SaTScan software (ver.9.1.1). Here, we preset the radius of circle variations from zero up to 50 % (default) of the total population in the SaTScan. Figure 3 illustrates the regions detected as significant hotspot. This hotspot $Z^*$is consisted of regions {C5, B6, C6, D6}, and this is $\log\lambda(Z^*)$ = 24.90, the relative risk = 2.18. In addition, the p-value based on 999 Monte Carlo replications was 0.001. Here, the total number of scanned regional patterns Z was 288.

4.2 All possible scanning method

Next, we apply the all possible scanning method for this mesh data. Figure 4 shows a flow of this. This figure describes the $Z \in \boldsymbol{Z}_2$ are obtained based on the neighbor information of $Z \in \boldsymbol{Z}_1$, and then, Z with the same shape in $\boldsymbol{Z}_2$ are all deleted except one. For example, regional patterns {A1, B1} and {A1, A2} are generated as Z in $\boldsymbol{Z}_2$ because the region {A1} is adjacent to the {B1} and the {A2}, respectively. Moreover, since the regional patterns $\{\{\mathrm{A1,B1}\},\{\mathrm{B1,A1}\}\} \in \boldsymbol{Z}_2$ are same shape, the {B1, A1} is deleted in in $\boldsymbol{Z}_2$. As a result, the number of all regional patterns $Z \in \boldsymbol{Z}_m (m = 1,2,\ldots,24)$ in this 6x4 mesh data was 1,168,587.

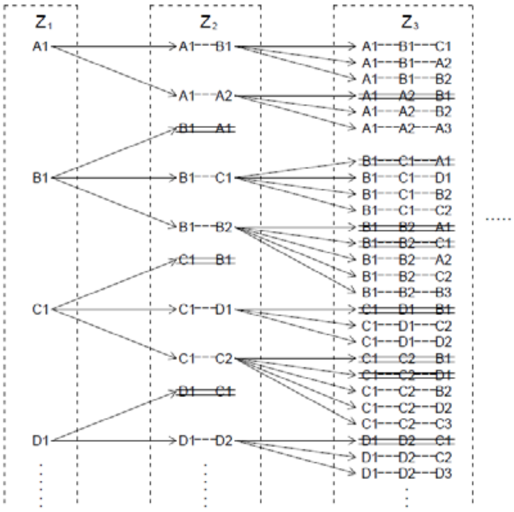


Fig.4 A flow of all possible scan for 6x4 mesh data

	A	B	C	D
1	2	5	15	9
2	7	21	18	4
3	4	3	6	5
4	6	5	4	2
5	3	1	9	4
6	18	27	21	24

Fig.5 The hotspot using all possible scan

We scanned the regions from zero up to 50 % of the total population to be a same condition of circular scan, and it was 198,806. As a result, we obtained the hotspot as $Z^* = \{\mathrm{C1,D1,B2,C2,C3,C4,C5,A6,B6,C6,D6}\}$ (see Figure 5) with $\log\lambda(Z^*) = 45.55$, the relative risk = 2.18 and *p*=0.001. This hotspot Z^* has the highest likelihood ratio compared with other regional patterns.

4.3 Echelon scanning method

Finally, we apply echelon scanning method to this mesh data. Figure 6 shows the echelon dendrograrm based on the above neighbor information and the relative risk for each region. This dendrogram has two big peaks consisting in {B6, D6, C6, A6, C5} and {B2, C2, C1, D1, A2, C3, B1, D3}, respectively. Table 1 shows the process of echelon scanning for this mesh data under the condition of scanning up to 50% of the total population.

Table.1 The process of echelon scanning for 6x4 mesh data

Z	Case	Expected	Relative risk	$\log\lambda(Z)$	*p*
B6	27	9.29	2.91	11.85	0.001
D6	24	9.29	2.58	8.58	0.005
B6, D6, C6	72	27.88	2.58	29.61	0.001
B6, D6, C6, A6	90	37.17	2.42	35.11	0.001
B6, D6, C6, A6, C5	99	46.46	2.13	31.09	0.001
B2	21	9.29	2.26	53.74	0.042
B2, C2	39	18.58	2.10	9.55	0.001
B2, C2, C1	54	27.88	1.94	11.42	0.001
B2, C2, C1, D1	63	37.17	1.70	9.30	0.001
B2, C2, C1, D1, A2	70	46.46	1.51	6.80	0.026
B2, C2, C1, D1, A2, C3	76	55.75	1.36	4.58	0.112
B2, C2, C1, D1, A2, C3, B1, D3	86	74.33	1.16	1.34	0.928
A4	6	9.29	0.65	0	>0.999
A4, B4	11	18.58	0.59	0	>0.999

As a result, $Z^* = \{\mathrm{B6, D6, C6, A6}\}$ was detected as a regional pattern with maximum log likelihood ratio (see Figure 7), and this is $\log\lambda(Z^*) = 35.11$, the relative risk = 2.42 and *p*=0.001. In addition, $Z^* = \{\mathrm{B2, C2, C1}\}$ was detected as regions with the second highest log likelihood ratio.

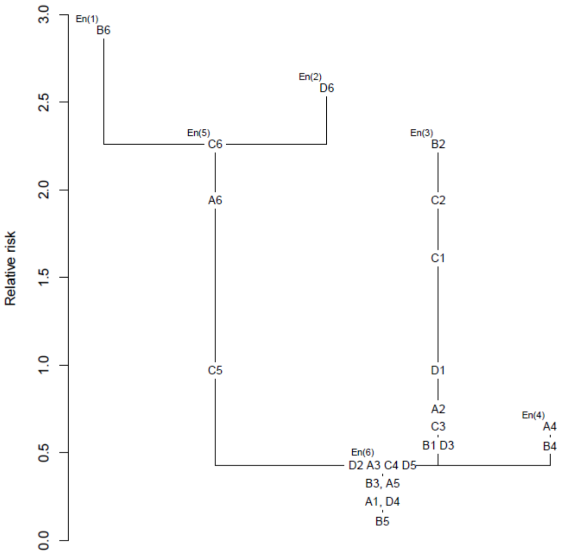
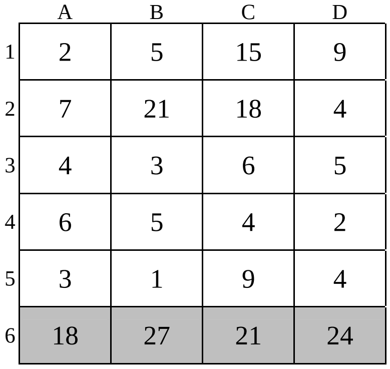


Fig.6 The echelon dendorgram for 6x4 mesh data



	A	B	C	D
1	2	5	15	9
2	7	21	18	4
3	4	3	6	5
4	6	5	4	2
5	3	1	9	4
6	18	27	21	24

Fig.7 The hotspot using echelon scan

5. Consideration

Table 2 shows the results for each scanning method. All of the methods were able to detect the significant hotspot. Echelon's hotspot was obtained the higher likelihood ratio and relative risk than circular's one although the total number of the scanned *Z* was only 14. It is because the regions consisting of peak in the dendorgram based on the relative risk are preferentially scanned in the echelon scan. In addition, we observe that it is a line shape's hotspot, which cannot be detected by circular shaped scanning. The echelon scanning method enables the detection of hotspot having various shapes and significant likelihood ratios, because its searches are based on the essential spatial structure of data.

On the other hand, the all possible scan was obtained the hotspot with the highest likelihood ratio. Nevertheless, it had a very low relative risk. This may be caused by the feature where the likelihood-ratio-based spatial scan statistic tends to detect a hotspot which is much larger than the true cluster, encompassing neighboring regions with nonelevated risks. In contrast, the echelon-based scan statistic has a small chance of exploring the lower value regions that are located about the bottom echelon, such as regions {C3, C4, C5} in this mesh data.

Table.2 The results for each scanning method for 6x4 mesh data

Hotspot regions	Cases	Expected	Relative risk	$\log\lambda(Z)$	*p*	no. scanned *Z*
Circular						
C5, B6, C6, D6	81	37.17	2.18	24.90	0.001	288
All possible						
C1,D1,B2,C2,C3,C4, C5,A6,B6,C6,D6	172	102.21	1.68	45.55	0.001	198,806
Echelon						
A6, B6, C6, D6	90	37.17	2.42	35.11	0.001	14

A further effectiveness of echelon scan can be verified using a plot of the log-likelihood ratio vs. relative risk shown in Figure 8. Each point means a regional patterns *Z* calculated by all possible scan; there are 198,806. Among them, 14 regional patterns scanned by echelon also plot using a square. As the plot indicates, the echelon scan could scan the *Z* having a high log-likelihood ratio in the higher range of the relative risk. One significant advantage of the echelon scanning method is that it can detect significant spatial hotspots with a high relative risk in spite of considerably diminished number of scanned windows *Z*. Accordingly, the echelon scanning method can be applied to extensive spatial regional data with thousands of regions.

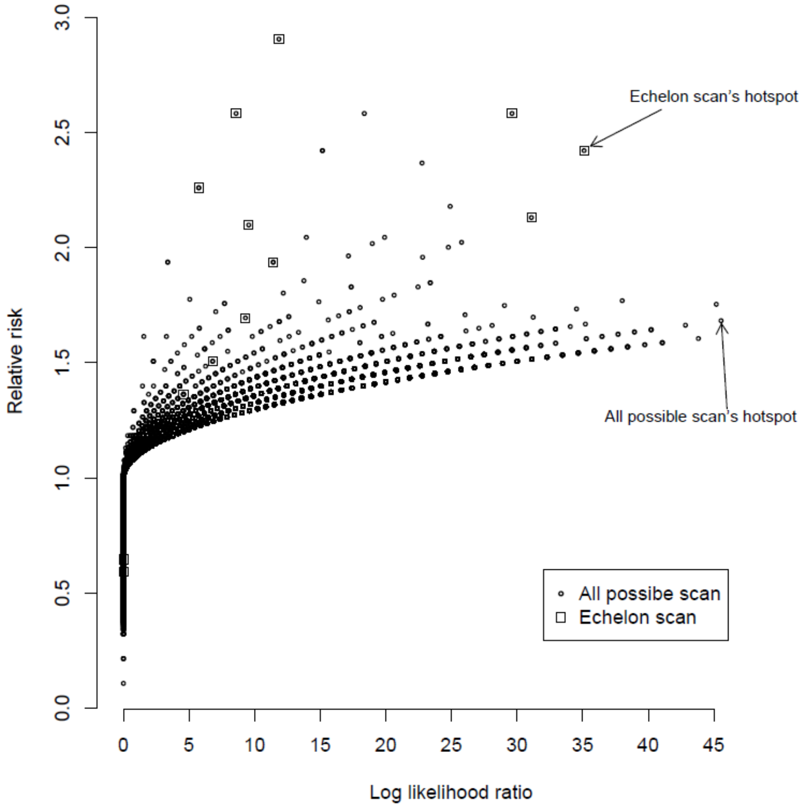


Fig.8 A plot of the log-likelihood ratio vs. the relative risk

References

Duczmal, L and Assunção, R.A. (2004) "A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters." *Computational Statistics and Data Analysis*, 45, 269–286.

Ishioka F. and Kurihara K. (2012) "Hotspot detection using scan method based on echelon analysis." *Proceedings of the Institute of Statistical Mathematics*, 60(1), 93–108 (in Japanese).

Kulldorff, M. (1997) "A spatial scan statistics." *Communications in Statistics, Theory andMethods*, 26, 1481–1496.

Kulldorff, M. and Information Management Services, Inc. (2011) SaTScan v9.1.1: Software for the spatial and space-time scan statistics. http://www.satscan.org/

Kurihara, K. (2004) "Classication of geospatial lattice data and their graphical representa-tion." *Classification, Clustering and Data Mining Applications (Edited by Banks, D. et el.)*, Springer, Berlin, Tokyo, 251–258.

Kurihara, K. and Ishioka, F. (2007) "Classication of spatial data based on the pattern of hierarchical structure and its applications." *Journal of the Japan Statistical Society*, 37(1), Series J, 113–132 (in Japanese).

Myers, W.L., Patil, G.P. and Joly, K. (1997) "Echelon approach to areas of concern in synoptic regional monitoring." *Environmental and Ecological Statistics*, 4, 131–152.

Patil, G.P. and Taillie, C. (2004) "Upper level set scan statistic for detecting arbitrarily shaped hotspots." *Environmental and Ecological Statistics*, 11, 183–197.

Tango, T. and Takahashi, K. (2005) "A exible spatial scan statistic for detecting clusters." *International Journal of Health Geographics*, 4, 11.