Statistical Approach for Environmental Problems based on Spatial Structure

Koji Kurihara[†] and Fumio Ishioka[‡]

[†]*The Graduate School of Environmental Science, Okayama University, 3-1-1 Tsushima-naka Okayama 700-8530, Japan, kurihara@ems.okayama-u.ac.jp*

[‡]*The School of Law, Okayama University, 3-1-1 Tsushima-naka Okayama 700-8530, Japan, fishioka@law.okayama-u.ac.jp*

1. Introduction

The increased level of interest is the concern in society for environmental issues and their relation to the health of individuals. It is important to investigate the areas of significant risk (hotspot) about the effect on the human health status to make early warning for infectious diseases and so on. Most environmental phenomena investigated by sampling geographic space have spatial components. The important roll of statistical analysis for spatial data is to build a model and to make clear the structure of data based on spatial information. There are some typical problems of spatial analysis for geostatistical data, lattice data and point patterns. We focus on lattice data over a fixed subset D of d-dimensional Euclidean space. We deal with the 0/1 event data over the entirety of a partitioned spatial region. Data can be collected directly within each region. These data are known as a kind of spatial epidemiological data, cellular data, irregular lattice data and so on. Several methods have been proposed to detect the hotspots areas. From the perspective of the spatial autocorrelation, Anselin (1995) proposed a local Moran's *I* statistics which was able to locate spatial associations. Recently, the hotspot detection by scan statistics based on the likelihood ratio is a popular method. Kulldorff (1997) detected the hotspots, significant cluster (zone) for the lattice data, based on spatial scan statistic with Binomial and Poisson models. The circular window zone for scanning is defined around one lattice (county) seat. The zone consists of counties whose county seat exists within the circle. Thus we can only detect the circular cluster based on this circular scan. Echelon analysis (Myers et al. 1997) is useful to investigate the cellular surface analysis by systematically and objectively determining topological structure and change. The echelon dendrogram represents the surface topology of lattice data and hierarchical structure of these data. Regional features such as hotspots and trends are shown in an echelon dendrogram. The candidates of hotspots are given as the top echelon in the dendrogram, and some extended approaches are proposed for health and environmental data (Ishioka et al. (2007), Kurihara (2004), Kurihara et al. (2000, 2006), Myers et al. (2006), Tomita et al. (2008)). Therefore we can detect the hotspots of any size and shape for spatially aggregated lattice data based on proposed technique with spatial scan statistic and echelon analysis. The purpose of this paper is to classify any types of lattice data based on their spatial hierarchical structure and to detect the hotspots with regional features. In section 2, we explain the spatial scan statistics. In section 3, we introduce echelon analysis

for some types of lattice data. In section 4, we demonstrate the proposed technique with some illustrations on environmental and epidemiological data.

2. Spatial Scan Statistics

The spatial scan statistics is a test statistics to detect the areas with significantly high or low rates. There is one area Z, which is a subset of whole area G. Individuals within area Z have population probability p_1 of the attribute, whereas the population probability for individuals outside of the area Z is p_2. The probabilities for all individuals are mutually independent. The null hypothesis is H_0: $p_1=p_2=p$, and the alternative hypothesis is H_1: $p_1>p_2$, then we have a high attribute rate in an area Z. Let $n(G)$ be the total population in whole area G, and $n(Z)$ be the population within area Z. The $c(G)$ is the total number of attributes in all of area G and $c(Z)$ is the number of the attributes within area Z. Then we consider the model based on the Poisson distribution. The probability of in the study area is given by

$$f(Z) = \exp[-p_1 n(Z) - p_2(n(G)-n(Z))]\frac{[p_1 n(Z) + p_2(n(G)-n(Z))]^{c(G)}}{c(G)\ !}. \tag{2.1}$$

The density function $f(x)$ of a specific point being observed at location x is

$$\begin{cases} \dfrac{p_1 n(x)}{p_1 n(Z) + p_2(n(G)-n(Z))} & if\ \ x \in Z \\ \dfrac{p_2 n(x)}{p_1 n(Z) + p_2(n(G)-n(Z))} & if\ \ x \notin Z \end{cases}. \tag{2.2}$$

We can therefore write the likelihood function as

$$L(Z, p_1, p_2) = \exp[-p_1 n(Z) - p_2(n(G)-n(Z))]\frac{[p_1 n(Z) + p_2(n(G)-n(Z))]^{c(G)}}{c(G)\ !}$$

$$\times \prod_{x_i \in Z} \frac{p_1 n(x)}{p_1 n(Z) + p_2(n(G)-n(Z))} \prod_{x_i \notin Z} \frac{p_2 n(x)}{p_1 n(Z) + p_2(n(G)-n(Z))}. \tag{2.3}$$

$$= \frac{\exp[-p_1 n(Z) - p_2(n(G)-n(Z))]}{c(G)\ !} {p_1}^{c(Z)} {p_2}^{c(G)-c(Z)} \prod_{x_i} n(x_i)$$

To maximize the likelihood function (2.3), we calculate the maximum likelihood function conditioned to area Z. The maximum likelihood estimator $\hat{p}_1 = \frac{c(Z)}{n(Z)}$ and $\hat{p}_2 = \frac{(c(G)-c(Z))}{(n(G)-n(Z))}$ are substituted.

$$L(Z)=\frac{\exp[-c(G)]}{c(G)\,!}\left(\frac{c(Z)}{n(Z)}\right)^{c(Z)}\left(\frac{c(G)-c(Z)}{n(G)-n(Z)}\right)^{c(G)-c(Z)}\prod_{x_i} n(x_i)\,. \tag{2.4}$$

The likelihood ratio $\lambda(Z)$ is maximized over all subset areas of whole areas to detect the hotspots.

$$\lambda(Z)=\underset{Z}{Max}\,L(Z)\Big/L_0=\frac{\left(\frac{c(Z)}{n(Z)}\right)^{c(Z)}\left(\frac{c(G)-c(Z)}{n(G)-n(Z)}\right)^{c(G)-c(Z)}}{\left(\frac{c(G)}{n(G)}\right)^{c(G)}}\,. \tag{2.5}$$

Here, L_0 is the following likelihood function under the null hypothesis.

$$L_0=\sup_p\frac{\exp[-pn(G)]}{c(G)\,!}\,p^{c(G)}\prod_{x_i} n(x_i)=\frac{\exp[-c(G)]}{c(G)\,!}\left(\frac{c(G)}{n(G)}\right)^{c(G)}\prod_{x_i} n(x_i)\,. \tag{2.6}$$

The test statistics $\lambda(Z)$ is also written as

$$\lambda(Z)=\left(\frac{c(Z)}{e(Z)}\right)^{c(Z)}\left(\frac{c(G)-c(Z)}{e(G)-e(Z)}\right)^{c(G)-c(Z)}, \tag{2.7}$$

where $e(Z)$ is the expected value of the attribute within area Z, and $e(G)=c(G)$. An area Z, where the value of λ becomes the maximum, is suitable as the hotspot.

3. Echelon Analysis

3.1 Basic idea

The echelon approach aggregates the areas in which the values have identical topological structure and produce a hierarchically related structure of these areas based on connective (neighbor) information among cells. One-dimensional spatial lattice data has the position (i) and the value h_i on the horizontal and vertical lines, respectively. For D_1 divided lattice (interval) data, data are taken at the interval $l_1(i)=(i-1,i],\ i=1,2,\ldots,D_1$. Table 1 shows the 25 intervals named from A to Y in order and their values (e.g., A=1 and Q=7).

Table 1: One-dimensional spatial lattice data.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
ID	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
h(i)	1	2	3	4	3	4	5	4	3	2	3	4	5	6	5	6	7	6	5	4	3	2	1	2	1

At first, we define the neighbor information of spatial lattice data $l_1(i)$, say $NB(i)$. The $NB(i)$ indicates the spatial positions between each cell, and it is given by

$$NB(i)=\begin{cases}\{i+1\}, & i=1\\ \{i-1,i+1\}, & 1<i<D_1\\ \{i-1\}, & i=D_1\end{cases}\cdot \tag{3.1}$$

We can make the cross sectional view of topographical map like Figure 1, based on $NB(i)$ and value of each cell. There are nine numbered parts with same topological structure in these hills. These parts are called echelons. These echelons consist of peaks, foundation of peaks and foundation of foundation. The numbers 1,2,3,4 and 5 are the peaks of hills. The numbers 6 and 7 are the foundations of two peaks. The number 8 is the foundation of two foundations. The number 9 is the foundation of foundation and peak and also called as the root. The graphical representation is given by the following dendrogram shown in Figure 2.

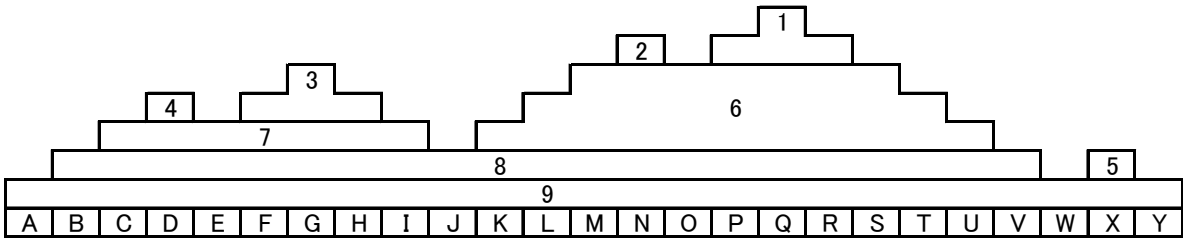

Figure 1: The hypothetical set of hillforms in one-dimensional spatial lattice data.

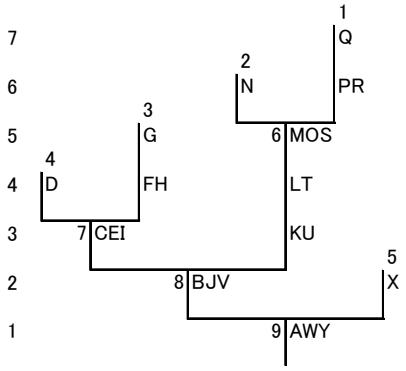

Figure 2: The echelon dendrogram for one-dimensional spatial lattice data.

3.2 Echelon analysis for two and three dimensional spatial lattice data

Two dimensional spatial lattice data, such as remote sensing data or mesh data, are given as the cells of digital value $h_{i,j}$ over the $D_1 \times D_2$ array data:

$$l_2(i,j)=\left\{(x,y)\mid x_{i-1}\le x\le x_i, y_{j-1}\le y\le y_j,\right\}\ i=1,2,\ldots,D_1,\ j=1,2,\ldots,D_2 \tag{3.2}$$

The neighbor information of cell $l_2(i,j)$ is given as

$$NB(l_2(i,j)) = \{(a,b) \mid i-1 \le a \le i+1,\ j-1 \le b \le j+1\} \cap \{(a,b) \mid 1 \le a \le D_1, 1 \le b \le D_2\} - \{(i,j)\} \quad (3.3)$$

where $A - B = A \cap \{B^c\}$ for the sets of *A* and *B*. Here, B^c denotes the complement of *B*. For such 2D data with a digital value over a 5×5 array shown in the left side of Figure 3, the echelon dendrogram shown in the right side of Figure 3 is produced by the following steps.

	A	B	C	D	E
1	2	24	8	15	3
2	10	1	14	22	5
3	4	13	19	23	25
4	20	21	12	11	17
5	16	6	9	18	7

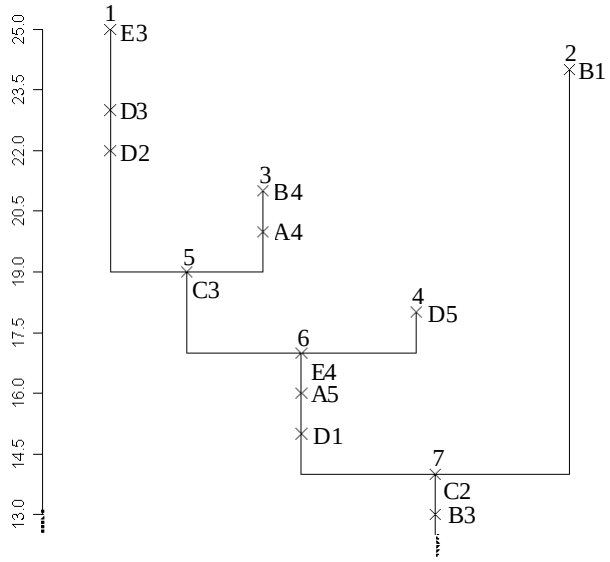

Figure 3: Digital data over 5×5 array and their echelon dendrogram.

Algorithm-E: To find the peaks and foundation of echelon

Step E-1) Find the peaks

The digital values in the peak are greater than the values of neighboring cell of same peak. There are four peaks in this 5-by-5 array. The maximum value in this array is 25. The value of 25 belongs to the fist peak. The maximum value among connected data to 25 is 23. The value of 23 is greater than the values of neighboring cell of 25 and 23. Thus the value of 23 belongs to the fist peak. The maximum value among connected data to 25 and 23 is 22. The value of 22 is greater than the values of neighboring cell of 25, 23 and 22. Thus the value of 22 belongs to the fist peak. The maximum value among connected data to 25, 23 and 22 is 19. But the value of 19 is not greater than 21 which is connected to 19. Thus the value of 19 does not belong to the first peak. As a result, the first peak consists of the values of 25, 23 and 22, and its echelon number is 1. These values are greater than the values of neighboring cell of first peak. In the same manner, second peak consists of 24, and third peak consists of 21 and 20, and the fourth peak consists of 18. These echelon numbers are 2, 3 and 4, respectively.

Step E-2) Find the foundations of the peaks and foundations

The maximum value except the values of four peaks is 19. The value of 19 is the foundation of the peaks whose echelon numbers are 1 and 3. The echelon number of this foundation is 5.

The echelon number 5 is a parent of echelon numbers of 1 and 3. This relationship is expressed as 5(1 3) using echelon numbers. Similarly, we can find the foundation 6 for echelon numbers 4 and 5, and foundation is 7 for echelon numbers of 2 and 6. These relationship is expressed as 7(2 6(5(1 3) 4)) using echelon numbers.

Three dimensional spatial lattice data consist of overlapped two-dimensional (2D) spatial data. These data are also considered as cubic data that consist of $D_1 \times D_2 \times D_3$. Therefore, neighbor information of cell $l_3(i, j, k)$ is given as

$$NB(l_3(i,\ j,k)) = \{(a,b,c)\,|\,i-1 \le a \le i+1,\ \ j-1 \le b \le j+1,\ \ k-1 \le c \le k+1\} \\ \cap\{(a,b,c)\,|\,1 \le a \le D_1,\ \ \ 1 \le b \le D_2,\ 1 \le c \le D_3\} - \{(i,\ j,k)\} \tag{3.4}$$

where i=1,2,…,D_1, j=1,2,…,D_2, and k=1,2,…,D_3.

For 3D data with a digital value over a $4 \times 4 \times 3$ array shown in the left side of Figure 4, the echelon dendrogram shown in the right side of Figure 4 is also produced by the similar steps of Algorithm-E.

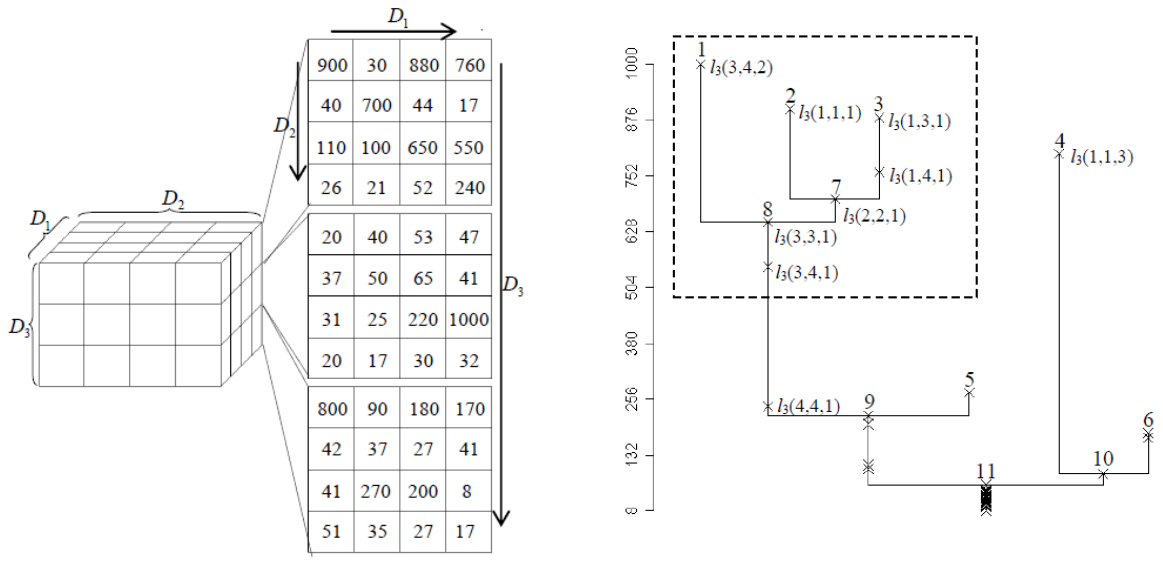

Figure 4: Digital data over $4 \times 4 \times 3$ array and their echelon dendrogram.

We use the echelon technique to seek areas Z where the test statistics is maximized. Through scanning based on echelon, we can detect the candidates of hotspots using the following procedures.

Algorithm-H: To detect the candidate of hotspots

Step H-1) Draw the echelon dendrogram for specified spatial data

Step H-2) Scan the areas from the upper echelon to the bottom, based on the hierarchical structure determined in Step H-1.

Step H-3) Detect the candidate of hotspots, which takes the maximum natural logarithm of test statistics $\lambda(Z)$.

3.3 Echelon analysis for multivariate spatial data using PCA

Echelon analysis has been applied only to univariate spatial data. As a result, it is impossible to detect the hotspots on the multivariate spatial data. We propose a technique of echelon analysis for multivariate spatial data with principal component analysis (PCA) which is the one of the multivariate dimension reduction techniques. We can detect the candidate of hotspots for multivariate spatial data using the following procedures.

Algorithm-M: To detect the candidate of hotspots for multivariate spatial data

Step M-1) Dimension reduction of the multivariate spatial data using the principal component analysis

Step M-2) Characterize the factors for each principal component

Step M-3) Draw the echelon dendrogram using principal component scores to detect the hotspots

Step M-4) Scan the regions from the upper echelon to the bottom, on the basis of the hierarchical structure of Step M-3

StepM-5) Detect the hotspots, which take the maximum $\log \lambda$ based on the likelihood ratio calculated to the factors detected on the StepM-2

4. Applications

4.1 Hotspot detection for 3D lattice data in the event of leachate-leaking accident

At final-disposal sites, the possibility exists of a leachate accident of waste material to groundwater because of liner sheet rupture. As an application of our study, we detect the hotspot with statistically significant pollution area for the simulated process of leachate advection-diffusion in the event of a leachate leaking accident. Figure 5 shows expected processes of leachate advection-diffusion in the event of a leachate-leaking accident at the three time points.

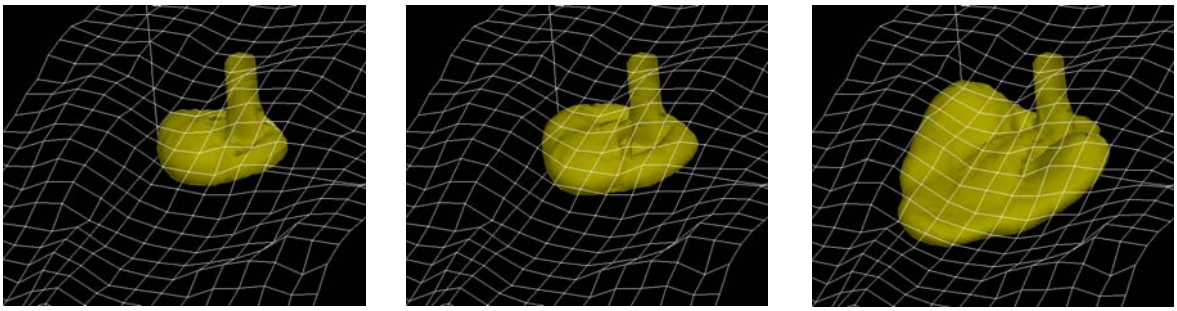

Figure 5: Simulation results of leachate advection-diffusion in the event of a leachate-leaking accident at the three time points. (t=1,2,3)

This consists of the physical space (x, y, z). Therefore, it can be said that these are 3D lattice data. These data consist of

$$\{D_1 \times D_2 \times D_3 \mid D_1 \in \{1,2,\ldots,61\},\ D_2 \in \{1,2,\ldots,101\},\ D_3 \in \{1,2,\ldots,41\}\}; \quad (4.1)$$

Each cell has leachate concentration C computed by solving the advection-diffusion equation. Because the volume of the leachate concentration data is huge, i.e. $61 \times 101 \times 41 = 252601$, we grouped the cells as

$$\{D_1 \times D_2 \times D_3 \mid D_1 \in \{1,2,\ldots,16\},\ D_2 \in \{1,2,\ldots,26\},\ D_3 \in \{1,2,...,11\}\} \quad (4.2)$$

and detected statistically significant pollution area (hotspot) based on echelon structure using spatial scan statistics. Table 2 shows the calculated result of the spatial scan statistics and the number of hotspot cells in each time.

Table 2: Result of hotspot detection for each time.

	$\log \lambda(Z)$	Number of hotspot cells	Ratio of hotspots
t=1	137.27	40	0.0087
t=2	183.86	47	0.0103
t=3	310.54	104	0.0227

4.2 Hotspots detection for 5 leading causes of death of Korea

We detect the candidate of hotspots for the 5 leading causes of death among the 16 counties of Korea in 2001. It is collected to promote national prosperity and to formulate the policy for health by Korea National Statistical Office in 2001. The mortality statistics were compiled in accordance with the World Health Organization (WHO) regulations, which specify that member nations classify causes of death. The 5 leading causes of death of Korea in 2001 are shown in Table 3.

(1) Malignant diseases (i.e. Cancer)
(2) Cerebrobascular diseases
(3) Diseases of heart
(4) Diabetes mellitus
(5) Chronic lower respiratory diseases

Thus we shall limit ourselves to these 5 leading causes in this study. We calculate the standardized mortality ratio (SMR) as a common intensity measurement and then apply the principal component analysis (PCA). We detect hotspots on the first principal component (PCA). Eigen values and vectors of PCA are shown in Table 4.

As the coefficients of the first component are all positive, it can be interpreted as an overall measure of the five variables. The echelon dendrogram using the score of the first component is shown in the left side of Figure 6. We calculate the spatial scan statistics according to Poisson model by aggregating regions for the echelon from the upper county in each peaks. The area Z with the maximum $\log \lambda(Z)$ becomes the candidates of hotspots. The result of hotspots for the PC1 are six counties (Busan, Ulsan, Gyeongnam, Gyeongbuk, Daegu and Jeonnam) in the first peak, where the spatial statistics $\lambda(Z)$

=1255.26. Using the spatial scan statistics, we can detect the candidates of hotspots which take maximum likelihood. We can also detect various shapes of hotspots. The second component represents a contrast between habitual disease (Heart diseases and Diabetes mellitus: positive sign) and the remaining variables (Cancer, Cerebrovascular diseases and Chronic lower respiratory diseases ; negative sign). Thus the habitual disease in PC2 can be interpreted as the main cause of death. In similar manner, the hotspots for PC2 turned out Seoul and Gyeonggi in the second peak, where the spatial statistics $\lambda(Z)$ =319.51 in the right side of Figure 6.

Table 3: The 5 leading causes of death of Korea in 2001.

counties	Total population	Cancer	Cerebrovascular diseases	Diseases of hearts	Diabetes mellitus	Chronic lower respiratory
Seoul	10263336	10077	5573	2823	1911	1139
Busan	3770536	4789	2853	1916	1000	688
Daegu	2525109	2903	1634	718	586	416
Incheon	2564598	2629	1813	724	518	373
Gwangju	1383765	1389	661	325	292	174
Daejeon	1403164	1370	868	352	250	217
Ulsan	1055618	915	557	281	209	167
Gyeonggi	9544494	9408	5953	2747	1870	1424
Gangwon	1552407	2274	1573	633	468	389
Chungbuk	1496520	2241	1511	466	373	365
Chungnam	1918137	3230	1981	740	553	514
Jeonbuk	2006454	3179	1979	692	570	552
Jeonnam	2099308	3953	1943	997	828	778
Gyeongbuk	2784704	4973	3329	1296	1005	992
Gyeongnam	3106502	4902	2735	1369	848	822
Jeju	546889	624	331	172	98	91

Table 4: Eigen values and vectors of PCA.

	I	II	III	IV	V
Cancer	0.4396	-0.5579	-0.2744	-0.4609	0.4559
Cerebrovascular diseases	0.4212	-0.1571	0.8764	0.1515	0.0824
Diseases of hearts	0.4156	0.6428	0.0400	-0.6086	-0.2053
Diabetes mellitus	0.4599	0.4085	-0.2898	0.5714	0.4596
Chronic lower respiratory	0.4951	-0.2900	-0.2664	0.2604	-0.7294
Eigen values	3.2463	0.7269	0.5166	0.3556	0.1547
Cumulative contribution ratio	0.6493	0.7946	0.8980	0.9691	1.0000

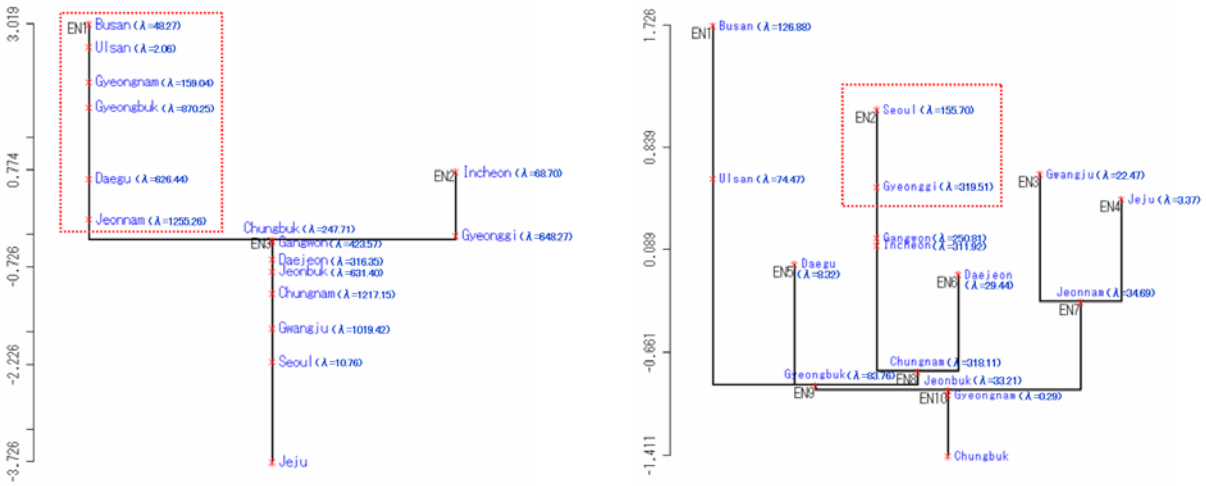

Figure 6: Echelon dendrogram of PC1(left) and PC2(right).

5. Conclusion

It is important to investigate the areas of significant risk (hotspots) about the effect on the human health status to make early warning for infectious diseases and so on. We can easily find the candidates of hotspots for any types of spatial data based on spatial scan statistics and echelon spatial hierarchical structure.

References

Anselin, L. (1995). Local indicators of spatial association-LISA., *Geographic Analysis*, 27, 93-115.

Ishioka, F., Kurihara, K., Suito, H., Horikawa, Y. and Ono Y. (2007). Detection of hotspots for three-dimensional spatial data and its application to environmental pollution. *Journal of Environmental Science for Sustainable Society*, 1, 15-24.

Kulldorff, M. (1997). A spatial scan statistics. *Communications in Statistics, Theory and Methods*, 26, 1481-1496.

Kurihara, K. (2004). Classification of geospatial lattice data and their graphical Representation. *Classification, Clustering, and Data Mining Applications* (Edited by D. Banks et al.), Springer, 251-258.

Kurihara, K., Myers, W.L. and Patil, G.P. (2000). Echelon analysis of the relationship between population and land cover patterns based on remote sensing data. *Community Ecology*, 1, 103-122.

Kurihara, K., Ishioka, F. and Moon, S. (2006). Detection of Hotspots on Spatial Data by Using Principal Component Analysis. *Journal of the Korean Data Analysis Society*, 8(2), 447-458.

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.

Myers, W. L., Kurihara, K., Patil, G. P. and Vraney, R. (2006). Finding upper level sets in cellular surface data using echelons and SaTScan. *Environmental and Ecological Statistics*, 13(4), 379-390.

Tomita, M., Hatsumichi, M. and Kurihara, K. (2008). Identify LD blocks based on hierarchical spatial data. *Computational Statistics & Data Analysis*, 52(4), 1806-1820.