Studies on Software Development to Detect Hotspot Cluster and Structural Analysis of Spatial Data Based on Echelon Analysis

SEPTEMBER 2021

Shoji KAJINISHI

Contents

1 Introduction

In spatial data analysis, it is very important to know "where the problem is occurring" in the analysis target area, and for that purpose, it is necessary to clearly understand the data structure (Cressie, 1993; Kurihara, 2000 , 2004). In the field of spatial epidemiology, it is necessary to spatially grasp the outbreak situation and evaluate it appropriately by using indicators such as mortality risk and disease morbidity risk. For example, indicators focusing on age composition and population differences between comparison target areas (Tango *et al.*, 2007) are often used. Furthermore, in recent years, a spatial hotspot cluster detection has been conducted to evaluate the state of accumulation of risks in a specific area (Kulldorff, 1997; Kurihara, 2003; Ishioka and Kurihara, 2012; Tango *et al.*, 2012). By this test, it is possible to obtain information on whether or not various risks that could not be judged only by the index value are statistically accumulated in a certain area. By analyzing with these methods used in the field of spatial epidemiology and utilizing the information obtained from the results, it is expected that measures will be taken to identify regional risk sources and eliminate them. For example, in the mid-19th century, John Snow thought the cause of the cholera epidemic in London was polluted water in wells and maped the residences and well locations of dead residents. As a result, it was clarified that the number of cholera patients was particularly high around a specific well. Based on the results obtained from the disease map, the end of cholera was seen when the water in the well was banned. In this example, if spatial hotspot cluster could be detected, it would have been possible to easily find a place where the risk of cholera mortality was significantly high, and it would have been possible to perform a more detailed analysis.

The software that can be applied to the analysis of spatial epidemiology is as follows. EcheScan (Kurihara *et al.*, 2020), DMS (Tango and Imai, 2013), SaTScan (Kulldorff *et al.*, 2020), EBPoiG (Takahashi, 2006). However, each of these tools is independent, and it is not possible to analyze risk index calculation, spatial hotspot cluster detection, and disease map output corresponding to those results in a series of flows. Therefore, it is necessary to use several independent tools individually to derive individual results, but the process is complicated because the specifications differ for each software. However, in the field of spatial epidemiology, it is common to carry out these analyzes in a series of steps, and we think that tools are needed to easily carry out these analyzes. Therefore, in this research, we construct an environment for comprehensively performing these series of analyzes using statistical software R.

As mentioned above, research to detect areas with statistically significantly higher values in spatial data analysis is very important. On the other hand, it is also an important

research field to grasp the distribution and structure of the obtained data and evaluate the complexity of the spatial data structure of the analysis target area. Spatial data has a data structure, and the data structure differs depending on the area to be analyzed, the type of data, the time when the data is aggregated, and so on. If the data structure can be grasped in detail, the spatial characteristics of the area can be known. Moreover, if the data structure can be evaluated quantitatively, the data structure can be compared and examined between regions and eras. This makes it possible to evaluate the characteristics of the data, which may help to find clues for research to investigate the causes of data structures that change with the times. In this way, what can be obtained from research on the structure of spatial data is considered to be very meaningful, but there are few such researches (Kurihara and Ishioka, 2007, Kurihara *et al.*, 2000), and there is no concrete method or evaluation method. The purpose of this study is to define indicators for assessing the structural complexity of spatial data. We will use the Echelon dendrogram generated by the Echelon analysis as a way to visually grasp the complexity of the spatial data structure. By confirming the structure of the generated dendrogram, it is possible to grasp the difference in the data structure and its change. However, there are various shapes in the Echelon dendrogram, and it is very difficult to grasp the differences and changes in the data structure just by looking at the shapes. In addition, no index that can be evaluated quantitatively is defined.

Therefore, in this study, we proposed an index that can evaluate the complexity of spatial data and defined the "stage" of dendrograms to compare the complexity of spatial data. First, the dendrogram is quantitatively expressed according to a certain rule, and the shape of the dendrogram is patterned. In the previous study, five indexes were defined for the dendrogram pattern, but in this paper, we proposed the index LV for judging more complicated shapes. By calculating these index values, it has become possible to evaluate the complexity of the dendrogram. Next, based on the idea of Echelon profiles of Echelon tree, we considered the complexity of the dendrogram from four scales and an index showing the structure called Cycle. Furthermore, we focused on the growth process of dendrogram complexity and defined the concept of dendrogram "stage". This makes it possible to evaluate changes in the structure of spatial data due to aging. However, the Echelon dendrogram used to grasp the structure of spatial data has the problem that the structure becomes very complicated as the number of regions increases. Therefore, we defined the "merge of peaks" that can simply express the structure while retaining the structural features of the dendrogram.

In this paper, Chapter 2 we introduce typical indicators of mortality risk(SMR, EB-SMR). Chapter 3 introduces Echelon analysis, which is useful for studying the topological structure of the surface of spatial data in a systematic and objective way. Chapter 4 describes how to detect hotspot clusters and how to scan them. Chapter 5 describes

how to use the developed software using specific examples. In Chapter 6 describes the indicators and methods for evaluating the complexity of the dendrogram. In Chapter 7, we will use actual data to confirm changes in the data structure over time. In Chapter 8 introduces the "merge of peaks" of dendrogram and considers its evaluation as a countermeasure when the number of regions is large and the spatial data structure becomes very complicated.

2 Indicators of mortality risk

2.1 SMR(Standardized Mortality Ratio)

In the field of spatial epidemiology, indicators calculated based on specific rules are often used when expressing the risk of illness or death. Furthermore, the situation can be grasped by visualizing the data using a disease map and spatially grasping the point of occurrence. Mortality is widely known as an indicator of mortality risk, but it is not always a suitable indicator when trying to compare regions. This is because it does not take into account the differences in population and age composition between the regions to be compared. In such cases, SMR(Standardized Mortality Ratio), which is an index that considers the influence of age composition, is often used.

Now, when the analysis target area consists of m areas in total, the SMR of the i-th area is given by the following equation;

$$\mathrm{SMR}_i = \frac{d_i}{e_i} \quad (i = 1, 2, \ldots, m). \tag{2.1}$$

Where, d_i represents the number of observed deaths in region i, and e_i represents the expected number of deaths in region i. In particular, the expected number of deaths e_i can be obtained by the following method considering the difference in age composition;

$$e_i = \sum_{k=1}^{K} n_{ik} P_k \quad (k = 1, 2, \ldots, K). \tag{2.2}$$

Where, n_{ik} represents the population of the k-th age group in the i region, P_k represents the mortality rate of the k-th age group of the expected population, and K represents the number of age groups. The SMR obtained in this way can be used for comparison between regions as an index considering the difference in age composition. However, although SMR is an index that takes into account differences in age composition, it has been pointed out that it is not appropriate to compare regional differences because it does not take into account differences in population between regions (Tango *et al.*, 2007). Therefore, I will explain Empirical Bayes estimate of SMR (EBSMR) considering the instability of less populated areas.

2.2 EBSMR(Empirical Bayes estimate of SMR)

The EBSMR in region i is given by the following equation using α and β estimated from the data.

$$\mathrm{EBSMR}_i = \frac{\beta + d_i}{\alpha + e_i} \tag{2.3}$$

Moment estimates or more precise maximum likelihood estimates are used for α and β. The EBSMR is explained in detail below.

Let d_i be the number of deaths in region $i(i = 1, 2, \ldots, m)$, let e_i be the expected number of deaths, and let θ_i be the unknown standardized mortality ratio. In general, the number of deaths is assumed to follow the Poisson distribution. In this case, the expected number of deaths multiplied by the standardized mortality ratio is considered to be the following Poisson distribution with $\theta_i e_i$ as the expected value.

$$d_i \sim Poisson(\theta_i e_i), \quad (d_i = 0, 1, 2, \ldots) \tag{2.4}$$

At this time, if θ_i is considered as a parameter, its maximum likelihood estimator is derived as follows.

$$\mathrm{SMR}_i = \hat{\theta}_i = \frac{d_i}{e_i} \quad (i = 1, 2, \ldots, m) \tag{2.5}$$

This is SMR. However, as mentioned earlier, SMR varies widely when the population is small, and it cannot be said to be an appropriate index when comparing regional differences. Therefore, to adjust for regional disparities in the population, we assume that $(\theta_1, \theta_2, \ldots, \theta_m)$ is a random variable that follows a continuous distribution, and each θ_i is considered a variable. The density function of the prior distribution is represented by $g(\theta|\eta)$, where η is a parameter that defines this distribution.

Since the number of deaths d_i follows the Poisson distribution, the probability density function of d_i is as follows.

$$f(d_i|\theta_i, e_i) = \frac{(\theta_i e_i)^{d_i} \exp(-\theta_i e_i)}{d_i!} \tag{2.6}$$

$$E(d_i) = \theta_i e_i, \qquad V(d_i) = \theta_i e_i$$

Therefore, using Bayes' theorem, the posterior distribution of θ_i is as follows.

$$h(\theta_i|e_i, d_i, \eta) = \frac{g(\theta_i|\eta) f(d_i|\theta_i, e_i)}{\int_0^\infty g(\theta_i|\eta) f(d_i|\theta_i, e_i) d\theta} \tag{2.7}$$

Therefore, the estimated value of SMR is as follows when the expected value from the posterior distribution is adopted.

$$\hat{\theta}_i \coloneqq E(\theta_i|e_i, d_i, \eta) = \int_0^\infty \theta h(\theta_i|e_i, d_i, \eta) d\theta = \frac{\int_0^\infty \theta g(\theta_i|\eta) f(d_i|\theta, e_i) d\theta}{\int_0^\infty g(\theta_i|\eta) f(d_i|\theta, e_i) d\theta} \tag{2.8}$$

Where, the problem is the estimation of the prior distribution parameters. One way to solve this is empirical Bayesian estimation based on the marginal likelihood of the following deaths d_i.

$$\prod_{i=1}^{m} Pr\{d_i|e_i, \eta\} = \prod_{i=1}^{m} \int_0^\infty g(\theta|\eta) f(d_i|\theta, e_i) d\theta \tag{2.9}$$

Where, we assume a Gamma distribution with $\eta = (\alpha, \beta)$ as the prior distribution of θ. That is, it becomes as follows.

$$g(\theta|\alpha,\beta) = \frac{\alpha(\alpha\theta)^{\beta-1}\exp(-\alpha\theta)}{\Gamma(\beta)} \tag{2.10}$$

$$E(\theta) = \frac{\beta}{\alpha}, \quad V(\theta) = \frac{\beta}{\alpha^2}$$

The reason why the Gamma distribution is often used for prior distribution is that the Gamma distribution is a prior distribution that is conjugate to the Poisson distribution. Therefore, from Bayes' theorem,

$$h(\theta_i|e_i,\alpha,\beta) = g(\theta|\beta+d_i,\alpha+e_i) \tag{2.11}$$

and posterior distribution are also gamma distributions, which is very convenient to calculate. This model is called the Poisson-Gamma model. In this case, the marginal likelihood of the number of deaths d_i has the following negative binomial distribution.

$$Pr\{d_i|e_i,\alpha,\beta\} = \frac{\Gamma(\beta+d_i)}{\Gamma(\beta)d_i!}\left(\frac{\alpha}{\alpha+e_i}\right)^{\beta}\left(\frac{e_i}{\alpha+e_i}\right)^{d_i} \tag{2.12}$$

$$E(d_k) = \frac{e_i\beta}{\alpha}, \quad V(d_k) = \frac{e_i(e_i+\alpha)\beta}{\alpha^2}$$

The moment estimator of (α, β) is as follows.

$$E\left\{\frac{1}{m}\sum_{i=1}^{m}\frac{d_i}{e_i}\right\} = \frac{\beta}{\alpha} \tag{2.13}$$

$$E\left\{\frac{1}{m}\sum_{i=1}^{m}\left(\frac{d_i}{e_i}-\frac{\beta}{\alpha}\right)^2\right\} = \frac{\beta}{\alpha^2} \tag{2.14}$$

Therefore, it is necessary to solve the following α and β.

$$\text{sample mean of SMR} = \frac{\beta}{\alpha}$$

$$\text{sample variance of SMR} = \frac{\beta}{\alpha^2}$$

The maximum likelihood estimator finds the α and β values ($\hat{\alpha}$ and $\hat{\beta}$) that maximize the following log-likelihood function.

$$\begin{aligned} l(\alpha,\beta) &= \log\sum_{i=1}^{m} Pr\{d_i|e_i,\alpha,\beta\} \\ &= \sum_{i=1}^{m}\sum_{s=0}^{d_i-1}\log(\beta+s) + m\beta\log\alpha - \beta\sum_{i=1}^{m}\log(\alpha+e_i) + \sum_{i=1}^{m}\{d_i\log e_i - d_i\log(\alpha+e_i)\} \end{aligned} \tag{2.15}$$

Therefore, it is necessary to solve the following likelihood function.

$$\frac{\partial l}{\partial \alpha} = \frac{\partial l}{\partial \beta} = 0 \tag{2.16}$$

This likelihood function can be solved using the Newton-Raphson method. Specifically, it is as follows.

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix}_{(k+1)} = \begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix}_{(k)} - \begin{bmatrix} \dfrac{\partial^2 l}{\partial^2 \alpha} & \dfrac{\partial^2 l}{\partial \alpha \partial \beta} \\ \dfrac{\partial^2 l}{\partial \alpha \partial \beta} & \dfrac{\partial^2 l}{\partial^2 \beta} \end{bmatrix}^{-1}_{(k)} \begin{bmatrix} \dfrac{\partial l}{\partial \alpha} \\ \dfrac{\partial l}{\partial \beta} \end{bmatrix}_{(k)} \tag{2.17}$$

$$\begin{cases} \dfrac{\partial l}{\partial \alpha} = m\dfrac{\beta}{\alpha} - \displaystyle\sum_{i=1}^{m} \frac{\beta + d_i}{\alpha + e_i} \\ \dfrac{\partial l}{\partial \beta} = \displaystyle\sum_{i=1}^{m} \sum_{s=0}^{d_i} \frac{1}{\beta + s} - \sum_{i=1}^{m} \log(1 + \frac{e_i}{\alpha}) \\ \dfrac{\partial^2 l}{\partial^2 \alpha} = -\dfrac{m\beta}{\alpha^2} + \displaystyle\sum_{i=1}^{m} \frac{\beta + d_i}{(\alpha + e_i)^2} \\ \dfrac{\partial^2 l}{\partial^2 \beta} = -\displaystyle\sum_{i=1}^{m} \sum_{s=0}^{d_i - 1} \frac{1}{(\beta + s)^2} \\ \dfrac{\partial^2 l}{\partial \alpha \partial \beta} = \dfrac{m}{\alpha} - \displaystyle\sum_{i=1}^{m} \frac{1}{\alpha + e_i} \end{cases}$$

For the initial value, use the following that can be calculated from $E(\theta) = \frac{\beta}{\alpha},\ V(\theta) = \frac{\beta}{\alpha^2}$.

$$\alpha = \frac{E(\theta)}{V(\theta)}, \quad \beta = \frac{E(\theta)^2}{V(\theta)}$$

Using the $\hat{\alpha}$ and $\hat{\beta}$ estimated in this way, the EBSMR of the i region can be calculated by the following equation.

$$\text{EBSMR}_i = \frac{\hat{\beta} + d_i}{\hat{\alpha} + e_i} \tag{2.18}$$

By transforming this equation as follows, the following features can be found.

$$\text{EBSMR}_i = \frac{\hat{\beta} + d_i}{\hat{\alpha} + e_i} = \frac{e_i}{\hat{\alpha} + e_i}\frac{d_i}{e_i} + \frac{\hat{\alpha}}{\hat{\alpha} + e_i}\frac{\hat{\beta}}{\hat{\alpha}} \tag{2.19}$$

1. EBSMR in populated areas approaches SMR($= d_i/e_i$) due to the higher expected number of deaths e_i.

2. EBSMR in less populated areas approaches the regional average ($= \hat{\beta}/\hat{\alpha}$) because the expected number of deaths e_i is smaller.

From these characteristics, it can be said that EBSMR is an index that is not easily affected by differences in population.

2.3 Comparison of SMR and EBSMR

In this section, the SMR and EBSMR introduced so far are visually expressed and evaluated. The data used here is Sudden Infant Death Syndrome (SIDS) in North Carolina, United States. This data will be explained in detail in Chapter 5. Figures 2.1 are map plots of SMR and EBSMR values. Areas with higher risk are painted red, and the lower the risk, the lighter the color. In addition, the Figures 2.2 are plots of the common logarithmic values of population on the horizontal axis and the risk values of SMR and EBSMR on the vertical axis, respectively. From these results, it can be seen that even with the same data, the magnitude relationship between the mortality risk value and the value between regions has changed. Furthermore, EBSMR can consider regions with a small population, and it can be seen that there is little variation in risk values between regions. Therefore, when dealing with mortality risk indicators, it is necessary to accurately capture and use data and regional characteristics.

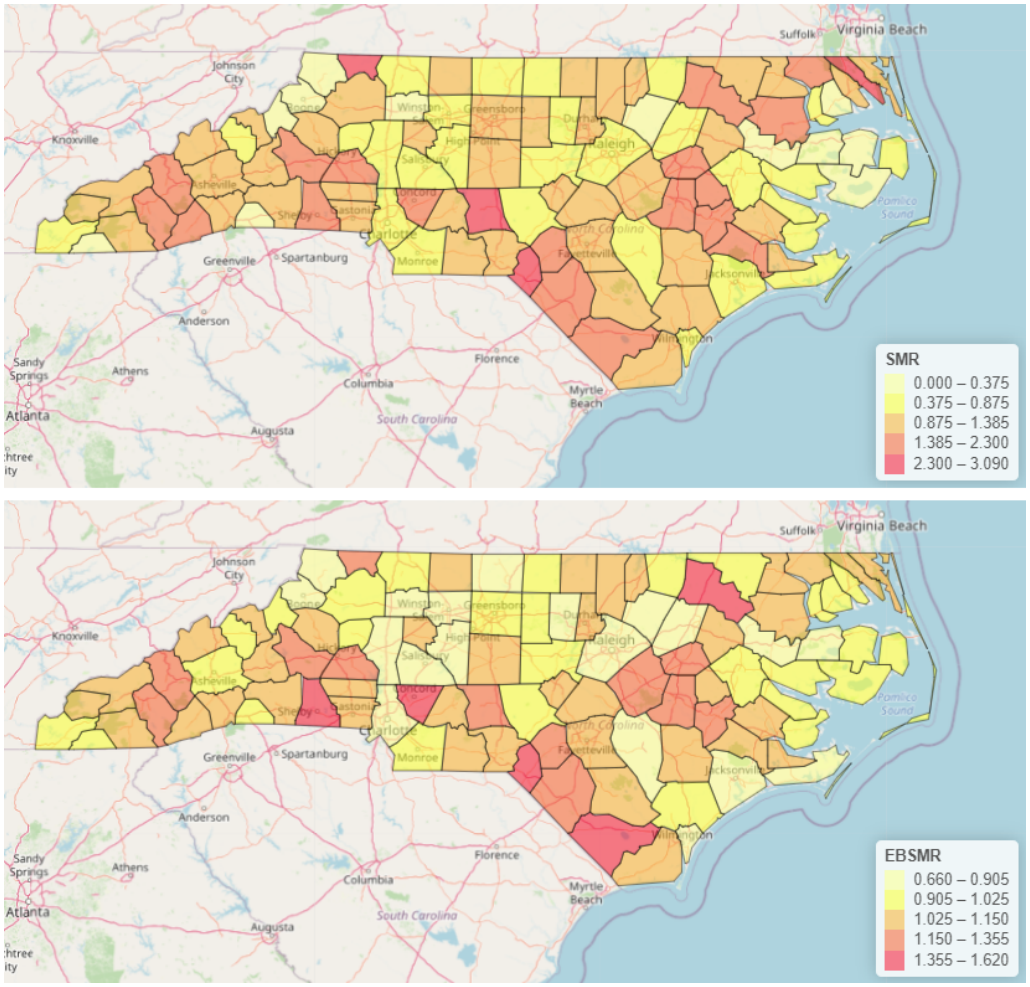


Figure 2.1: Disease map of SMR and EBSMR.

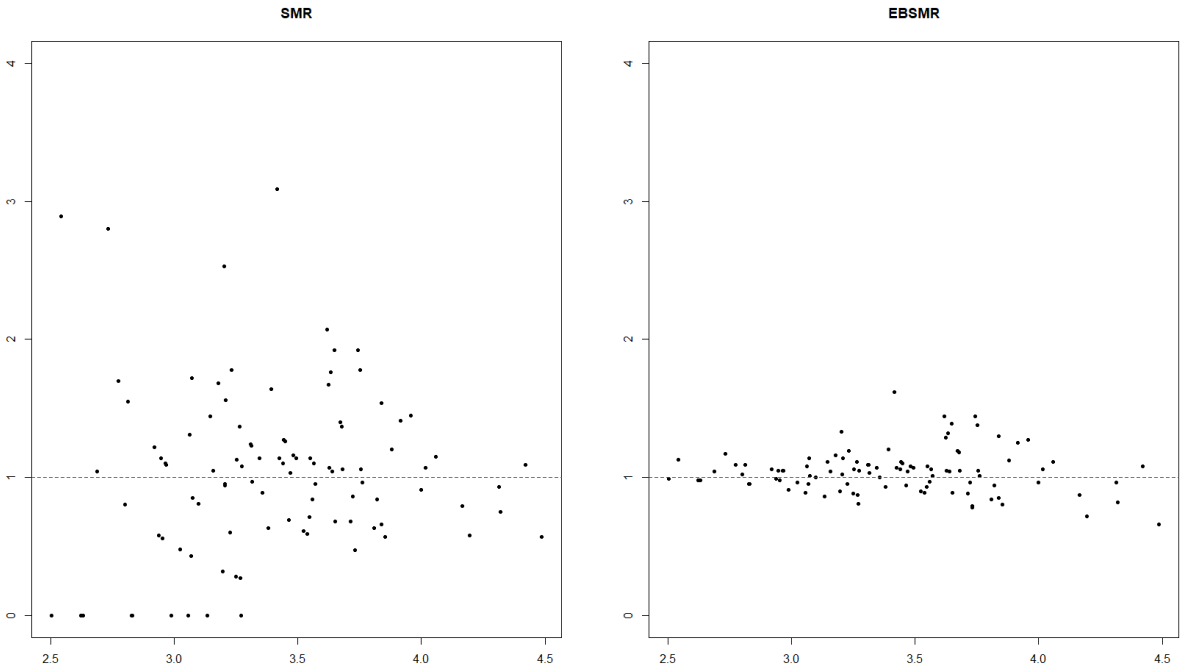


Figure 2.2: Scatter plot of SMR and EBSMR.

3 Echelon analysis

This chapter introduces Echelon analysis. Echelon analysis (Myers *et al.*, 1997) are useful techniques to study the topological structure of a surface in a systematic and objective manner. The Echelon dendrogram generated by the analysis expresses the surface topology and is useful in a wide range of fields. Now we consider the grid data of a fixed subset D of d-dimensional Euclidean space. Lattice data is observed over the entire spatial region, such as the cancer incidence in each region and the acquisition of the earth's surface by pixel remote sensing via satellite. Let $\mathbf{s} \in \mathbb{R}^{\mathbf{d}}$ be the data position in d-dimensional Euclidean space and $\mathbf{H}(\mathbf{s})$ be the random quantity of space position $\mathbf{s}$. Therefore, the location $\mathbf{s}$ produces a multivariate random field $\{\mathbf{H}(\mathbf{s}) : \mathbf{s} \in \mathbf{D}\}$ on a fixed non-random set $D \subset \mathbb{R}^d$. Observations are expressed by $\{\mathbf{h}(\mathbf{s}) : \mathbf{s} \in \mathbf{D}\}$. One-dimensional lattice data has a position ($x \in \mathbb{R}^1$) on the horizontal line and a value ($h(x)$) on the vertical line. For regularly divided lattice data in NL, these have an interval $s_i = l_1(i) = [i - 0.5, i + 0.5], i = 1, 2, \ldots, NL$. This lattice specifies an identifying set $D_1 = \{x_i | i = 1, 2, \ldots, NL\}$ using an index for a real seat $x_i = i$. Table 3.1 shows the intervals for $NL = 15$ from A to O, their values being $h(i)$.

Table 3.1: One dimensional spatial lattice data.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ID	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
$h(i)$	1	2	3	4	3	4	5	4	3	2	3	2	1	2	1

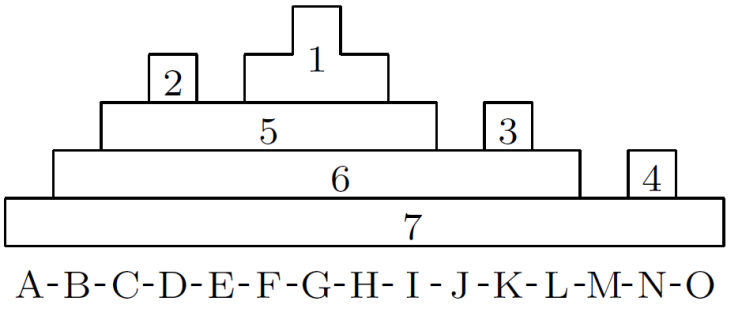


Figure 3.1: Topological subset on the surface.

Figure 3.1 shows the hypothetical topological subset on the surface level of one-dimensional lattice data. A total of seven numbered parts with the same topological

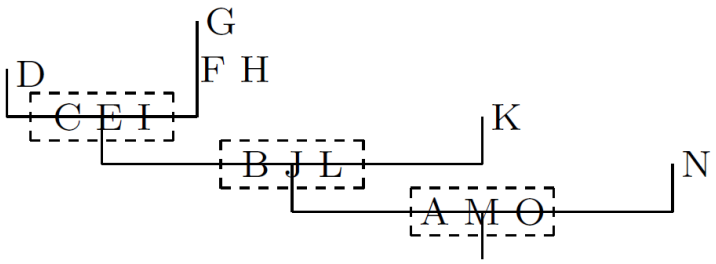


Figure 3.2: Hierarchical relation of echelons.

structure are in these hillforms. These classified parts are called echelons (Myers *et al.*, 1997), which consist of peaks, foundation, and root. Echelons $1, 2, 3$, and 4 are the peaks of hillforms. Echelon 5 is the foundation of two peaks. Echelons 6 and 7 are the foundations of foundation and peak. Echelon 7, which is the foundation of everything, is also called the route. The arrangement of echelons is recorded as a variation on parent and child relations. The family for an echelon consists of children of children through all overlying generations. Figure 3.2 shows a hierarchical representation of the topology structure given by the dendrogram. Specific numerical analyzes by Echelon analysis have already been proposed in various papers (Kurihara *et al.*, 2000, 2004, 2006; Ishioka *et al.*, 2007, 2012; Tomita *et al.*, 2008). Tables 1 and 2 show Algorithm 1 for finding peaks and Algorithm 2 for finding foundations in Echelon analysis. Algorithms 1 and 2 can find the echelons of the peaks and foundations with output values for input values in Table 3.2 and Table 3.3.

Table 3.2: Input and output values of Algorithm 1.

I / O	Name	Elements	Notes
I	LCT	$\{ i \mid i = 1, 2, \ldots, NL\}$	Counter of lattice
I	$H(i)$	$\{ h \mid h$ is the value of lattice $i\}$	Given
I	$NB(i)$	$\{ k \mid k$ is the neighbor of lattice $i\}$	Given
O	NP	Number of echelons (peaks)	1st-order echelon
O	$EN(i)$	$\{ k \mid k$ belongs to the i-th echelon$\}$	Lattice index
O	$NB(EN(j))$	$\{ k \mid k$ is the neighbor of the j-th echelon$\}$ $= \cup_{l \in EN(j)} NB(l) - EN(j)$	Lattice index

Algorithm 1 To find the i-th echelon $EN(i)$ of peak

Ensure: Find peaks

$i \Leftarrow 0$
while $LCT \neq \phi$ **do**
  $i \Leftarrow i + 1$
  $EN(i) \Leftarrow \phi$
  $M(i) \Leftarrow \arg \max_{j \in LCT} H(j)$
  while $H(M(i)) > \max_{j \in \{NB(M(i)) - EN(i)\}} H(j)$ **do**
    $EN(i) \Leftarrow EN(i) \cup \{M(i)\}$
    $LCT \Leftarrow LCT - \{M(i)\}$
    $M(i) \Leftarrow \arg \max_{j \in NB(EN(i))} H(j)$
  end while
  $LCT \Leftarrow LCT - \{M(i)\}$
  if $EN(i) = \phi$ **then**
    $i \Leftarrow i - 1$
  end if
  $NP \Leftarrow i$
end while

Algorithm 2 To find the i-th echelon $EN(i)$ of foundation

Ensure: Find foundations

$i \Leftarrow NP$
while $LCT \neq \phi$ **do**
  $i \Leftarrow i + 1$
  $EN(i) \Leftarrow \phi$
  $M(i) \Leftarrow \arg \max_{j \in LCT} H(j)$
  $CN \Leftarrow \{j \mid NB(M(i)) \cap FM(EN(j)) \neq \phi,\ j \in ECT\}$
  $FM(EN(i)) \Leftarrow \cup_{j \in CN} FM(EN(j)) \cup \{M(i)\}$
  $ECT \Leftarrow ECT \cup \{i\} - CN$
  while $\{NB(FM(EN(i))) - FM(EN(i))\} \neq \phi$ **do**
    while $H(M(i)) > \max_{j \in \{NB(FM(EN(i))) - FM(EN(i))\}} H(j)$ **do**
      $LCT \Leftarrow LCT - \{M(i)\}$
      $EN(i) \Leftarrow EN(i) \cup \{M(i)\}$
      $M(i) \Leftarrow \arg \max_{j \in \{NB(FM(EN(i))) \cap LCT\}} H(j)$
      $FM(EN(i)) \Leftarrow FM(EN(i)) \cup \{M(i)\}$
    end while
  end while
  if $LCT \neq \phi$ **then**
    $FM(EN(i)) \Leftarrow FM(EN(i)) - \{M(i)\}$
  end if
  $NE \Leftarrow i$
end while

Table 3.3: Input and output values of Algorithm 2.

I / O	Name	Elements	Notes
I	NP	Number of peaks	Obtained by Algorithm 1
I	LCT	$\{ i \mid i = 1, 2, \ldots, NL\} - \cup_{j=1}^{NP} EN(j)$	Counter of lattice
I	ECT	$\{ i \mid i = 1, 2, \ldots, NP\}$	Counter of echelon
I	$H(i)$	$\{ h \mid h$ is the value of lattice $i\}$	Given
I	$NB(i)$	$\{ k \mid k$ is the neighbor of lattice $i\}$	Given
I	$NB(EN(j))$	$\{ k \mid k$ is the neighbor of the j-th echelon$\}$ $= \cup_{l \in EN(j)} NB(l) - EN(j)$	Input and output
I	$FM(EN(j))$	$EN(j)$	Initial set for peaks
I	$NB(FM(EN(j)))$	$\{ k \mid k$ is the neighbor of the j-th family$\}$ $= \cup_{l \in FM(EN(j))} NB(l) - FM(EN(j))$	Initial set for family
O	NE	Number of echelons	Peaks and foundations
O	$EN(i)$	$\{ k \mid k$ belongs to the i-th echelon$\}$	Lattice index
O	$NB(EN(j))$	$\{ k \mid k$ is the neighbor of the j-th echelon$\}$ $= \cup_{l \in EN(j)} NB(l) - EN(j)$	Lattice index
O	$FM(EN(j))$	$\{ k \mid k$ is the family of the j-th echelon$\}$	Lattice index

4 Hotspot cluster detection

4.1 Hotspot cluster

From the results of disease maps drawn using indicators such as SMR and EBSMR, it is possible to spatially grasp the outbreak situation. However, it is very difficult to judge whether or not areas showing high index values are concentrated in a certain area. In general, there are always areas where the index value is relatively high, and it is possible that the high and low of the index value may be the result of observation within a chance range. In this chapter, we describe a method for identifying areas (hotspot clusters) where statistically significantly higher values are accumulated by the spatial hotspot cluster test.

4.2 Spatial scan statistics

Spatial scan statistic is a test method using a statistic based on the likelihood ratio proposed by Kulldorff (1997). Now, let Z be the area that is a candidate for the hotspot cluster, and let G be the entire area to be analyzed. Furthermore, let the mortality rate of Z is $p(Z)$, and the mortality rate of Z^c is $p(Z^c)$. Where, the hypothesis for detecting whether Z is a hotspot cluster is as follows.

$$\begin{cases} H_0 : p(Z) = p(Z^c) \\ \\ H_1 : p(Z) > p(Z^c) \end{cases} \tag{4.1}$$

Let $d(Z)$ be the number of deaths inside Z, $d(G)$ be the number of deaths in the entire area, $n(Z)$ be the population inside Z, and $n(G)$ be the population of the entire area. The probability of deaths $d(G)$ in all regions is expressed as follows based on the Poisson distribution.

$$\frac{(p(Z)n(Z) + p(Z^c)n(Z^c))^{d(G)}}{d(G)!} \exp(-(p(Z)n(Z) + p(Z^c)n(Z^c))) \tag{4.2}$$

In addition, the density $f(x)$ of the number of deaths at point x in all areas is as follows.

$$f(x) = \begin{cases} \dfrac{p(Z)n(x)}{p(Z)n(Z) + p(Z^c)n(Z^c)} & \text{if } x \in Z \\ \\ \dfrac{p(Z^c)n(x)}{p(Z)n(Z) + p(Z^c)n(Z^c)} & \text{if } x \in Z^c \end{cases} \tag{4.3}$$

Therefore, the likelihood function is given as follows.

$$\begin{aligned} L(Z, p(Z), p(Z^c)) &= \frac{(p(Z)n(Z) + p(Z^c)n(Z^c))^{d(G)}}{d(G)!} \exp(-(p(Z)n(Z) + p(Z^c)n(Z^c))) \\ &\quad \times \prod_{x \in Z} \frac{p(Z)n(x)}{p(Z)n(Z) + p(Z^c)n(Z^c)} \prod_{x \in Z^c} \frac{p(Z^c)n(x)}{p(Z)n(Z) + p(Z^c)n(Z^c)} \qquad (4.4) \\ &= \frac{1}{d(G)!} \exp(-(p(Z)n(Z) + p(Z^c)n(Z^c)))p(Z)^{d(Z)}p(Z^c)^{d(Z^c)} \prod_x n(x) \end{aligned}$$

In order to maximize this likelihood function, we solve the maximum likelihood function in the situation given the region Z.

Substituting the maximum likelihood estimators $\hat{p}(Z) = \frac{d(Z)}{n(Z)}$ and $\hat{p}(Z^c) = \frac{d(Z^c)}{n(Z^c)}$ into formula (4.4) gives the following.

$$L(Z) = \frac{1}{d(G)!} \exp\left(-d(G)\right) \left(\frac{d(Z)}{n(Z)}\right)^{d(Z)} \left(\frac{d(Z^c)}{n(Z^c)}\right)^{d(Z^c)} \prod_x n(x) \qquad (4.5)$$

On the other hand, substituting maximum likelihood estimators $\hat{p}(Z) = \hat{p}(Z^c) = \frac{d(G)}{n(G)}$ under H_0 gives the following equation.

$$L_0 = \frac{1}{d(G)!} \exp\left(-d(G)\right) \left(\frac{d(G)}{n(G)}\right)^{d(G)} \prod_x n(x) \qquad (4.6)$$

Therefore, when $\frac{d(Z)}{n(Z)} > \frac{d(Z^c)}{n(Z^c)}$, the likelihood ratio test amount λ is as follows.

$$\lambda(Z) = \frac{L(Z)}{L_0} = \frac{\left(\frac{d(Z)}{n(Z)}\right)^{d(Z)} \left(\frac{d(Z^c)}{n(Z^c)}\right)^{d(Z^c)}}{\left(\frac{d(G)}{n(G)}\right)^{d(G)}}, \qquad \left(\frac{d(Z)}{n(Z)} > \frac{d(Z^c)}{n(Z^c)}\right) \qquad (4.7)$$

Where, using the age-adjusted expected value e, from $e(G) = d(G)$, equation (4.7) is expressed as follows when $d(Z) > e(Z)$.

$$\lambda(Z) = \left(\frac{d(Z)}{e(Z)}\right)^{d(Z)} \left(\frac{d(Z^c)}{e(Z^c)}\right)^{d(Z^c)} \qquad (4.8)$$

At this time, the region Z that maximizes the likelihood ratio $\lambda(Z)$ is considered as a candidate for the hotspot cluster. The significance of hotspot clusters is evaluated by the Monte Carlo test.

4.3 Scan method

In order to detect hotspot cluster, it is necessary to find the region group Z that maximizes the likelihood ratio $\lambda(Z)$ in formula (4.8) (this is called scanning region group Z). However, unless the number of regions is extremely small, it is generally impossible to scan all the patterns of region group Z when the number of regions becomes too large. Where, we introduce the following two methods proposed to scan Z efficiently.

4.3.1 Echelon scan method

The Echelon scan method is a scan method that finds hotspot cluster using Echelon analysis (Myers, 1997). This method scans the area based on the structure of the spatial data obtained by the Echelon dendrogram. As a feature, non-circular hotspot cluster can also be identified, and since scanning is performed from the top of the hierarchical structure of the data, the calculation cost can be suppressed. Therefore, it can be applied to large-scale data (Kurihara, 2003; Ishioka and Kurihara, 2012; Ishioka *et al.*, 2019). The algorithm of the echelon scan technique is proposed in Algorithm 3 with output values for input values in Table 4.1.

Algorithm 3 To find the maximum $\log \lambda(Z)$ based on Echelon scan

Ensure: Find maximum LLR($\log \lambda(Z)$)
  $MAXLLR \Leftarrow -\infty$
  $MAXZ \Leftarrow \phi$
  $i \Leftarrow 1$
  while $i \leq NE$ **do**
    $j \Leftarrow 1$
    while $j \leq N(i)$ **and** $HE(i,j) > MAXHV$ **do**
      if $LLR(HE(ZE(i,j))) > MAXLLR$ **then**
        $MAXLLR \Leftarrow LLR(HE(ZE(i,j)))$
        $MAXZ \Leftarrow ZE(i,j)$
      end if
      $j \Leftarrow j+1$
    end while
    $i \Leftarrow i+1$
  end while

4.3.2 Circular scan method

Kulldorff (1997) proposed a circular scan method that scans concentrically as a method for determining candidate Z for hotspot cluster. In this method, one representative point of a certain area i (location of government office, center of gravity of the area, etc.) is determined, and a concentric circle with radius r is drawn around that point. When the representative points of the region $j(\neq i)$ are included in the circle, i and j are merged to obtain hotspot cluster candidate $Z = \{i, j\}$. The radius r is expanded until it reaches a certain critical value (the total population in Z, the length of r, the number of regions included in Z, etc.) predetermined by the analyst. In the entire set of Z scanned, Z that maximizes the value of $\lambda(Z)$ is detected as a hotspot cluster. It has been pointed out that this method is excellent in detecting circular hotspot cluster because it scans in a circular shape, but is not suitable for detecting non-circular hotspot cluster.

Table 4.1: Input and output values of Algorithm 3.

I O	Name	Elements	Notes
I	NE	Number of echelons	Obtained by Algorithms 1 and 2
I	NP	Number of peaks	Obtained by Algorithm 1
I	$N(i)$	Number of lattice elements for the i-th echelon	$NL = \sum_{i=1}^{NE} N(i)$
I	$HE(i,j)$	Value of the j-th element of i-th echelon $HE(i,1) > \cdots > HE(i,N(i))$	Given
I	$MAXHV$	Value of the $[NL/2]$-th lattice in order	Specified value
I	$ZE(i,j)$	Scan window of the upper j lattices for the i-th echelon $\cup_{k=1}^{j}\{(EN(i),k)\}$	Scan window for peak $i = 1,\ldots,NP$, $j = 1,\ldots,N(i)$
I	$ZE(i,j)$	Scan window of the upper j lattices for the i-th echelon $ZE(CH(i),CH(N(i))) \cup_{k=1}^{j} \{(EN(i),k)\}$ where $ZE(CH(i),CH(N(i)))$ $= \cup_{EN(j)\in CH(EN(i))} \cup_{k=1}^{N(j)} \{(EN(j)),k)\}$	Scan window for foundation $i = NP+1,\ldots,NE$, $j = 1,\ldots,N(i)$
I	$HE(ZE(i,j))$	Value of upper j lattices for the i-th echelon window $ZE(i,j)$	Value for peak and foundation $i = 1,2,\ldots,NE$, $j = 1,2,\ldots,N(i)$
O	$MAXLLR$	Log likelihood for the candidate of hotspot cluster	log LR(Z)
O	$MAXZ$	Window for the candidate of hotspot cluster	

5 Software development using R Shiny

In this paper, we constructed the software that can perform the series of analysis described so far using the R package shiny. There are three reasons for using shiny.

1. The user interface (UI) for web application development can be flexibly designed.
2. The only means by which the Echelon scan method can be performed is the R echelon package (Ishioka, 2020).
3. We are in a position to use the source code of the echelon package, so we are in an environment where data input, parameter setting, analysis result output, etc. necessary for executing the echelon scan method can be implemented on shiny.

Regarding (1) the UI of shiny can interactively change various parameters set for analysis using slider bars and text boxes, so the result of recalculation with a simple operation can be output.

With conventional software, it was necessary to perform calculation of risk indicators such as SMR and EBSMR, detection of spatial hotspot cluster, map drawing etc in independent environments. However, in the field of spatial epidemiology, it is desirable to perform these as a series of analyzes. In addition, most of the map drawing uses the map information prepared in the software in advance, and few can use the map information (shapefile, etc.) of an arbitrary analysis target area requested by the user. The software developed in this research can execute these series of analyzes on the web for any area given by the user using a shapefile. This makes it possible to easily check the geographical information of the analysis target area, such as whether or not the high-risk area is in the city center and the condition of the transportation network around it. In addition, it is expected that displaying the results of hotspot cluster on a map will help to find clues for investigating the cause of cluster occurrence.

The beta version of the developed software is available at the following URL(https://fishi.ems.okayama-u.ac.jp/kajinishi/ver0.1/). The analysis example shown in this paper can be reproduced with the example data on the top page of the above site. This chapter describes the software.

5.1 Shiny

Shiny is a package that allows you to create interactive web applications. Shiny has the following features.

1. You can build a web application with just a few lines of code that does not require JavaScript.
2. You can dynamically draw a spreadsheet table like Excel.
3. All UI can be built with R, and more flexible with HTML, CSS, and JavaScript.
4. You can use R's integrated development environment (R Console, RGui for Windows or Mac, RStudio, etc.).
5. Widgets (parts) for displaying inputs such as sliders and buttons and outputs such as charts are prepared as R objects.

5.2 Structure of data required for analysis

This software can perform analysis using data arbitrarily prepared by the analyst. Table5.1 shows the input contents required for the software, and Table 5.2 shows the output contents.

The shapefile is a map data file consisting of topographical information and attribute information proposed by ESRI, and is composed of multiple files. This software uses four files consisting of "shp(main file to store feature geometry)", "shx(index file that stores the index of feature geometry)", "dbf(dBASE table that stores feature attribute information)", and "prj(file that stores coordinate system information)". In addition, observation data (Case data) and population data (Population data) will be prepared in CSV format. The structure of the data using in the analysis is the first column consists of IDs that can identify the area, and the second and subsequent columns consist of observed data as shown in Figure 5.1. Select one from the attribute information given in the shapefile and describe it in the ID of the first column. For example, the number or area name given to each area. The method of checking the attribute information in the shapefile will be described in the next section. The data in Figure 5.1 shows a part of the data on the number of male suicides in the Chugoku region in 2016. The first column of the data is the city code given by city, and the second and subsequent columns are the number of suicides by age group in each area. In this example, it is divided into eight age groups in total, but it is also possible to give data that is not divided by age group. However, the age groups of the observational data and the population data must be the same.

5.3 Analysis example using software

In this section, we will introduce an example of analysis using software. The first case shows data that is not divided by age group, and the second case shows an example that

Table 5.1: Input content.

Input	Input content
Shapefile	shp, shx, dbf, prj files
Area id	ID
Area name	Region name
Case data	Observation data(file format : csv)
Population data	Population data(file format : csv)

Table 5.2: Output content.

Tab	Output content(External file output format)
SMR & EBSMR	Disease map Result table(csv)
Echelon scan	Disease map Result table(csv) Echelon dendrogram(png, pdf, eps)
Circular scan	Disease map Result table(csv)

ID	～20age	20～30age	30～40age	40～50age	50～60age	60～70age	70～80age	80age～
31201	0	0	5	5	1	1	0	4
31202	0	1	4	3	1	4	1	1
31203	0	0	2	3	3	1	0	2
31204	0	0	0	1	1	1	0	0
31302	0	0	0	3	0	0	0	0
31325	0	0	0	0	0	0	0	0
31328	0	0	0	0	0	0	0	2
31329	0	0	0	0	1	1	0	1
31364	0	0	0	0	0	0	1	1
31370	0	1	0	1	0	1	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 5.1: Structure of data to prepare.

uses data that is divided into eight age groups.

5.3.1 Analysis of data not classified by age group

We use data from Sudden Infant Death Syndrome (SIDS) in North Carolina, United States. This data is a compilation of the number of sudden infant deaths in each region of North Carolina from July 1974 to June 1978, and is not classified by age group. This is the case when $K = 1$ in the formula (2.2).

The analysis procedure of this software using this data is shown below.
(**Procedure1**) Access the page of this software and click "Browse" of [Shapefile] on the left side of the screen as shown in Figure 5.2 to load the four files that make up the shapefile shown in Section 5.2 at the same time. The attribute information in the read shapefile can be confirmed as shown in Figure 5.3 by clicking [Check].

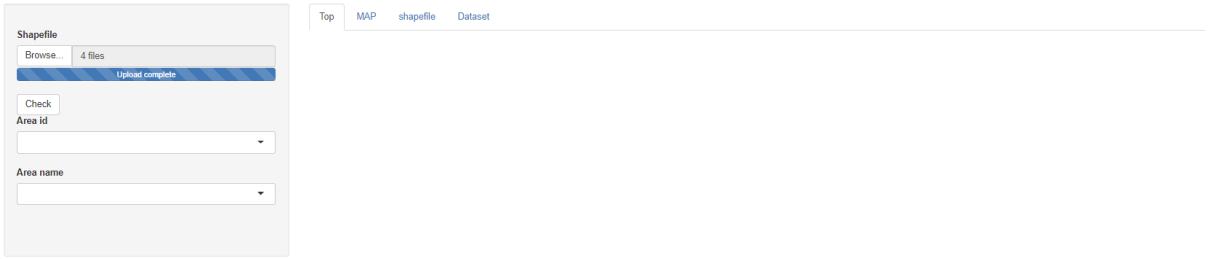


Figure 5.2: Screen when software is started and shapefile is read.

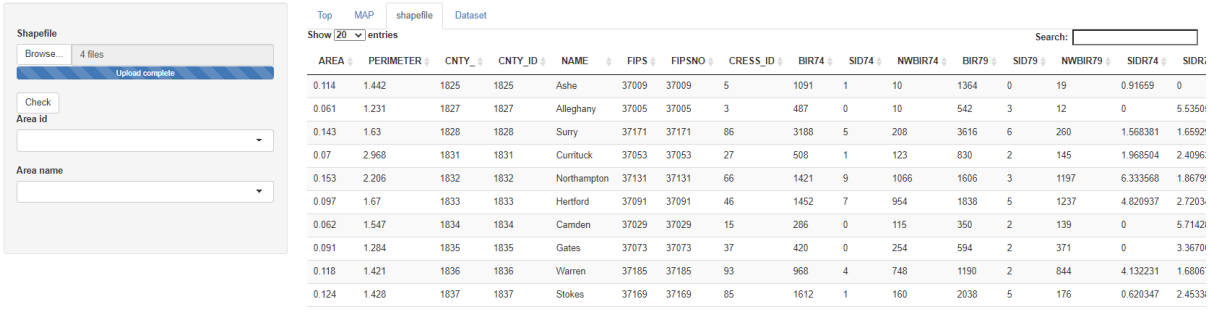


Figure 5.3: Screen when attribute information is displayed from [Check].

(**Procedure2**) Select Area id and Area name from the attribute information (Figure 5.3) of the shapefile as shown in Figure 5.4. For [Area id], select the variable name corresponding to the ID in the first column of the prepared observation data and population data (here, select CRESS ID). For [Area name], select the area name you want to display in the software (here, select NAME).

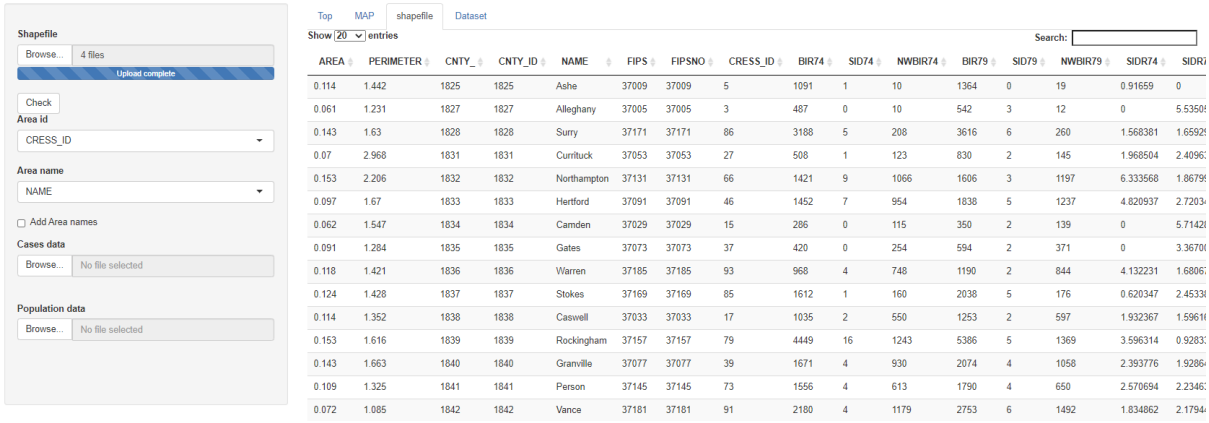


Figure 5.4: Screen with Area id and Area name selected.

(**Procedure3**) Read the prepared data. Enter observation data in [Case data] and population data in [Population data].
(**Procedure4**) Click the [RUN] button at the bottom of the page to start the analysis. The display of the result can be switched by the [SMR & EBSMR], [Echelon scan], and [Circular scan] tabs at the bottom of the map.

Figure 5.5 shows the disease map displayed on the [SMR & EBSMR] tab by pressing the [RUN] button. Table 5.3 shows the parameters that can be changed on the [SMR & EBSMR] tab. The parameters can be changed interactively. For example, the color density can be from 0 (= transmittance 100%) to 1 (= transmittance 0%), and the color coding can be from 3 to a maximum of 10 and the color coding method is quantile (equal classification), equal (equally spaced classification), fisher (natural class classification), pretty (visually easy-to-understand classification) can be selected. The results for SMR and EBSMR can be confirmed in the table at the bottom of the page as shown in Figure 5.6. This analysis result can be output to an external file (csv format) from [Download]. Table 5.4 shows the details of each item in the result table output in Figure 5.6.

Table 5.3: Parameter description on the [SMR & EBSMR] tab.

Parameter name	Contents	Changeable range	Parameter type
Polygons	Fill polygon	ON / OFF	Checkbox
Legend	Displaying the legend	ON / OFF	Checkbox
Shade of color	Color intensity	$0 \sim 1(0, 0.1, \ldots, 1)$	Slider bar
SMR&EBSMR	Which index to color	SMR / EBSMR	Select box
Class interval	Color coding method	quantile, equal, fisher, pretty	Select box
Number of class of color	Number of color coded divisions	$3 \sim 10(3, 4, \ldots, 10)$	Slider bar

Figure 5.7 shows the situation when the [Echelon scan] tab is selected. The hotspot cluster detected by Echelon scan are displayed on the map. In the [Echelon scan] tab as

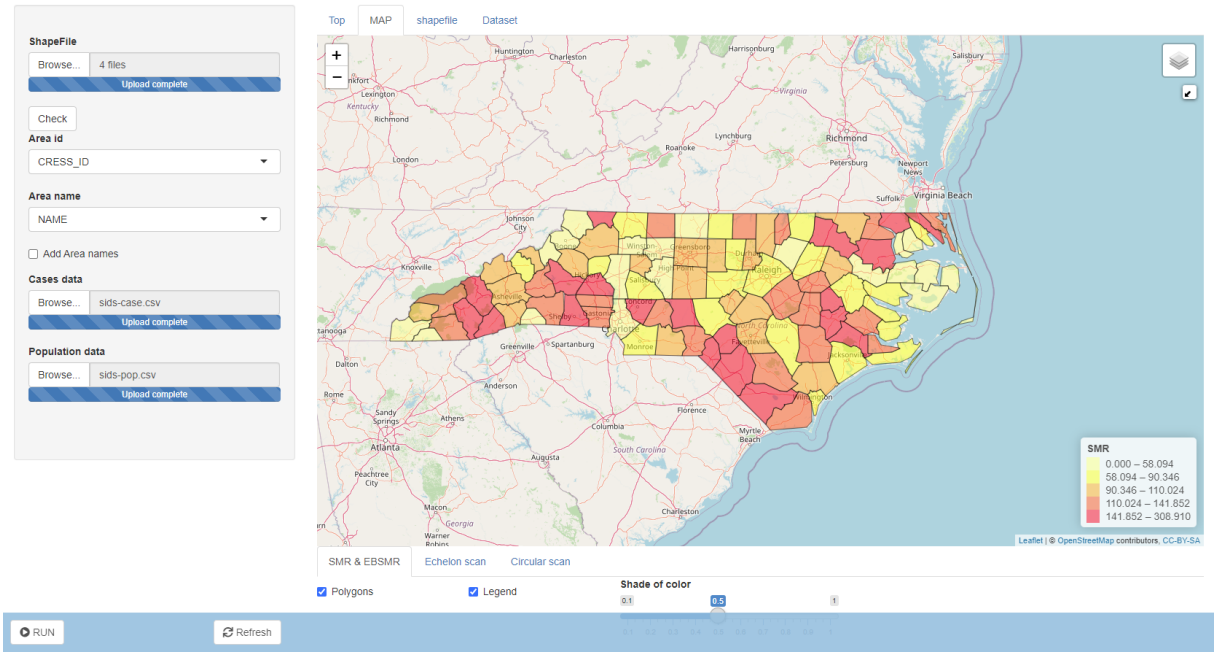
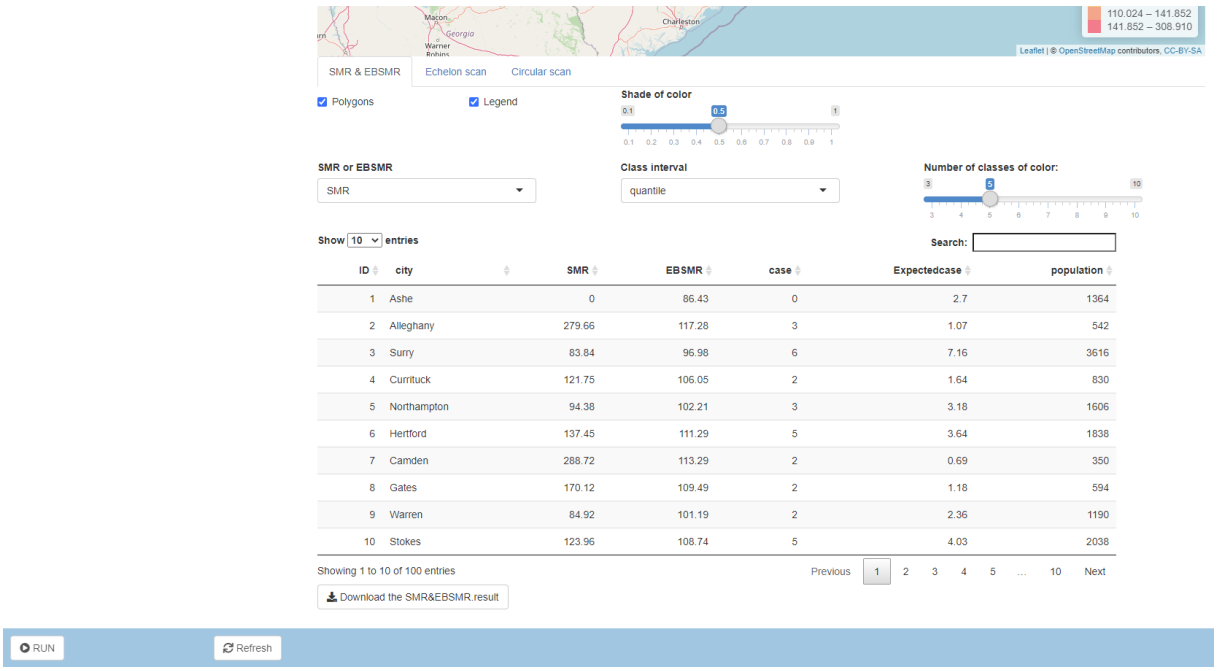


Figure 5.5: Screen immediately after pressing the [RUN] button.



ID	city	SMR	EBSMR	case	Expectedcase	population
1	Ashe	0	86.43	0	2.7	1364
2	Alleghany	279.66	117.28	3	1.07	542
3	Surry	83.84	96.98	6	7.16	3616
4	Currituck	121.75	106.05	2	1.64	830
5	Northampton	94.38	102.21	3	3.18	1606
6	Hertford	137.45	111.29	5	3.64	1838
7	Camden	288.72	113.29	2	0.69	350
8	Gates	170.12	109.49	2	1.18	594
9	Warren	84.92	101.19	2	2.36	1190
10	Stokes	123.96	108.74	5	4.03	2038

Figure 5.6: Result table on the [SMR & EBSMR] tab.

Table 5.4: Item description of [SMR & EBSMR] result table.

Item name	Contents
ID	Regional number
city	Region name
SMR	Value of SMR
EBSMR	Value of EBSMR
case	Number of observations
Expectedcase	Expected number of observations
population	Population

well, various parameters can be changed dynamically as shown in Figure 5.8. The details of the parameters are shown in Table 5.5.

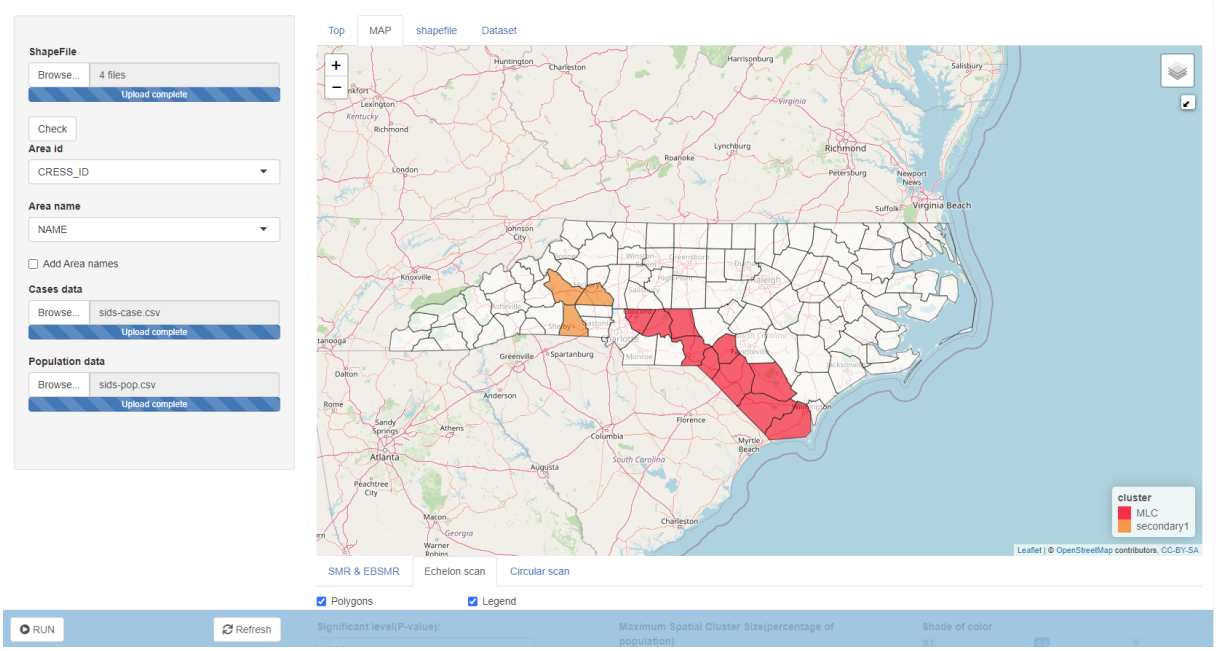


Figure 5.7: Screen when the Echelon scan tab is executed.

Information on the detected hotspot cluster is displayed at the bottom of the parameters. Similar to the table on the [SMR & EBSMR] tab, the results can be output to an external file. The explanation of each item in this result table is as shown in Table 5.6. In this table, the $\log \lambda(Z)$ of the formula (4.8) is output in descending order. Furthermore, on the [Echelon scan] tab, you can see the Echelon dendrogram that represents the structure of the data topologically, as shown in Figure 5.9. Table 5.7 shows the details of the parameters related to the dendrogram. For example, by adjusting the $x-$axis and $y-$axis ranges, the structure of the entire dendrogram can be confirmed, and even the parts that are dense and difficult to confirm in detail can be enlarged and displayed in detail. Dendrogram can be output to an external file in png, pdf, eps format. The [Circular scan]

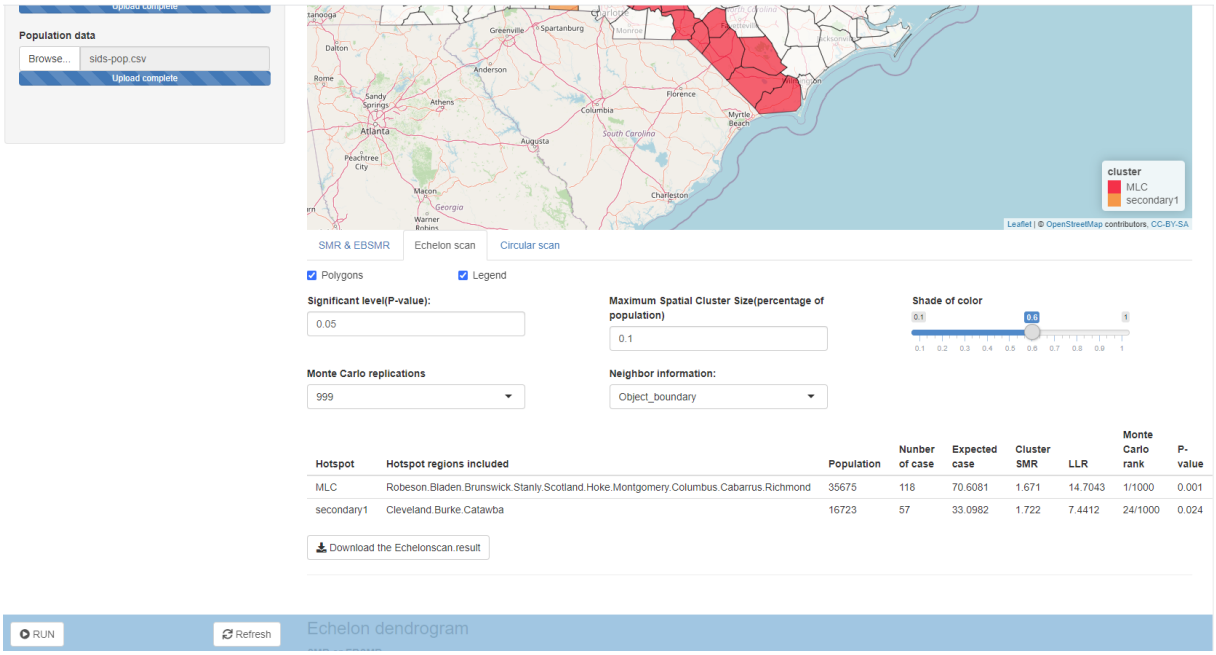


Figure 5.8: Screen of various parameters and scan result table.

Table 5.5: Echelon scan parameter description.

Parameter name	Contents	Changeable range	Parameter type
Polygons	Fill polygon	ON / OFF	Check box
Legend	Displaying the legend	ON / OFF	Check box
Significant level	Acceptable p value	0.01∼0.99	Select box
Maximum Spatial Cluster size	Maximum range to scan	0.1∼0.9	Select box
Shade of color	Color intensity	$0.1 \sim 1(0.1, 0.2, \ldots, 1)$	Slider bar
Monte Carlo repications	Number of simulations	999 / 9999	Select box
Neighbor infomation	Definition of neighborhood information	Object boundary, Distance, Delaunay triangle	Select box

tab has the same specifications as the [Echelon scan] tab, and analysis can be performed with various settings by changing various parameters.

Table 5.6: Item description of Echelon scan result table.

Item name	Contents
Hotspot	Hotspot cluster ranking
Hotspot regions included	Region name detected as hotspot cluster
Population	Population in hotspot cluster
Number of case	Number of observations in hotspot cluster
Expected case	Expected number of observations in hotspot cluster
Cluster SMR	SMR in hotspot cluster
LLR	Log likelihood ratio
Monte Carlo rank	Monte Carlo ranking
P-value	P value

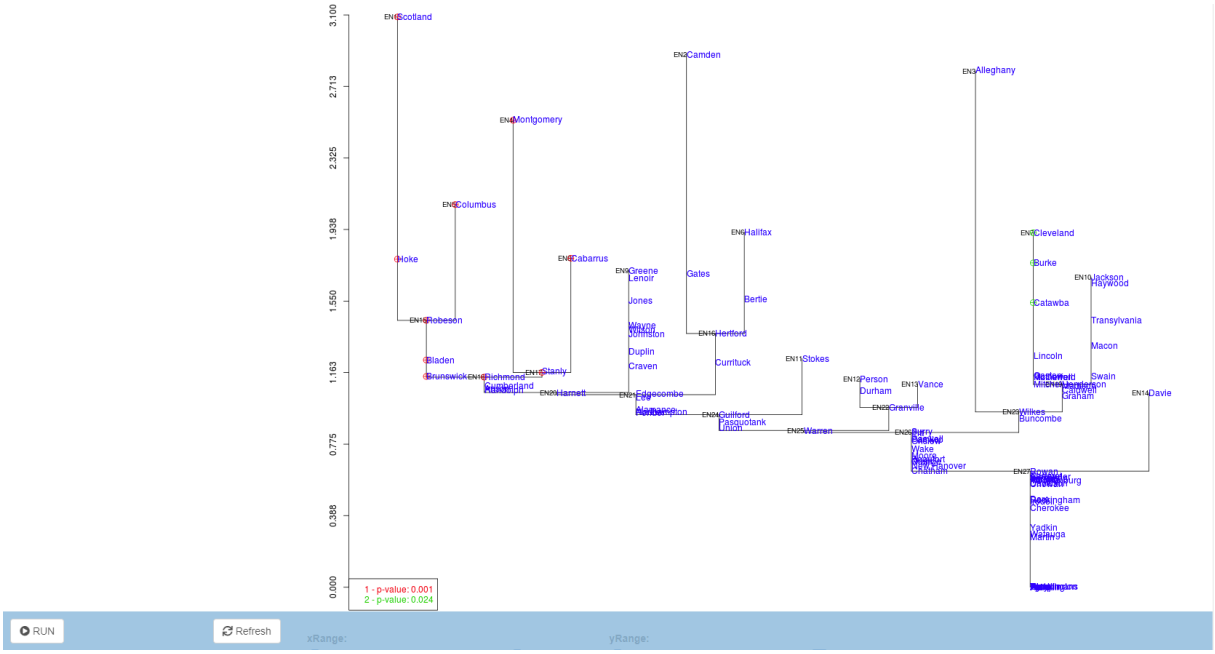


Figure 5.9: Screen displaying dendrogram.

5.3.2 Analysis of data by age group

We will analyze the number of male suicides in the Chugoku region in 2016. This data is divided into eight age groups, and is the case of $K = 8$ in the formula (2.2). In this data, we show an example of analyzing the following two cases with different age groups.

1. When not considering the number of age groups ($K = 1$)

Table 5.7: Parameter description in the Echelon dendrogram.

Parameter name	Contents	Changeable range	Parameter type
SMR or EBSMR	Index to draw dendrogram	SMR or EBSMR	Select box
xRange	Changeable horizontal axis range	0〜1	Slider bar
yRange	Changeable vertical axis range	0〜Maximum value of data	Slider bar
Region Names	Display of region name	ON/OFF	Checkbox
Symbol	Display of symbols	ON/OFF	Checkbox
font size	Character size	0〜1(0.1, 0.2, . . . ,1)	Slider bar
Download figure	Dendrogram download format	eps, pdf, png	Checkbox

2. When considering the number of age groups ($K = 8$)

Figure 5.10 and 5.11 show the calculation results of the mortality risk output by the software for "when age group is not considered ($K = 1$)" and "when age group is considered ($K = 8$)", respectively. As is clear from these figures, the expected number of deaths e_i for formula (2.2) changes depending on whether or not the age group is taken into consideration, so the SMR results for formula (2.1) also change. As a result, the two dendrograms (Figure 5.12 and Figure 5.14) created based on each SMR differ in their shape and the regions that make up each hierarchy. Differences also appear in the hotspot cluster detected by Echelon scan (Figure 5.13 and Figure 5.15). Considering the age group, three regions were excluded from the hotspot cluster. These areas have a large population in the age group with a large number of suicides, unlike the population composition ratio of the entire analysis target area. Therefore, it is considered that the SMR in these three regions became relatively small, resulting in such a result.

ID	city	SMR	EBSMR	case	Expectedcase	population
1	鳥取県鳥取市	0.53	0.72	16	30.11	190960
2	鳥取県米子市	0.64	0.82	15	23.56	149407
3	鳥取県倉吉市	1.44	1.19	11	7.62	48340
4	鳥取県境港市	0.55	0.94	3	5.49	34813
5	鳥取県岩美町	1.6	1.13	3	1.87	11891
6	鳥取県若桜町	0	1.04	0	0.53	3377
7	鳥取県智頭町	1.73	1.12	2	1.16	7348
8	鳥取県八頭町	1.08	1.08	3	2.78	17629
9	鳥取県三朝町	1.89	1.13	2	1.06	6720
10	鳥取県湯梨浜町	1.49	1.14	4	2.69	17083

Showing 1 to 10 of 107 entries

Previous 1 2 3 4 5 ... 11 Next

Download the SMR&EBSMR.result

Figure 5.10: Results of SMR and EBSMR when age group is not considered ($K = 1$).

ID	city	SMR	EBSMR	case	Expectedcase	population
1	鳥取県鳥取市	0.53	0.74	16	29.92	190960
2	鳥取県米子市	0.64	0.83	15	23.28	149407
3	鳥取県倉吉市	1.43	1.16	11	7.71	48340
4	鳥取県境港市	0.54	0.94	3	5.54	34813
5	鳥取県岩美町	1.54	1.1	3	1.95	11891
6	鳥取県若桜町	0	1.03	0	0.59	3377
7	鳥取県智頭町	1.62	1.09	2	1.24	7348
8	鳥取県八頭町	1.05	1.06	3	2.85	17629
9	鳥取県三朝町	1.81	1.1	2	1.1	6720
10	鳥取県湯梨浜町	1.49	1.11	4	2.69	17083

Showing 1 to 10 of 107 entries

Previous 1 2 3 4 5 ... 11 Next

Download the SMR&EBSMR.result

Figure 5.11: Results of SMR and EBSMR when age group is considered ($K = 8$).

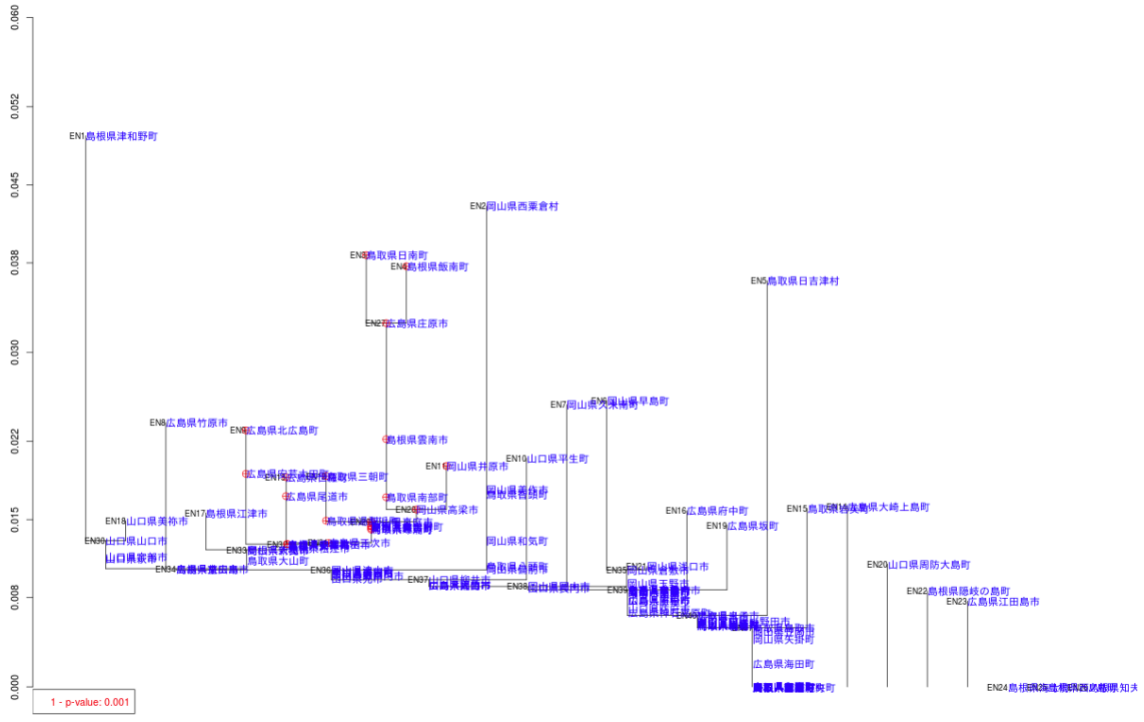


Figure 5.12: Echelon dendrogram when the age group is not considered ($K = 1$).

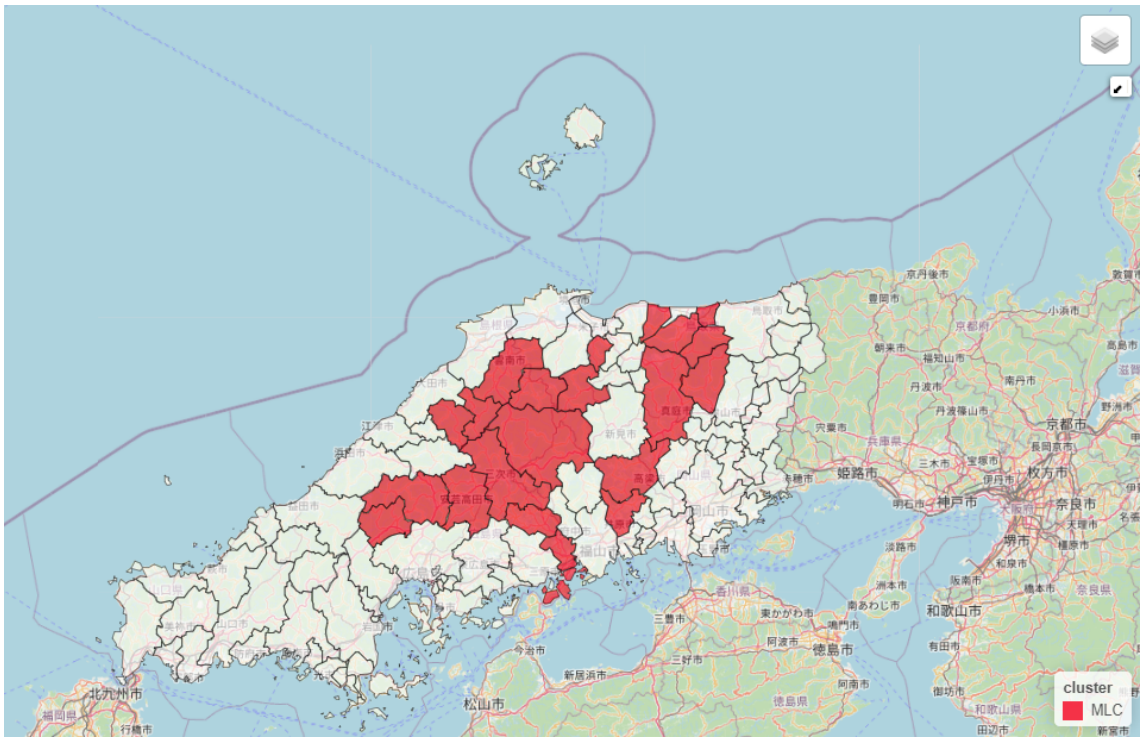


Figure 5.13: Map corresponding to the detected hotspot cluster area when the age group is not considered ($K = 1$).

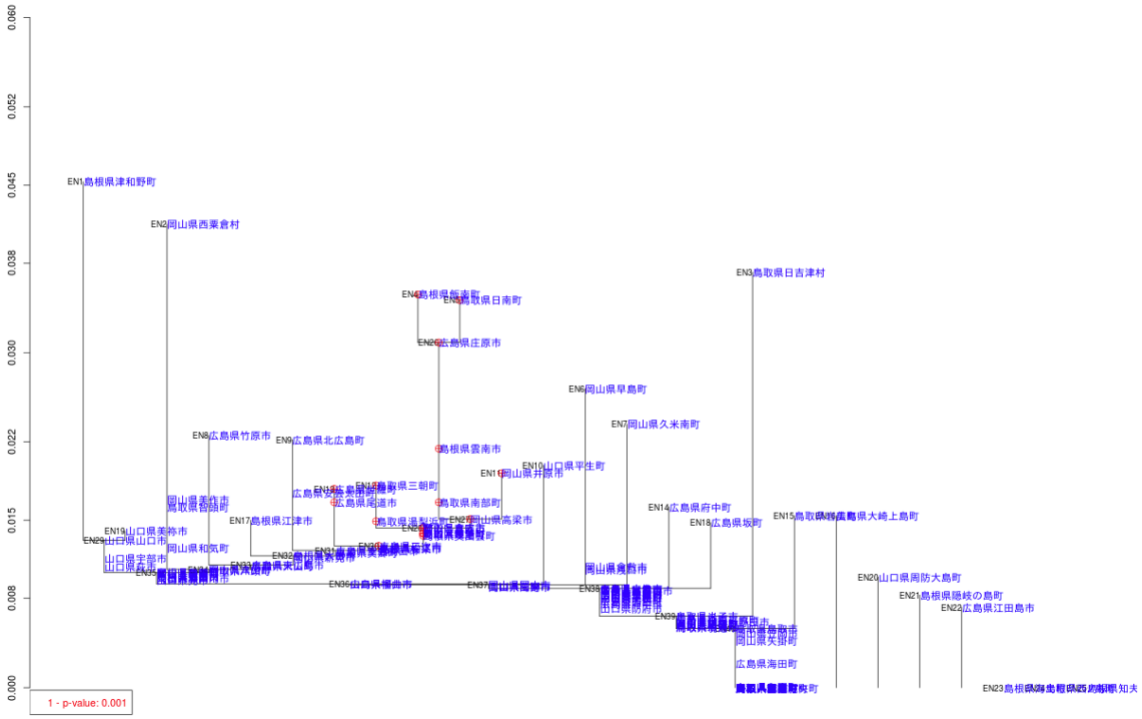


Figure 5.14: Echelon dendrogram when the age group is considered ($K = 8$).

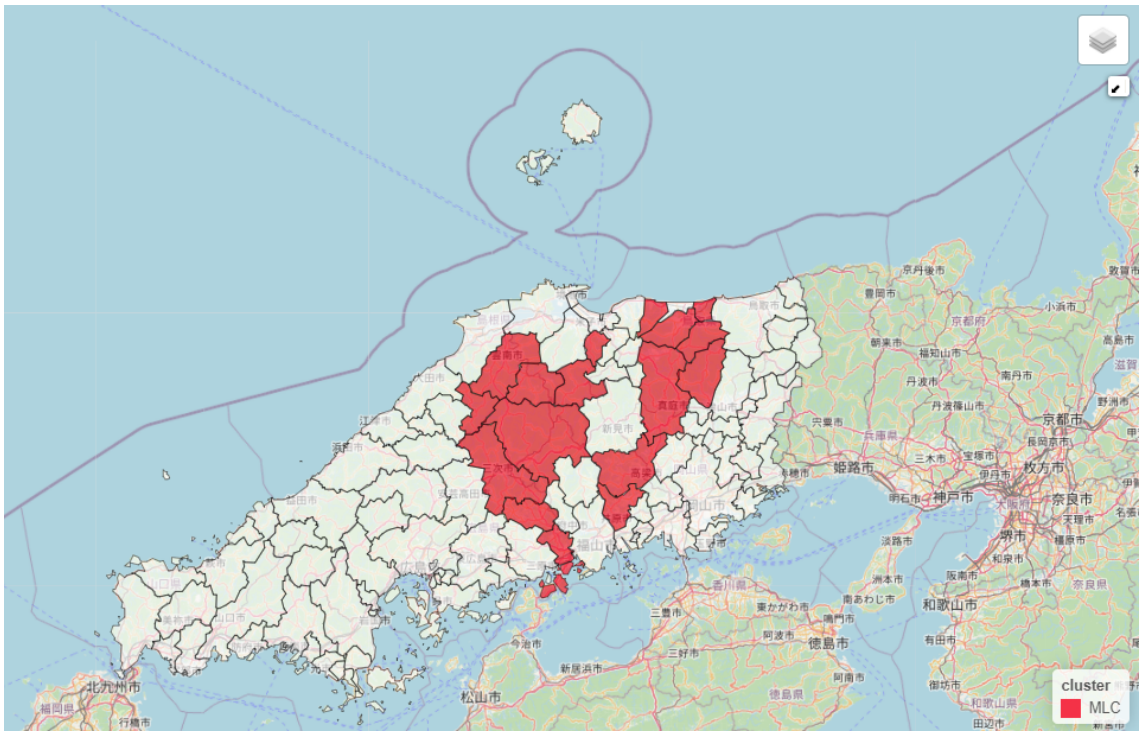


Figure 5.15: Map corresponding to the detected hotspot cluster area when the age group is considered ($K = 8$).

6 Assessment of dendrogram complexity

6.1 Patterning the Echelon dendrogram

The Echelon dendrogram generated by the Echelon analysis is a graph that accurately represents the structure of spatial data. By visually expressing the structure of data, various information possessed by spatial data can be obtained and features can be found. However, when trying to compare data structures between regions and eras, it is not possible to evaluate them because there are no clear evaluation criteria. Moreover, as the number of regions to be analyzed increases, the shape of the dendrogram becomes complicated, and it is very difficult to compare them. This is because the number of regions, neighborhood information, and the values of each region differ depending on the data to be analyzed, so that the positional relationship and height of the dendrogram peak and foundation also change. That is, it is extremely rare that a dendrogram having exactly the same shape is formed, and it is not possible to evaluate the data structure with clear evaluation criteria.

Therefore, we "pattern" the Echelon dendrogram according to certain rules and generalize the shape of the dendrogram without compromising the characteristics of the dendrogram. Dendrogram patterning has been proposed as an approach for the simplicity and complexity of data structures (Kurihara and Ishioka, 2007). The patterning is determined by finding the hierarchical structure of the Echelon dendrogram and using indicators such as the total number of layers and the total number of peaks of the shape pattern. For example, the three dendrograms shown in Figure 6.1 are "patterned" into the shape shown in Figure 6.2. The reason why these three dendrograms are unified into one shape is that when viewed from the root echelon, the first foundation has two branches, one of which forms a peak. On the other hand, the other is because the dendrogram is formed by repeating bifurcation at the next foundation again. A "patterned" dendrogram is represented by aligning the heights of all peaks and keeping the spacing between each foundation constant. Next, the flow of dendrogram patterning will be introduced using an example of 3 by 3 mesh data as shown in Figure 6.3.

Consider a situation where the data is randomly given to a 3 by 3 mesh so that values 1 to 9 do not overlap. Where, there are 11 possible patterns of dendrograms, as shown in Figure 6.4. Five indicators are defined for the patterning of the Echelon dendrogram (Kurihara and Ishioka, 2007), and the characteristics of the dendrogram are extracted using these indicators. There are five indicators: NE, NP, MF, MP, and LU. The details of the five indicators are as follows.

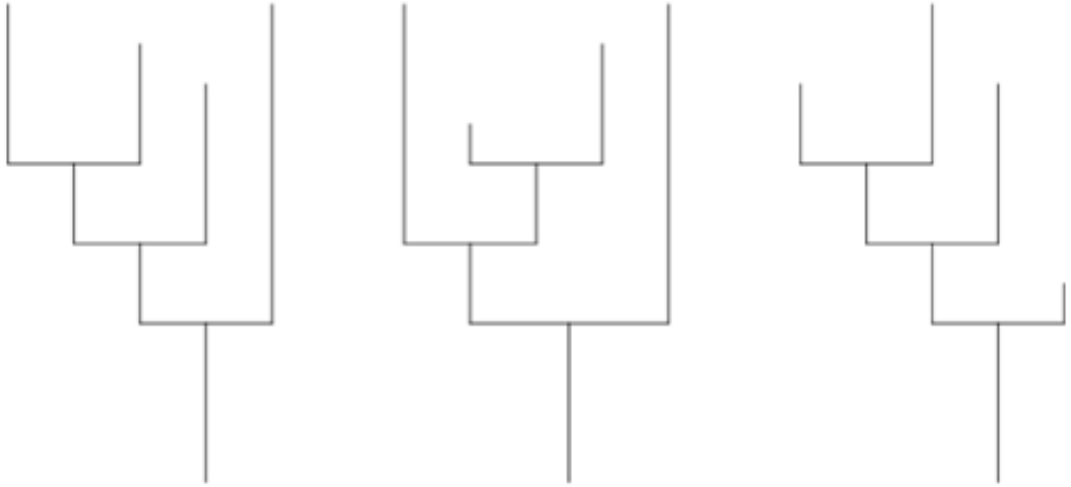


Figure 6.1: Three dendrograms with different shapes.

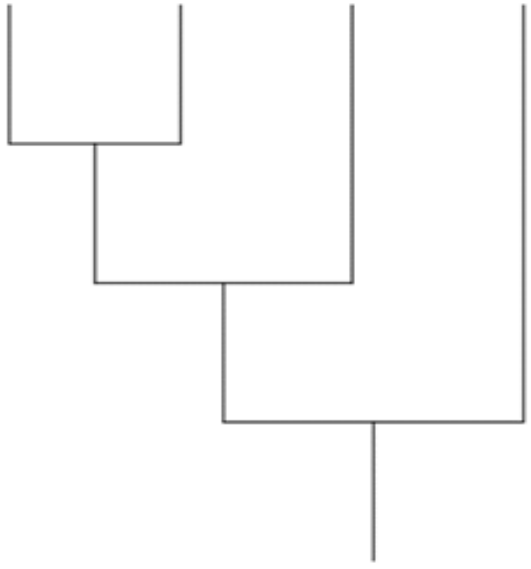


Figure 6.2: Dendrogram resulting from patterning the three dendrograms in Figure 6.1.

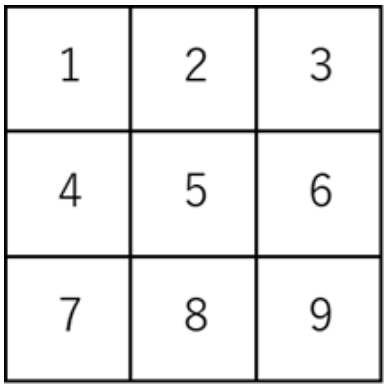


Figure 6.3: 3 by 3 mesh data.

- NE: Number of Echelons
- NP: Number of Peaks
- MF: Largest clan excluding root
- MP: Maximum number of Children
- LU: Total Peaks − Number of Peaks with Common Parents + Number of Parents (Parents with all their own children peaking)

The five indicators in the case of Figure 6.2 are $NE = 7$, $NP = 4$, $MF = 5$, $MP = 2$, and $LU = 3$. In addition, the five indicators of the dendrogram of the 11 patterns in Figure 6.4 are shown in Table 6.1. The five indicators were defined to discriminate the dendrogram, but these indicators were defined only for the 3 by 3 mesh data mentioned above. Therefore, it was confirmed that when the number of regions increases and the dendrogram becomes more complicated than this, it becomes difficult to distinguish with the five indicators. For example, it can be seen that the shape of the dendrogram pattern is different even though the five indicators of the four patterned dendrograms shown in Figure 6.5 is exactly the same. The patterns of these dendrograms are $NE = 12$, $NP = 7$, $MF = 10$, $MP = 3$, $LU = 5$, and the five indicators are exactly the same, so it is not possible to judge the difference in shape. Therefore, in this study, LV was defined as the sixth index. LV is the number given to each layer of the Echelon dendrogram multiplied by the number of peaks. For example, in the case of Figure 6.2, $LV = 3\times2+2\times1+1\times1 = 9$. Using this LV, as shown in Table 6.2, it can be seen that it was possible to discriminate the patterns of the four dendrograms that could not be discriminated by the five indexes.

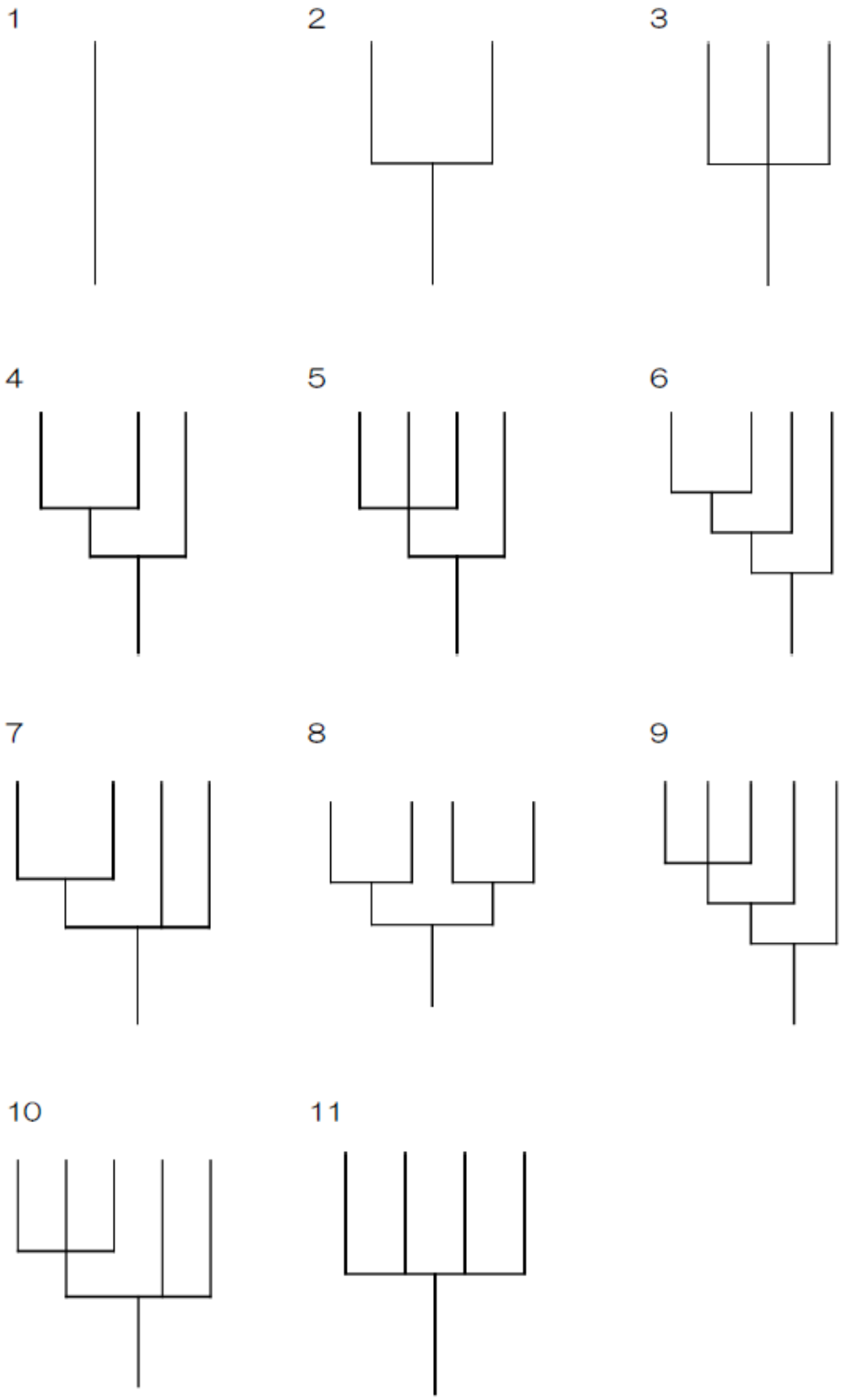


Figure 6.4: 11 dendrogram patterns generated from 3 by 3 mesh data.

Table 6.1: 5 indicators for each of the 11 patterns of dendrograms.

	NE	NP	MF	MP	LU
Pattern 1	1	1	0	1	1
Pattern 2	3	2	1	2	1
Pattern 3	4	3	1	3	1
Pattern 4	5	3	3	2	2
Pattern 5	6	4	4	3	2
Pattern 6	7	4	5	2	3
Pattern 7	6	4	3	3	3
Pattern 8	7	4	3	2	2
Pattern 9	8	5	6	3	3
Pattern 10	7	5	4	3	3
Pattern 11	5	4	1	4	1

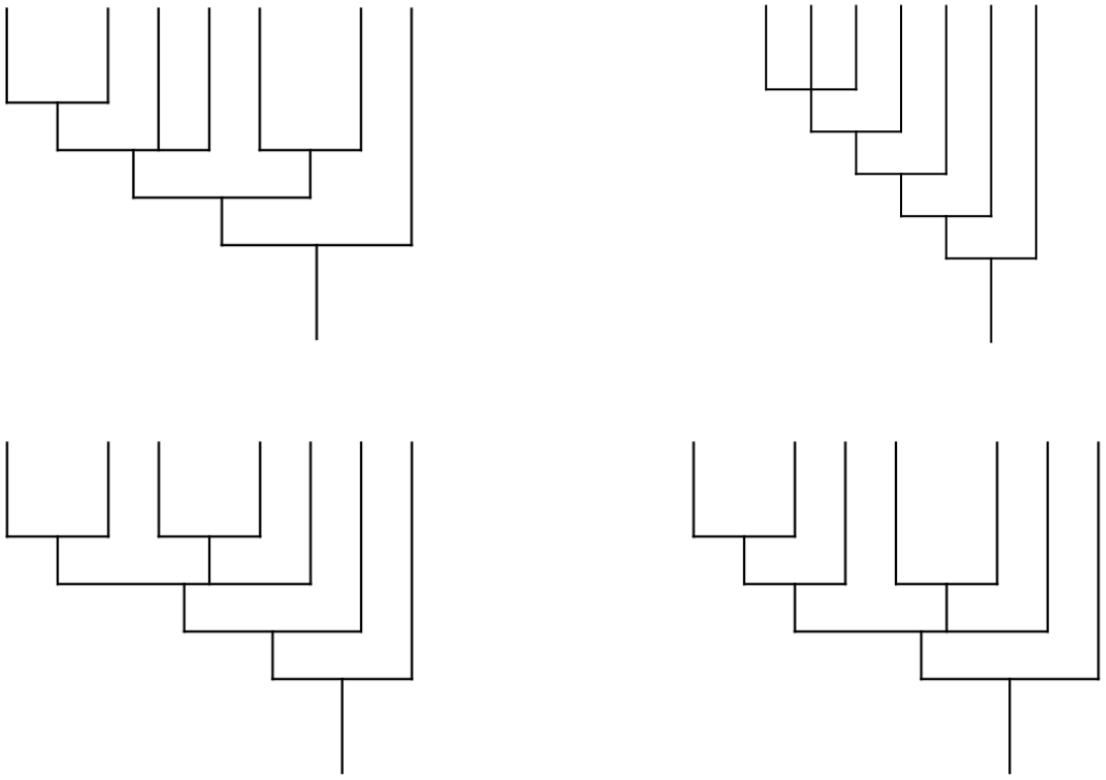


Figure 6.5: Dendrogram pattern with the same 5 values.

Table 6.2: 6 indicators of 4 dendrograms.

	NE	NP	MF	MP	LU	LV
Top left of Figure 6.5	12	7	10	3	5	21
Top right of Figure 6.5	12	7	10	3	5	25
Bottom left of Figure 6.5	12	7	10	3	5	22
Bottom right of Figure 6.5	12	7	10	3	5	20

6.2 Echelon tree

This section describes a method for evaluating complexity by focusing on the hierarchical structure of dendrograms and their relationships. Although it is possible to classify dendrograms using the six indicators calculated with the dendrogram patterning introduced in section 6.1, there is no clear indicator for assessing the complexity of dendrograms. Therefore, in considering the complexity of the dendrogram, we use the idea of the Echelon tree (Kurihara *et al.*, 2000). As shown in Figure 6.6 to Figure 6.8, the Echelon tree focuses only on the relationship between each node of the dendrogram and expresses only the relationship of the spatial data structure. In addition, the Echelon

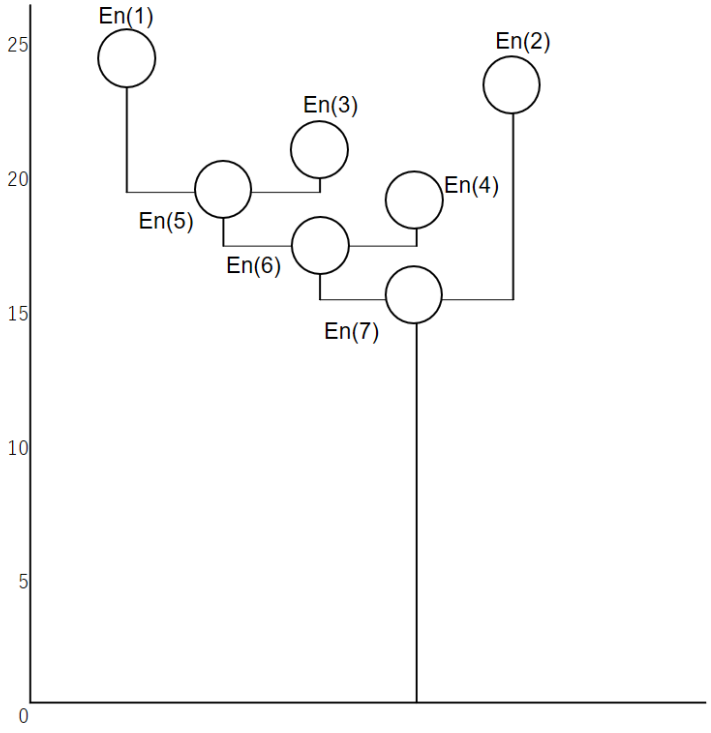


Figure 6.6: Flow of Echelon tree creation (1).

tree is polarized into a "Vine tree" and a "Binary tree" as shown in Figure 6.9 and Figure 6.10, and each tree has the following characteristics.

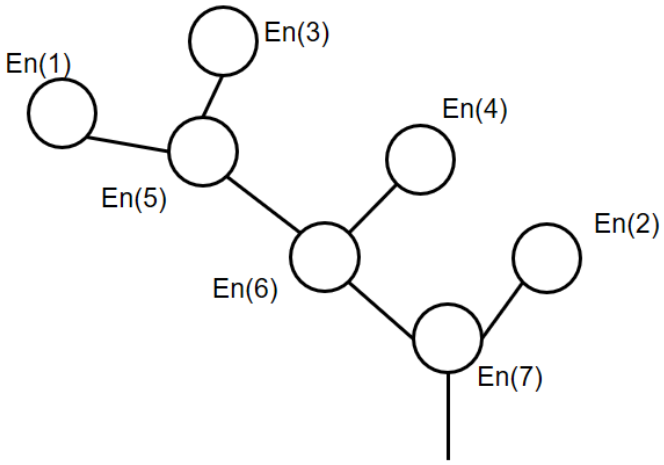


Figure 6.7: Flow of Echelon tree creation (2).

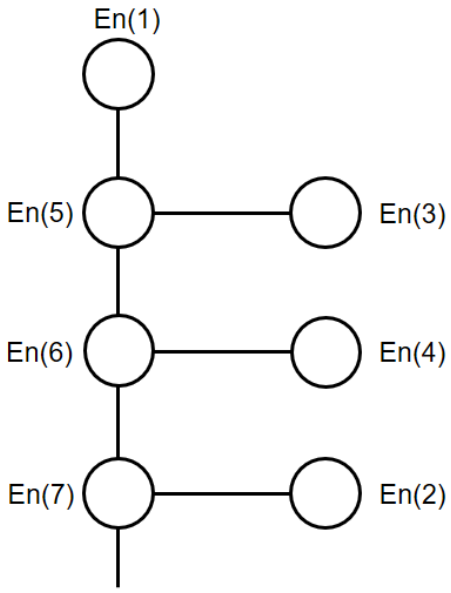


Figure 6.8: Flow of Echelon tree creation (3).

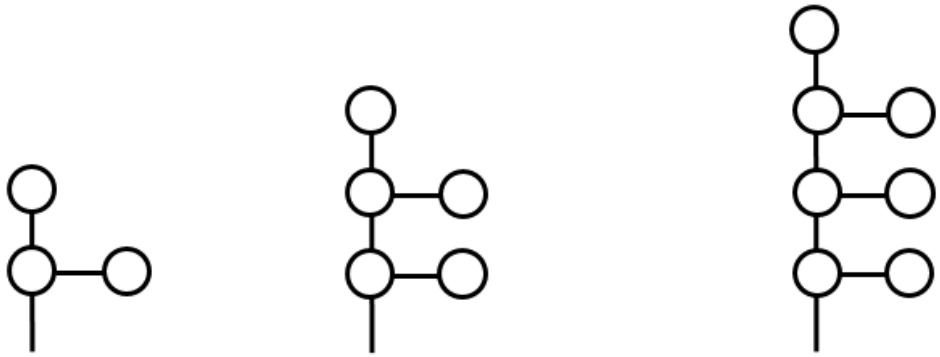


Figure 6.9: Vine tree.

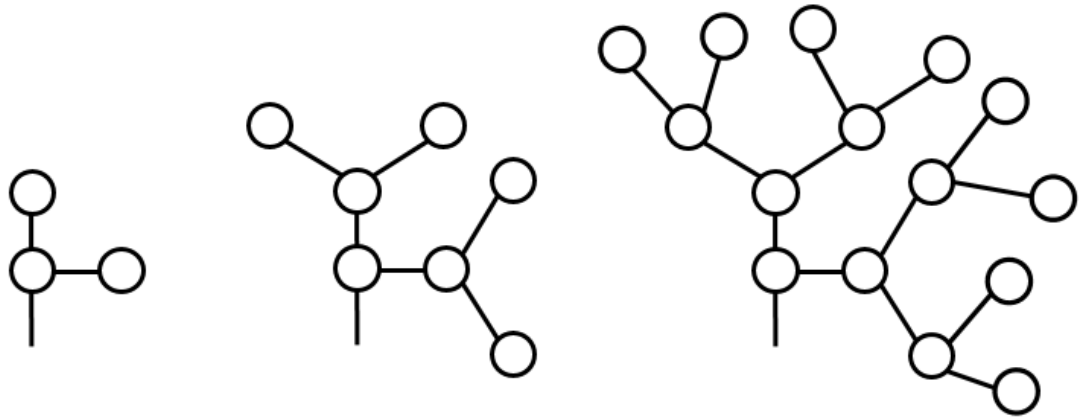


Figure 6.10: Binary tree.

- Vine tree ... There are multiple peaks between the 1st peak and the root.
- Binary tree ... There is a mountain containing multiple peaks.

Echelon profiles are defined as a structural analysis of the Echelon tree (Kurihara *et al.*, 2000). The Echelon profiles are obtained by the following method by decomposing the tree based on the pruning process in pattern recognition. First, as a pretreatment, "limb" is extracted from the Echelon tree. Count the number of nodes from all peaks to the root, and let the node group with the largest number of common nodes be the 1st limb.

1. Let the node generated from the 1st "limb" be "bough".
2. Decompose each "bough" into "limb" and "bough" in the same way as 1. A 2nd limb and a new bough are generated from the bough.
3. After that, repeat until no bough is generated. The number of times limb is detected is defined as cycle. The larger the number of cycles, the more complicated the spatial data structure.

In addition, the following four scales are defined in the Echelon tree.

$$\mathrm{Di}(i) = \frac{\sum \mathrm{limb}(i)}{\text{total nodes}}$$
$$\mathrm{Sc}(i) = \frac{\sum \text{cells of limb}(i)}{\text{total cells}}$$
$$\mathrm{Bu}(i) = \frac{\mathrm{bough}(i)}{\text{total peaks}}$$
$$\mathrm{St}(i) = \frac{\sum \text{peaks of limb}(i)}{\text{total peaks}}$$

Divergence (Di)

Represents the proportion of limb nodes in each cycle. The lower this rate, the more stable the number of limb nodes.

Scope (Sc)

Represents the percentage of cells in the limb of each cycle. It indicates whether the surface complexity is concentrated in a particular sector.

Bunching (Bu)

Represents the ratio of bough to all peaks in each cycle. The lower this rate, the closer to the binary tree.

Stacking (St)

Represents the percentage of peaks in the limb of each cycle. Echelons that share a common foundation indicate whether they tend to have the same degree.

In addition, the four scales are calculated for each cycle according to the flow shown below.

STEP1

Divide the tree into limb and bough based on the Pruning process.

STEP2

Calculate four scales using limb and bough.

Figure 6.11 and Figure 6.12 shows that the vine tree and binary tree in Figure 6.9 and Figure 6.10 are divided into limb and bough in each cycle, and four scales are calculated.

Figure 6.13 shows an example of calculating the four scales of Echelon profiles and

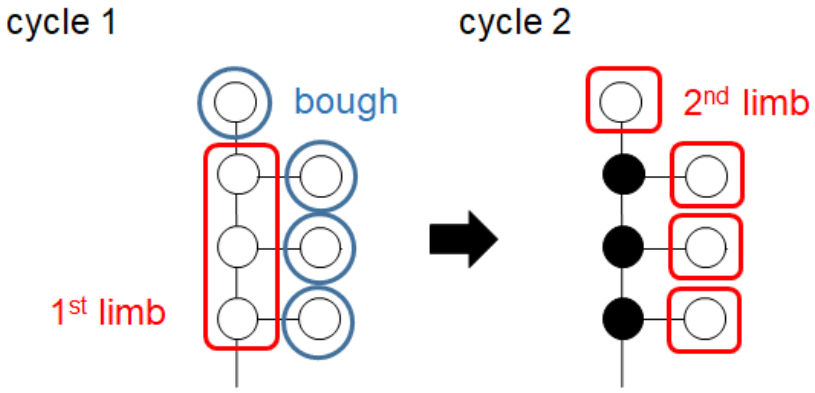


Figure 6.11: Progress of Cycle in Vine tree.

showing their values in a graph. From this graph, it is possible to grasp the state of the dendrogram at each cycle. In Cycle 1, Scope is about 0.6 and Bunching is about 0.45, so a dendrogram including about 60% of the entire analysis target area is formed, and it can be seen that there is a tendency of a binary tree type. The value of Echelon profiles will eventually be 1 for Divergence, Scope, Stacking and 0 for Bunching. In Cycle 2, Divergence is about 0.58, so we can see that the limb node exceeds 50% of the total. Also, since Scope is about 0.8, it can be seen that the dendrogram includes 80% of the entire region. Furthermore, since Stacking is about 0.4, it can be seen that nearly 40% of all peaks are included. After that, the number of Cycles increased, and in Cycle 4, which is one before the last cycle, the values are almost the same as Cycle 5, and it can be seen that the dendrogram can be almost explained. In Cycle 5, Divergence, Scope,

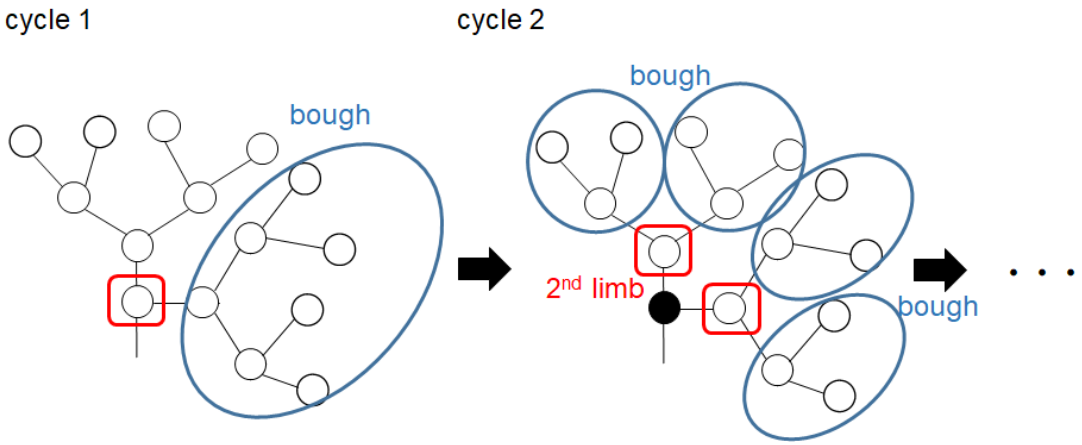


Figure 6.12: Progress of Cycle in Binary tree.

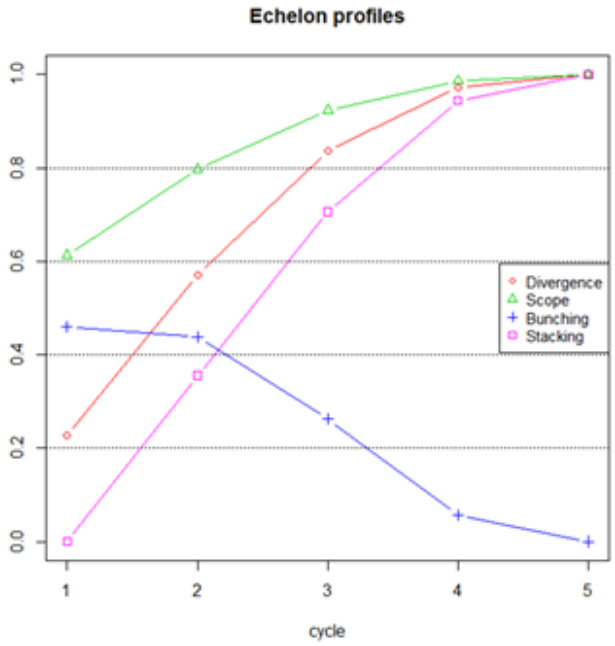


Figure 6.13: Calculation results and their graphs for four scales.

and Stacking are 1 and Bunching is 0. It has been proposed to use the number of cycles to represent the complexity of the Echelon tree (Kurihara *et al.*, 2000). Since the number of cycles represents the number of detections of limb and bough in the Echelon tree, it can be an index expressing the complexity of the dendrogram. However, it is considered that there is some complexity that cannot be evaluated only by the number of cycles. For example, in the case of two tree as shown in Figure 6.14.

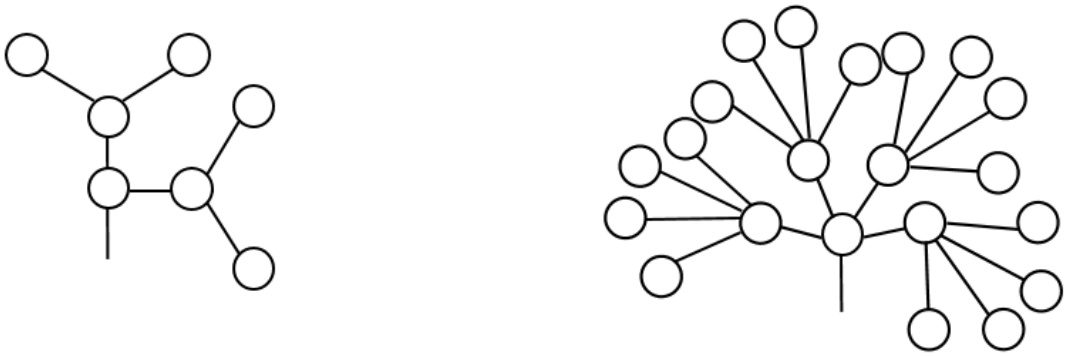


Figure 6.14: Echelon tree with the same cycle.

When trying to compare Figure 6.14 left and right, Figure 6.14 right seems to be more complicated than left, but from the viewpoint of the number of cycles, left and right are both have 3 cycles. Assessing the complexity of a dendrogram using the number of cycles may differ from the apparent complexity. Although the number of cycles is one index showing complexity, it cannot be said to be sufficient as an evaluation index because of the problems mentioned above. Therefore, in section 6.3, we define the concept of "stage" of the dendrogram, and consider an index that can evaluate the complexity of the dendrogram pattern by considering it together with the information on the number of cycles.

6.3 Stage of dendrogram

In this section, we define the concept of "stage" to evaluate the complexity of the dendrogram. In the 4 by 4 mesh data, if you try to arrange the numbers 1 to 16 so that they do not overlap, it will be as $16!(= 2.092279e + 13)$, and there are a huge number of combinations. Since it is unrealistic to generate all the data, this time we will randomly generate 1 million kinds of data from this combination of data. Echelon analysis was performed on the generated data to create a dendrogram. When the patterns of this dendrogram were examined, 181 patterns were detected as shown in Table 6.3, and 6

indexes were calculated. Next, the results of principal component analysis are shown in Figure 6.15. In this figure, the horizontal axis represents the first principal component and the vertical axis represents the second principal component, and the principal component scores are plotted. The cumulative contribution rate up to the second principal

Table 6.3: Value of 6 indicators of 181 generated dendrograms.

	NE	NP	MF	MP	LU	LV
1	11	7	9	4	4	22
2	5	3	3	2	2	5
3	9	5	7	2	4	14
4	7	4	3	2	2	8
5	7	4	5	2	3	9
⋮			⋮			
179	13	8	10	3	6	26
180	13	8	6	3	6	20
181	12	7	10	3	4	19

component is 86.98%. Looking at this result, it can be seen that it is roughly divided into four groups, it is expected that the MP value has a large effect on the second principal component on the vertical axis. In addition, it can be seen that the horizontal axis of the first principal component is influenced by the number of peaks in dendrograms such as NE and NP and the number of echelons. Furthermore, when compared with the results of the cluster analysis in Figure 6.16, it can be seen that the clustering is performed as shown in Figure 6.17. Here, we focus on the dendrogram pattern of the MP = 4 group. When the patterns of these dendrograms were confirmed in detail, it was found that the shape of the dendrogram tended to change as shown in Figure 6.18. Looking at the shape of the dendrogram pattern, a simple dendrogram is placed on the left side, and it seems that the shape becomes more complicated as it goes to the right side in order. When we confirmed the six patterned indicators for these dendrograms, we found that if the number of NP and NE was large, as shown in Figure 6.18, they tended to be placed on the right side of the figure. From these results, it is considered that the complexity of the patterned dendrogram is greatly influenced by the value of NE, so we consider an index focusing on the size of NE.

In this paper, we define the "stage" of the dendrogram from the viewpoint of the growth of the data structure considering the increase and decrease of NE. The stage focuses on the growth process of the dendrogram, and the stage goes up as it becomes more complicated. Specifically, the stage is determined according to the NP (number of peaks). The stage progresses in one of two ways as shown in Figure 6.19 and Figure 6.20.

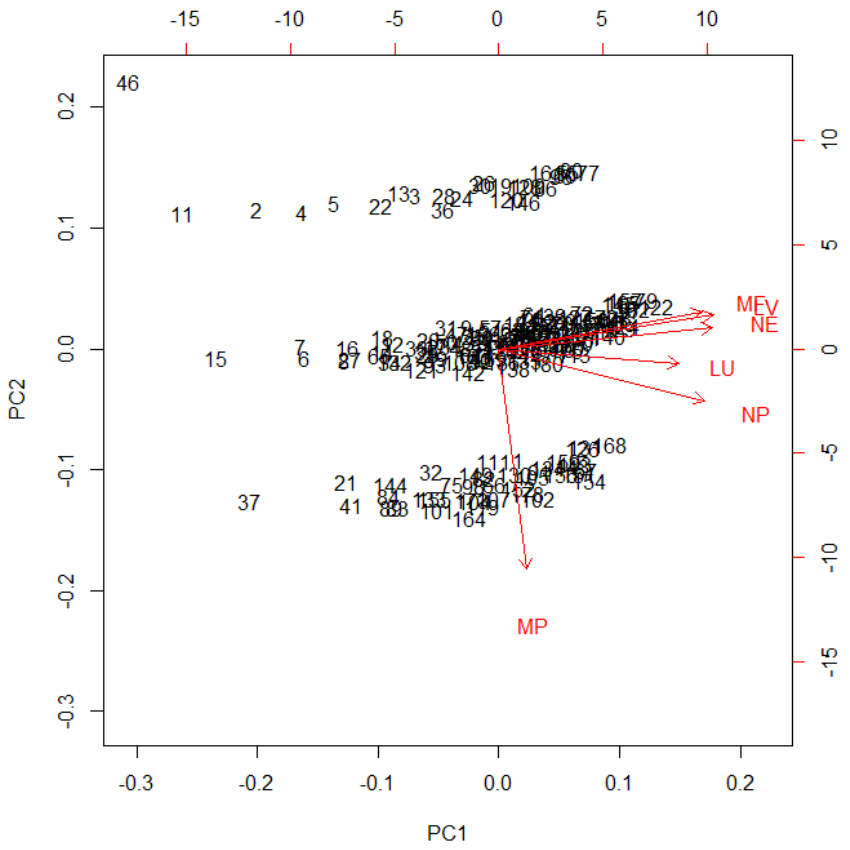


Figure 6.15: Results of principal component analysis.

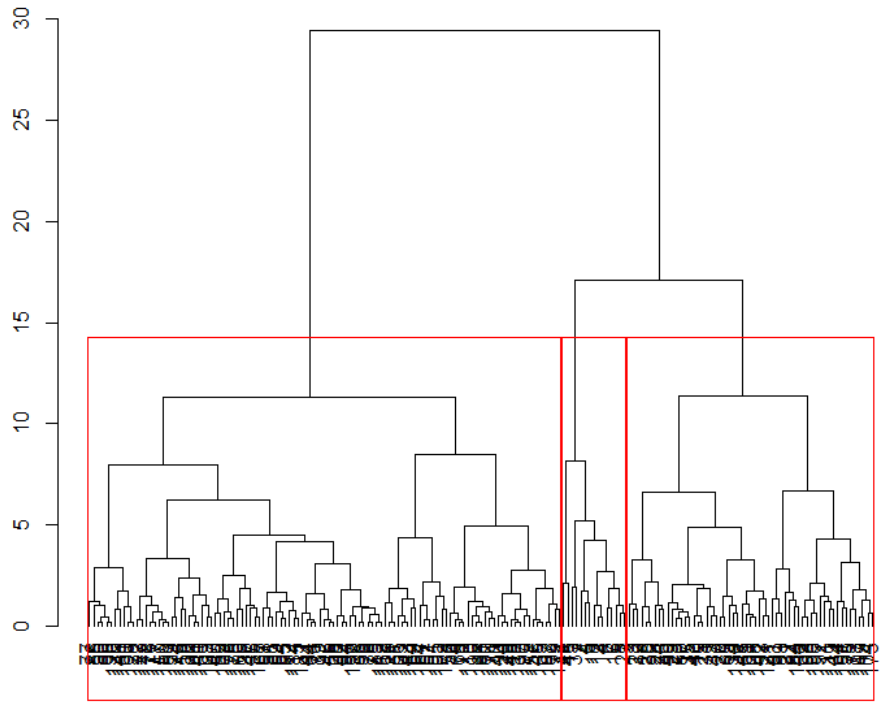


Figure 6.16: Results of cluster analysis.

1. The area that constitutes the peak becomes the foundation, and another peak appears.

2. Another peak appears from the common foundation.

The stage I dendrogram is the simplest shape with NE = 1, but it becomes more complicated as the stage progresses. Figure 6.21 shows how the stage progresses comprehensively from the stage I dendrogram. By defining the stage, it is possible to evaluate the dendrogram with the same number of cycles, and even within the same stage, it is possible to evaluate the complexity from the difference in the number of cycles. Dendrogram "stage" was defined to evaluate data whose data structure changes over time. For example, assume data that changes over time, such as population, population ratio, pollutant concentration, and plant reproduction. The purpose of analysis using "stage" is to evaluate whether or not the structure of data has changed to a complex one. In the next chapter, we will introduce an example in which the spatial data structure changes with time using actual data, and consider the results.

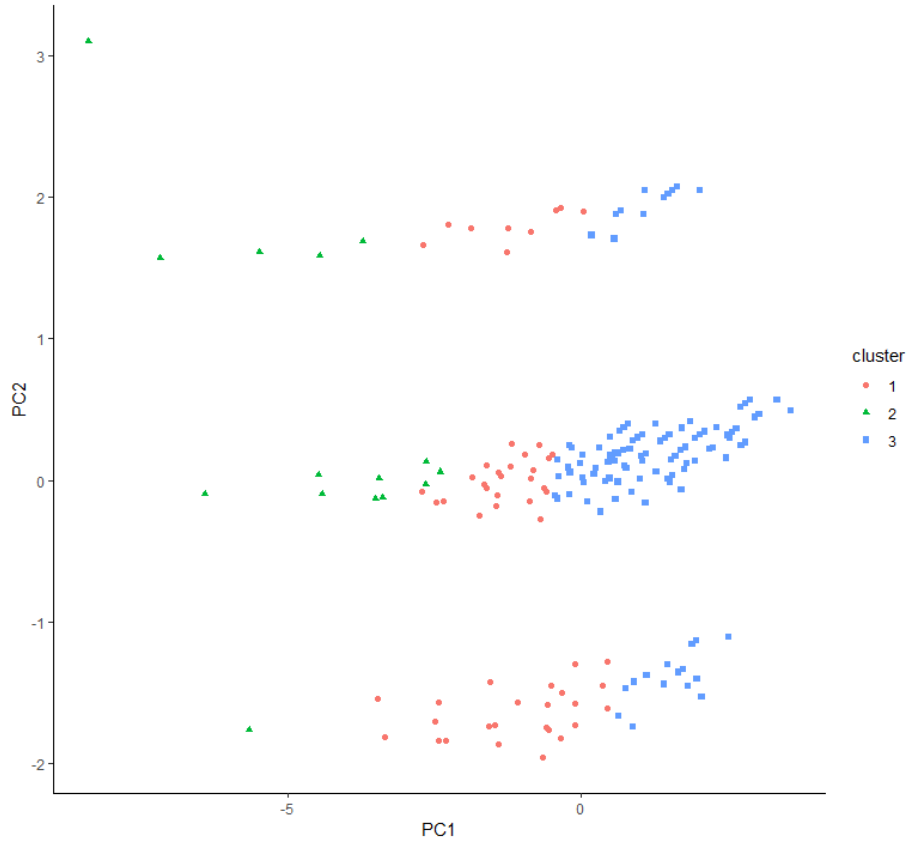


Figure 6.17: Results of principal component analysis corresponding to the results of cluster analysis.

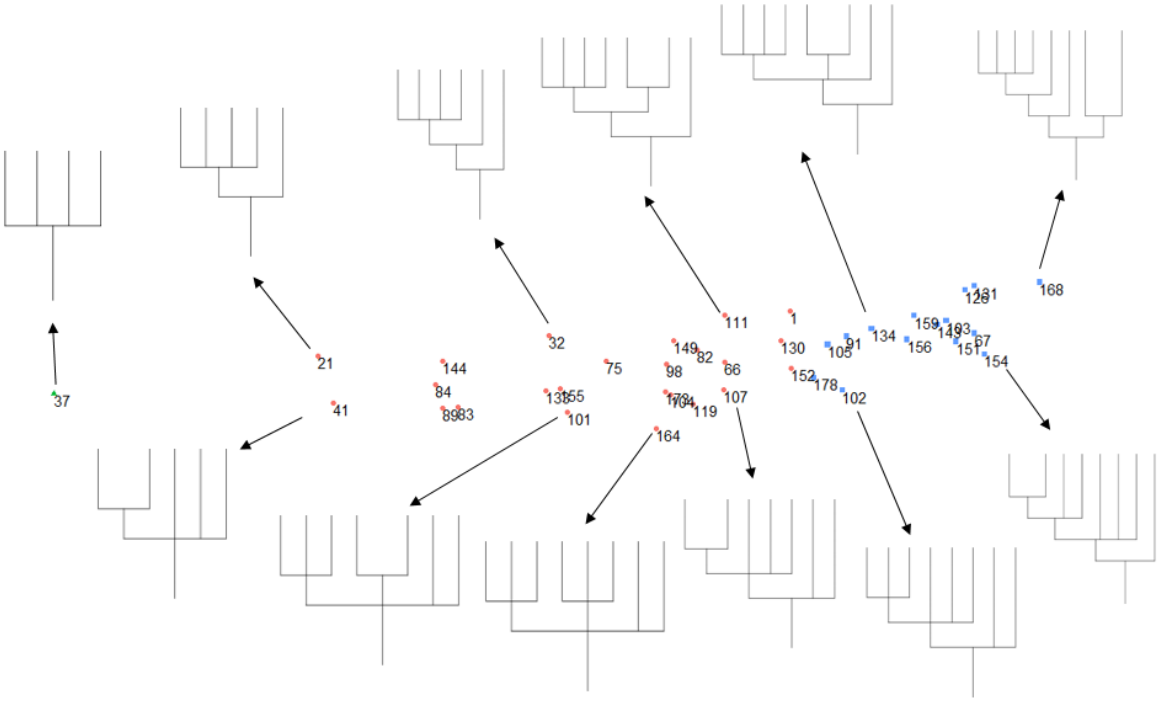


Figure 6.18: Breakdown of dendrogram with MP = 4.

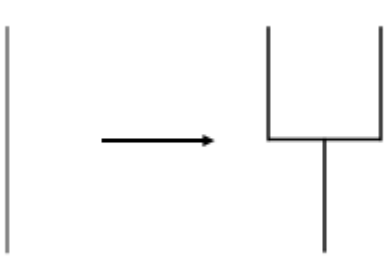


Figure 6.19: Stage progress pattern (1).

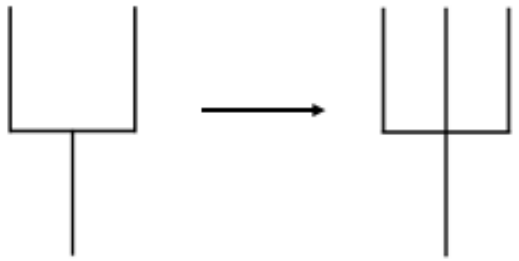


Figure 6.20: Stage progress pattern (2).

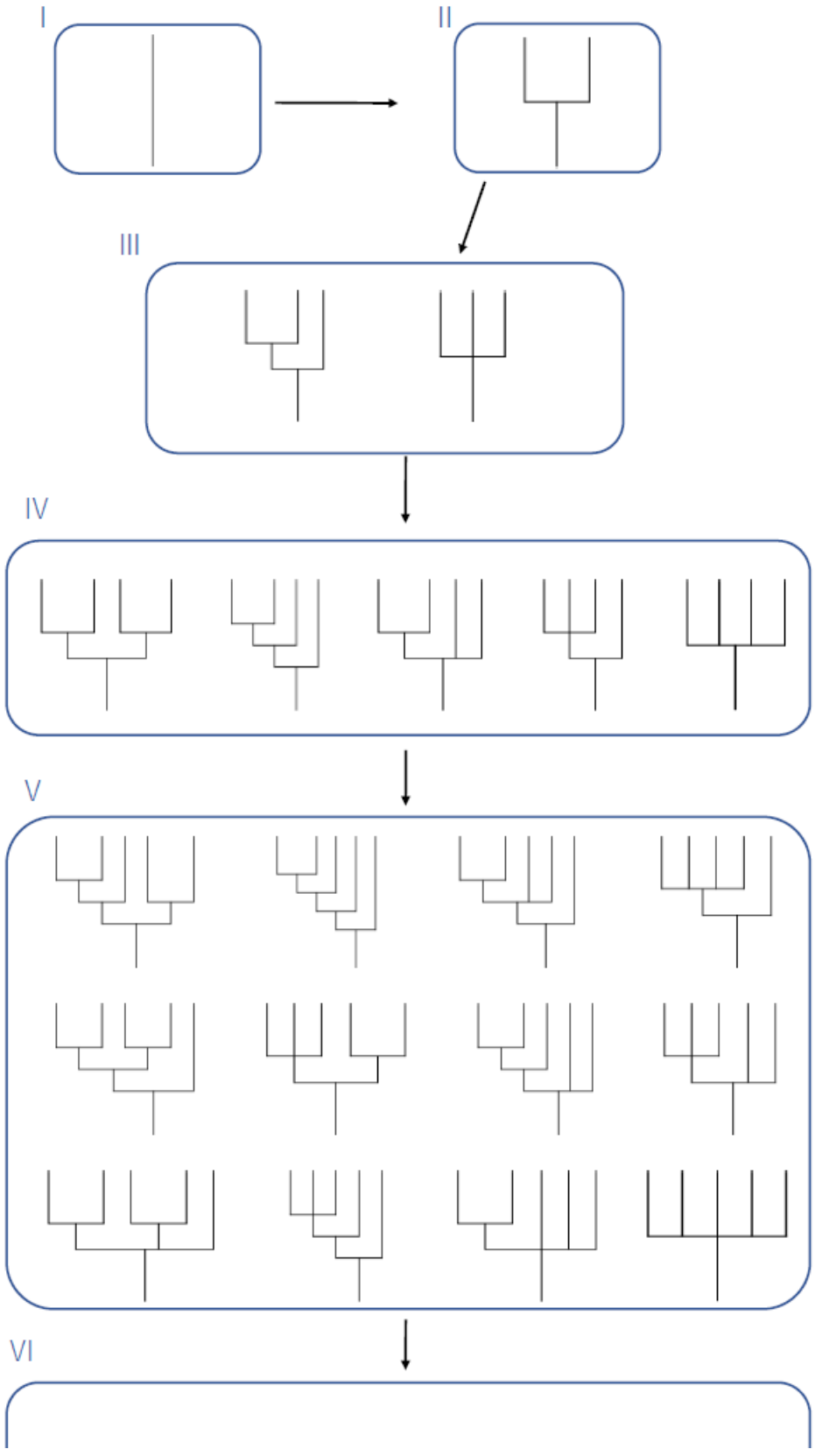


Figure 6.21: The figure which comprehensively expressed the stage progress of a dendrogram (up to stage 5).

7 Analysis example with actual data

In this chapter, we evaluate the complexity of the spatial structure of actual data using the "staged" evaluation criteria defined in section 6.3 and consider its usefulness. The data used are data on the proportion of the population aged 65 and over in the 23 wards of Tokyo from 1985 to 2018. The results obtained from this data are shown in Figures 7.1 to 7.12, and the contents of each figure are as follows. The figure on the upper left is a color-coded map in descending order of the proportion of the population aged 65 and over. The figure on the upper right is the dendrogram generated as a result of the Echelon analysis. The bottom left figure shows the shape of the patterned dendrogram. The figure at the bottom right shows a color-coded map for each dendrogram pattern hierarchy.

In 1985, it can be seen that areas with high values tend to be concentrated near the center of Tokyo's 23 wards. On the other hand, at first, areas with high values tended to concentrate in the center of the 23 wards, but as time passed, it can be seen that areas with high values were shifting to the outside of the areas. From these results, it can be seen that the proportion of the population aged 65 and over in the 23 wards of Tokyo has changed from a structure concentrated in the city center to a structure concentrated outside the 23 wards as time changes. In addition, the pattern of the Echelon dendrogram is simple because the number of regions is not large, but the shape of the dendrogram has changed over time in stages I to Ⅳ. From 1994 to 2000, it is the simplest stage I, and it can be seen that the data structure has changed since this period. The first stage and the last stage are both dendrograms of Ⅲ, but the order of the regions that make up them is very different. You can see this clearly in the map (bottom right) painted for each layer of the Echelon dendrogram.

Looking at the map painted for each dendrogram level, the tendency to move from near the center to the outside with time is similar to the map painted according to the proportion of the population. However, from the point of view of the data hierarchy, it can be seen that there is a different tendency. Around 2018, the map of the population ratio does not show a very high value in the southern region, but from the viewpoint of the hierarchical structure of the data, it can be seen that the southern region of Tokyo's 23 wards is in a high hierarchy. From this, it is currently showing a low value, but since it is a peak, it may become a high value in the future. Therefore, it will be necessary to carefully observe the progress in the future. In addition, it is thought that it will lead to grasping the regional characteristics that areas with high values or areas that are expected to increase in the future are waiting, and to find an opportunity to take measures from that tendency.

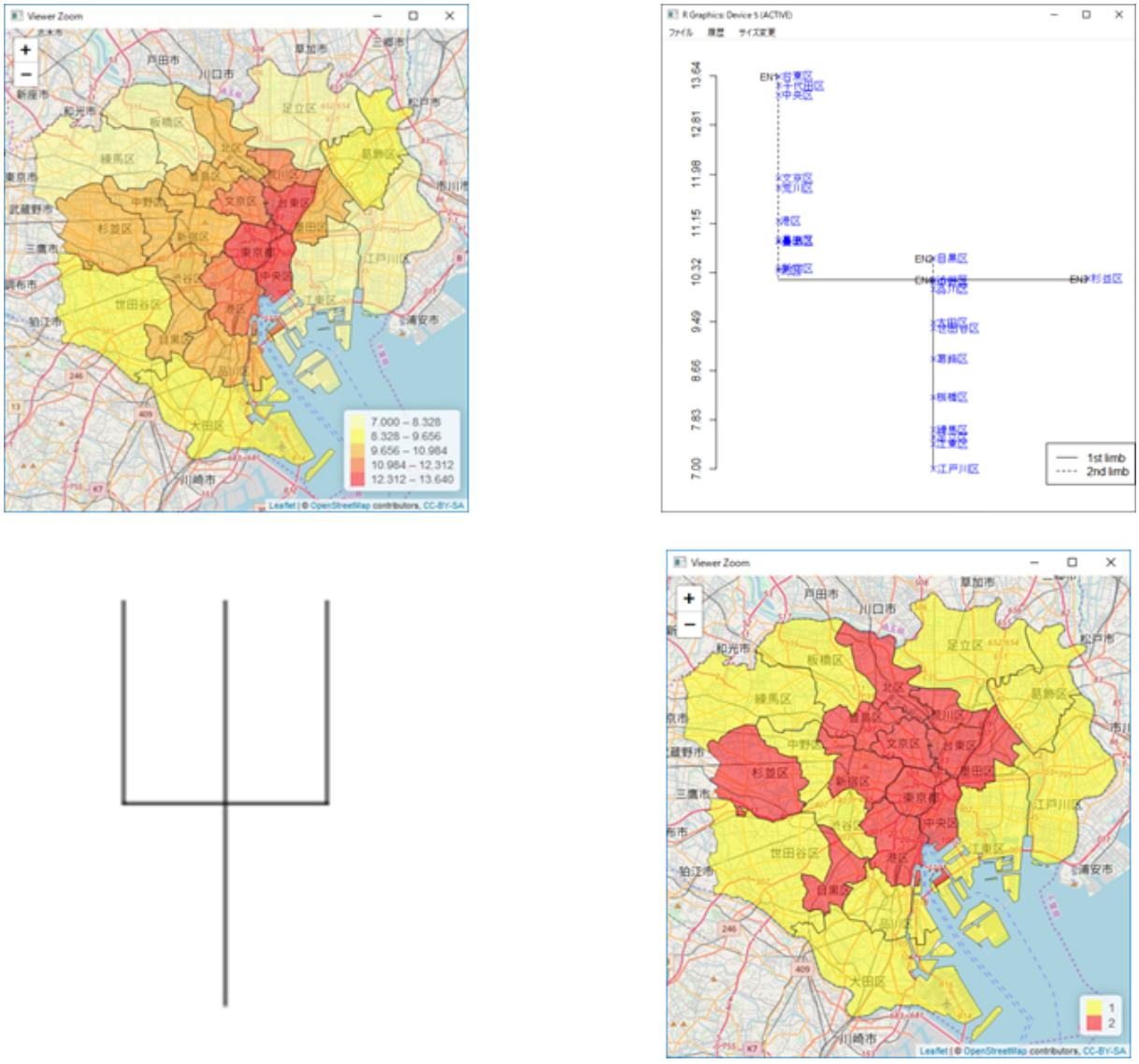
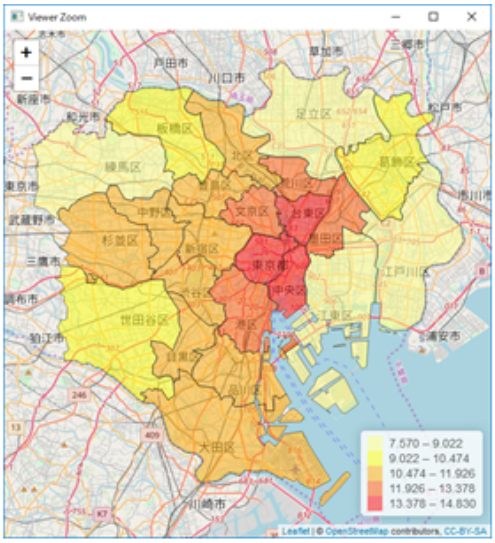
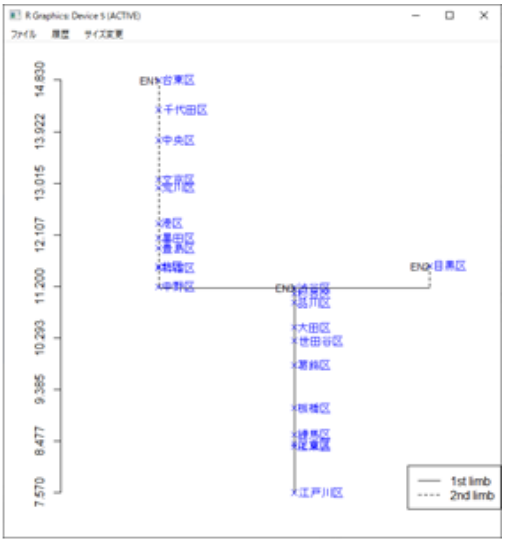
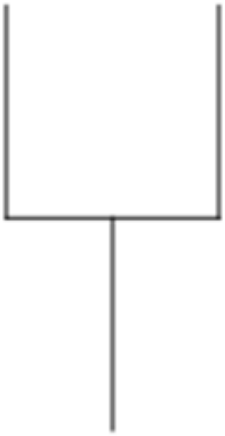


Figure 7.1: Analysis results (using 1985 data).







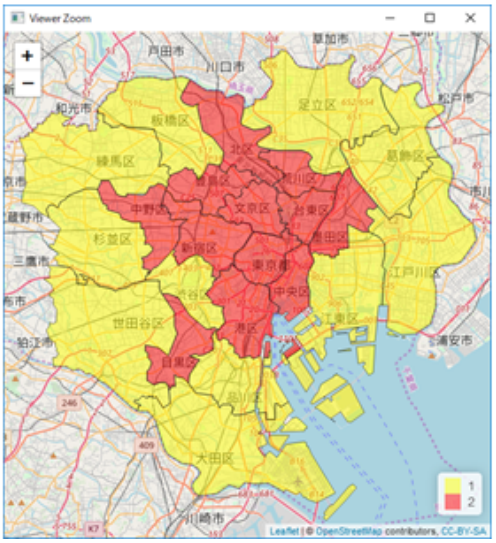
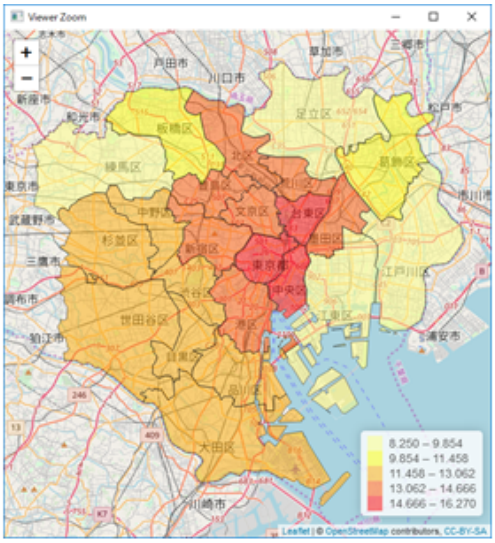
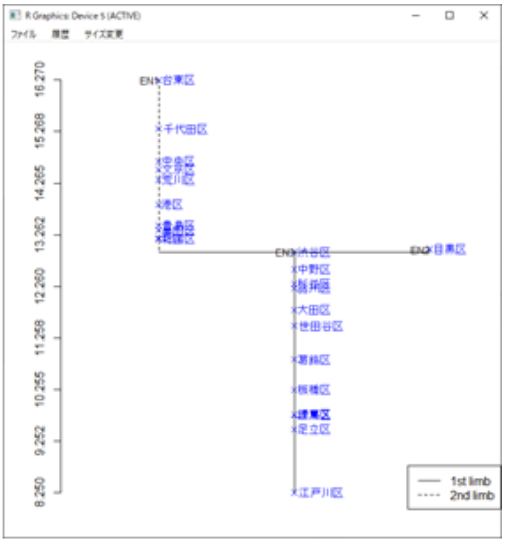
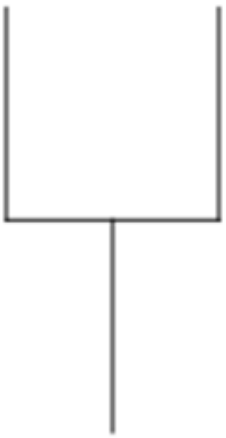


Figure 7.2: Analysis results (using 1988 data).







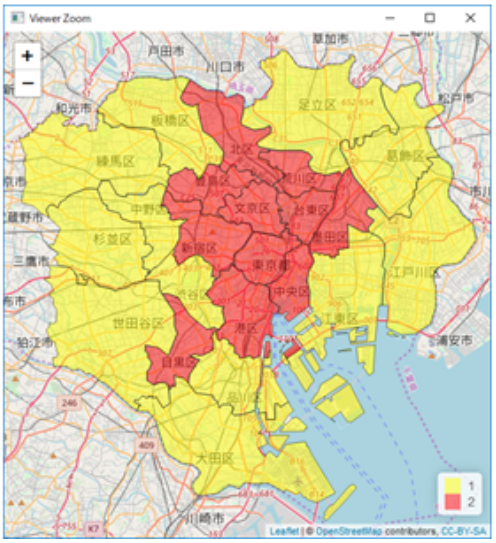


Figure 7.3: Analysis results (using 1991 data).

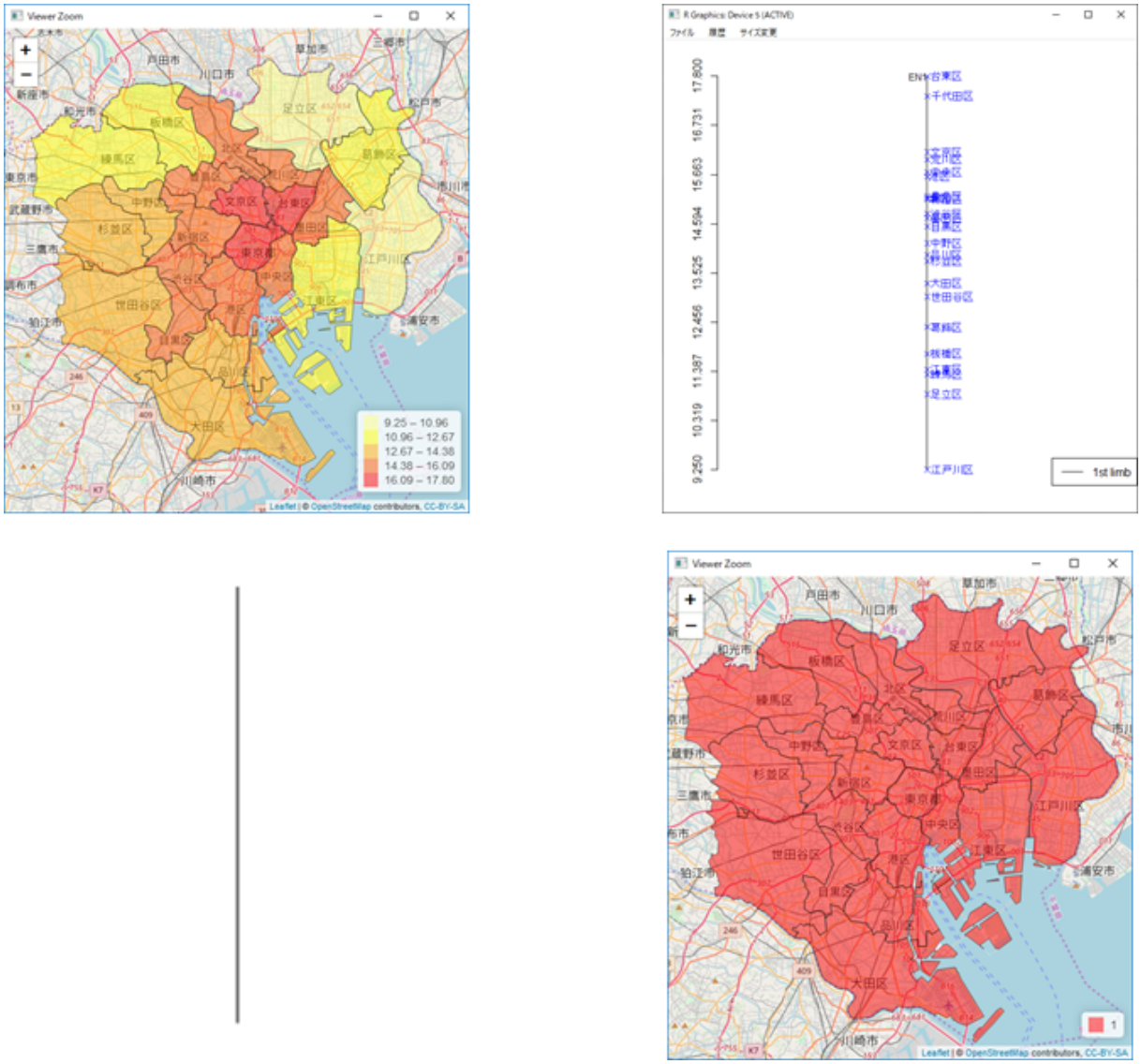


Figure 7.4: Analysis results (using 1994 data).

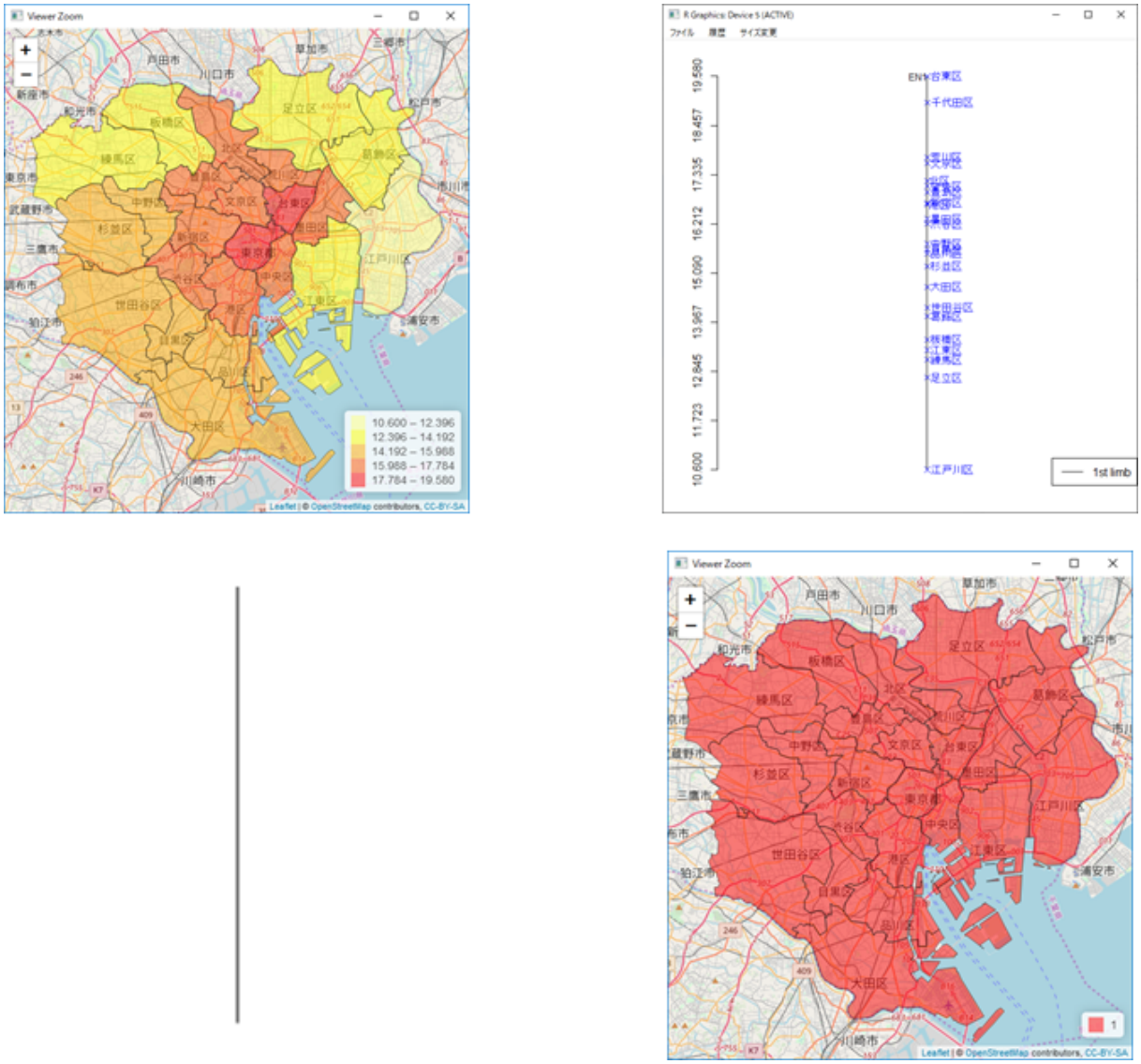


Figure 7.5: Analysis results (using 1997 data).

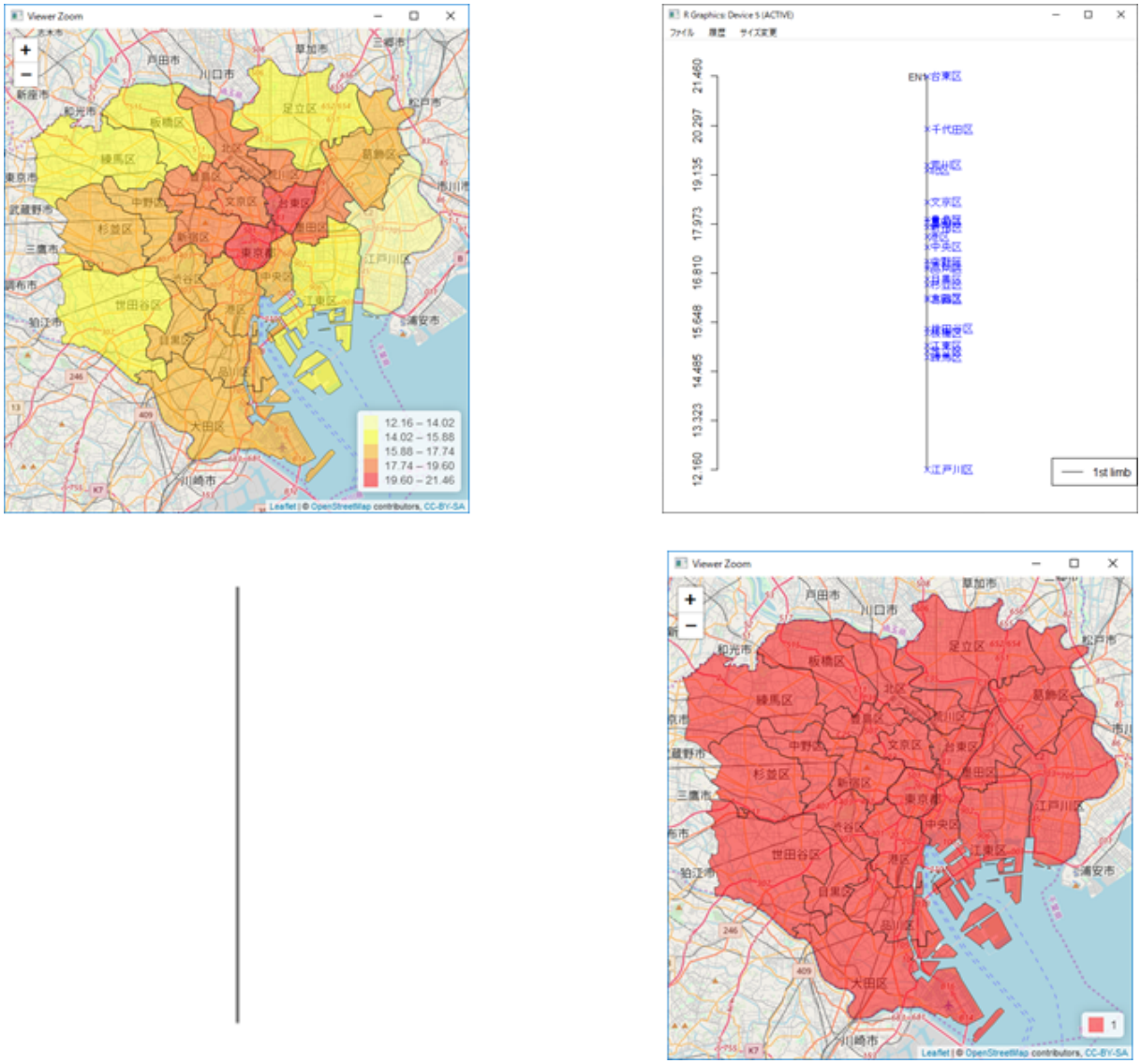


Figure 7.6: Analysis results (using 2000 data).

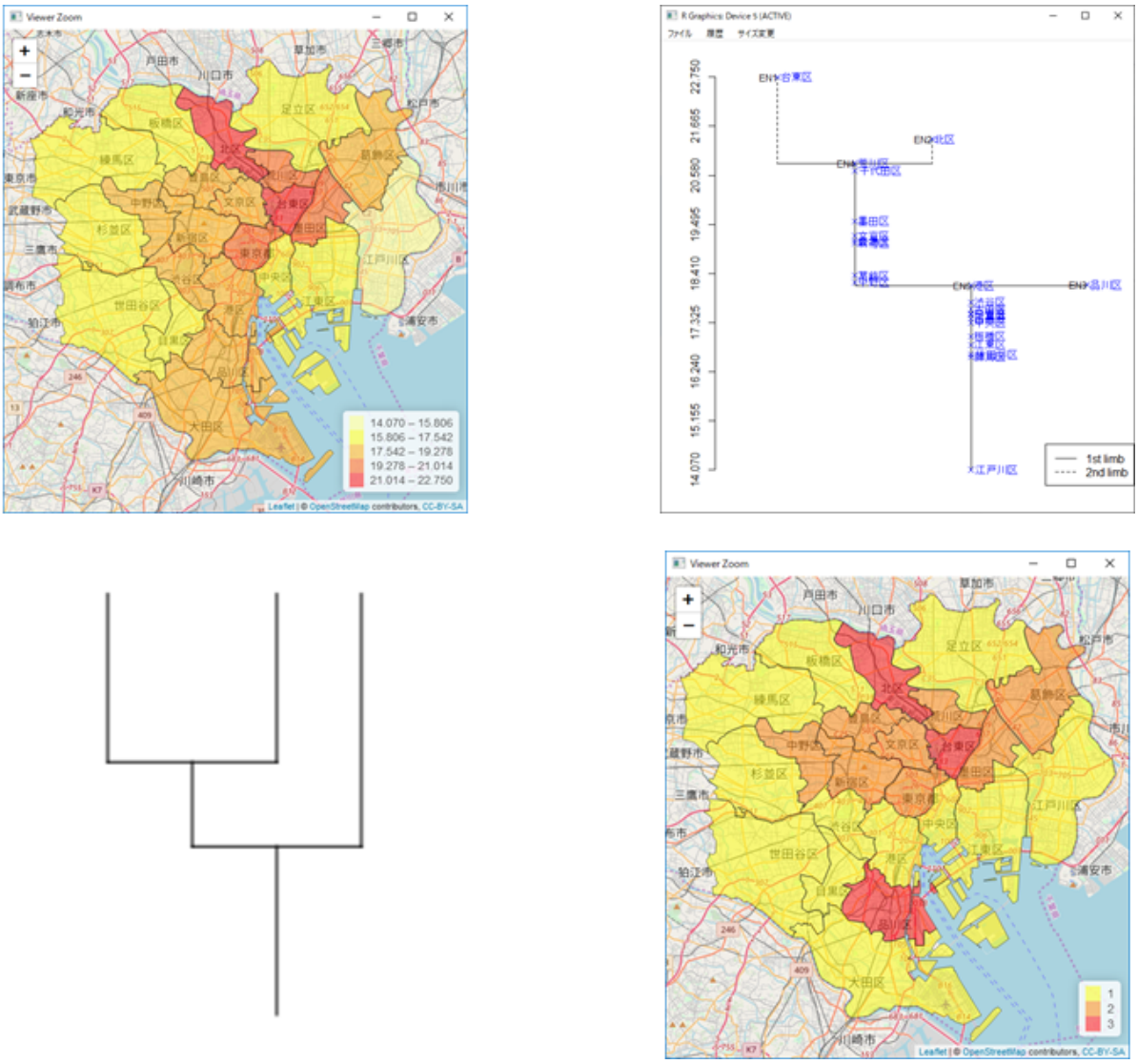


Figure 7.7: Analysis results (using 2003 data).

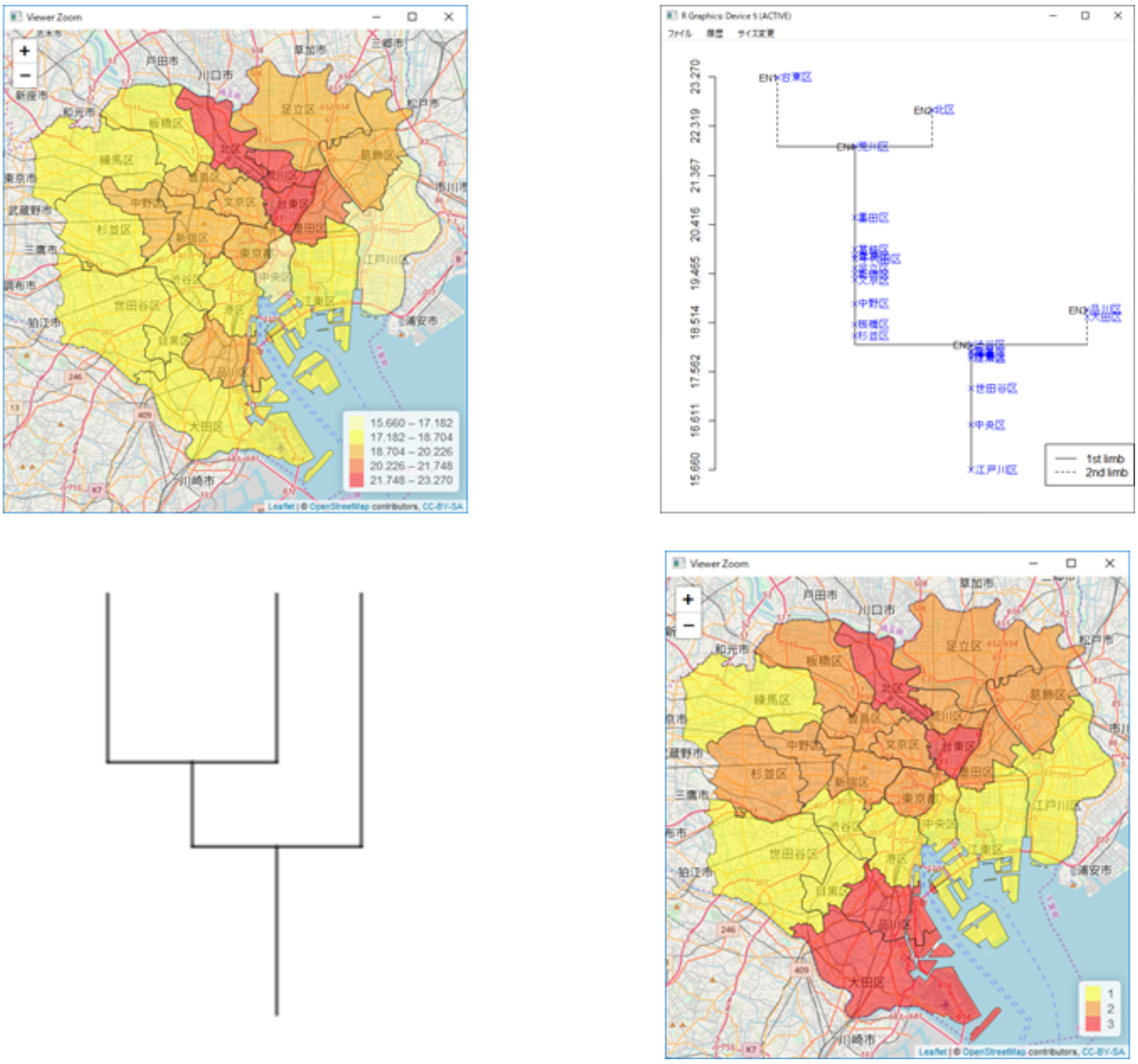


Figure 7.8: Analysis results (using 2006 data).

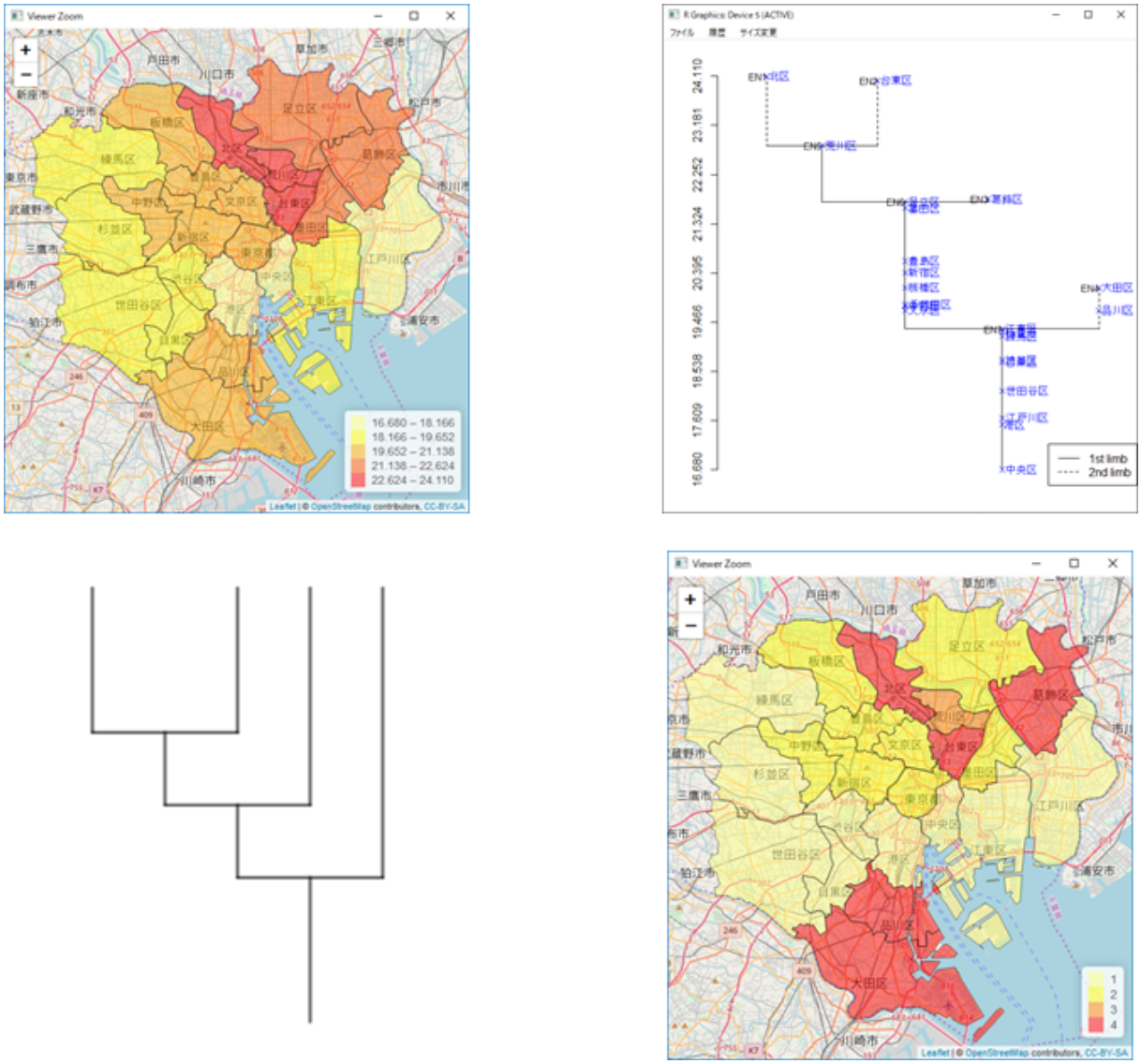


Figure 7.9: Analysis results (using 2009 data).

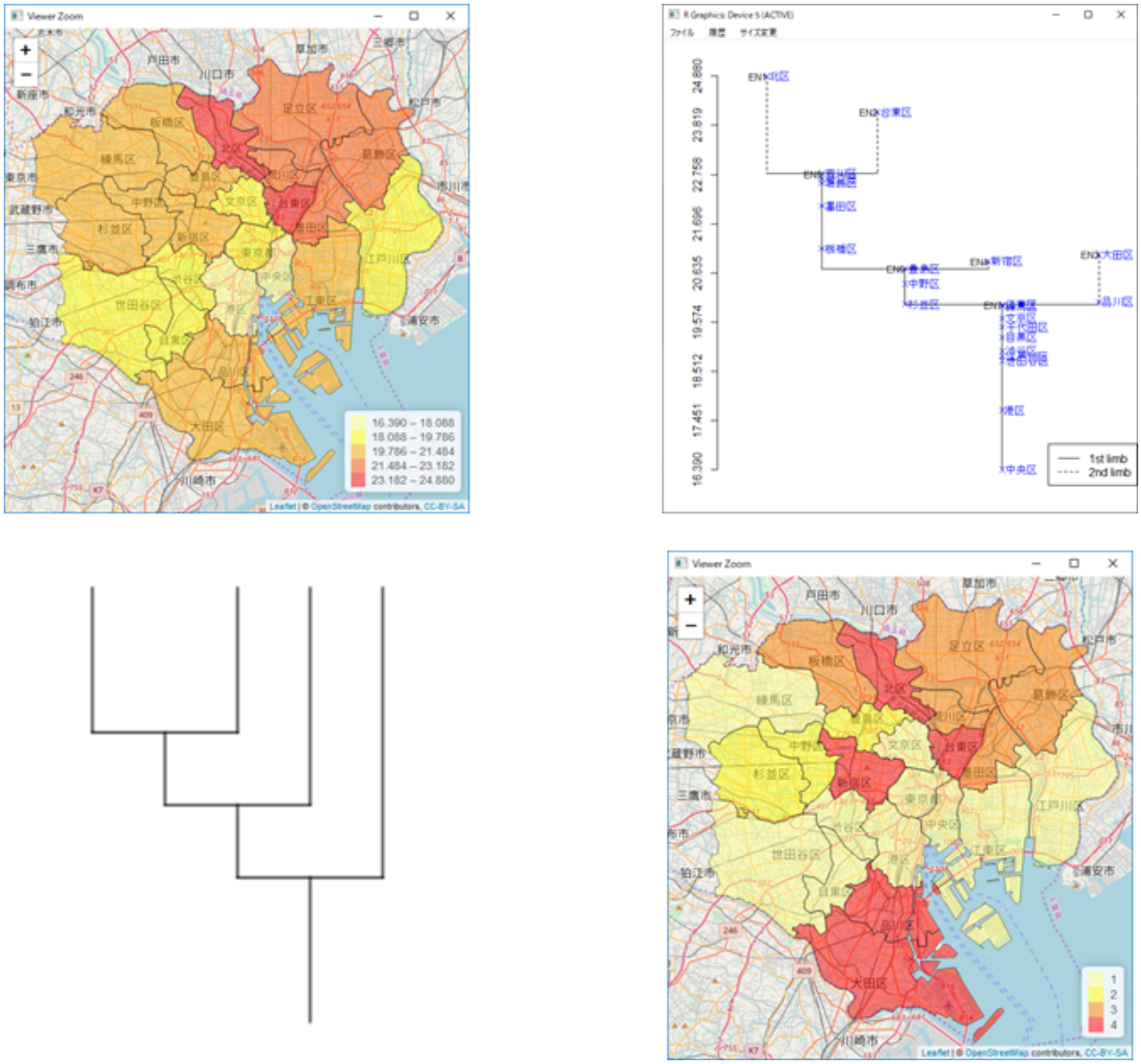


Figure 7.10: Analysis results (using 2012 data).

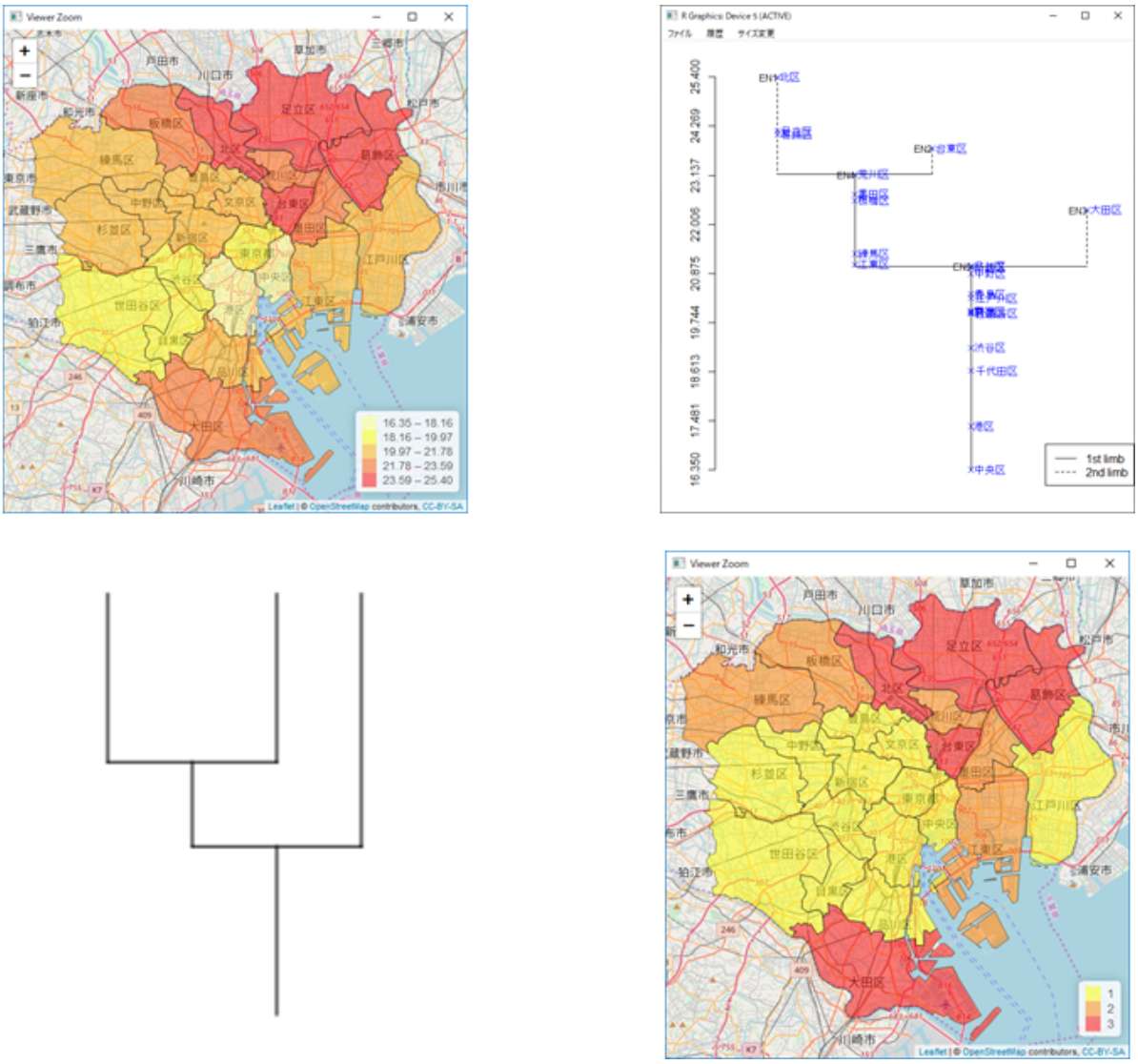


Figure 7.11: Analysis results (using 2015 data).

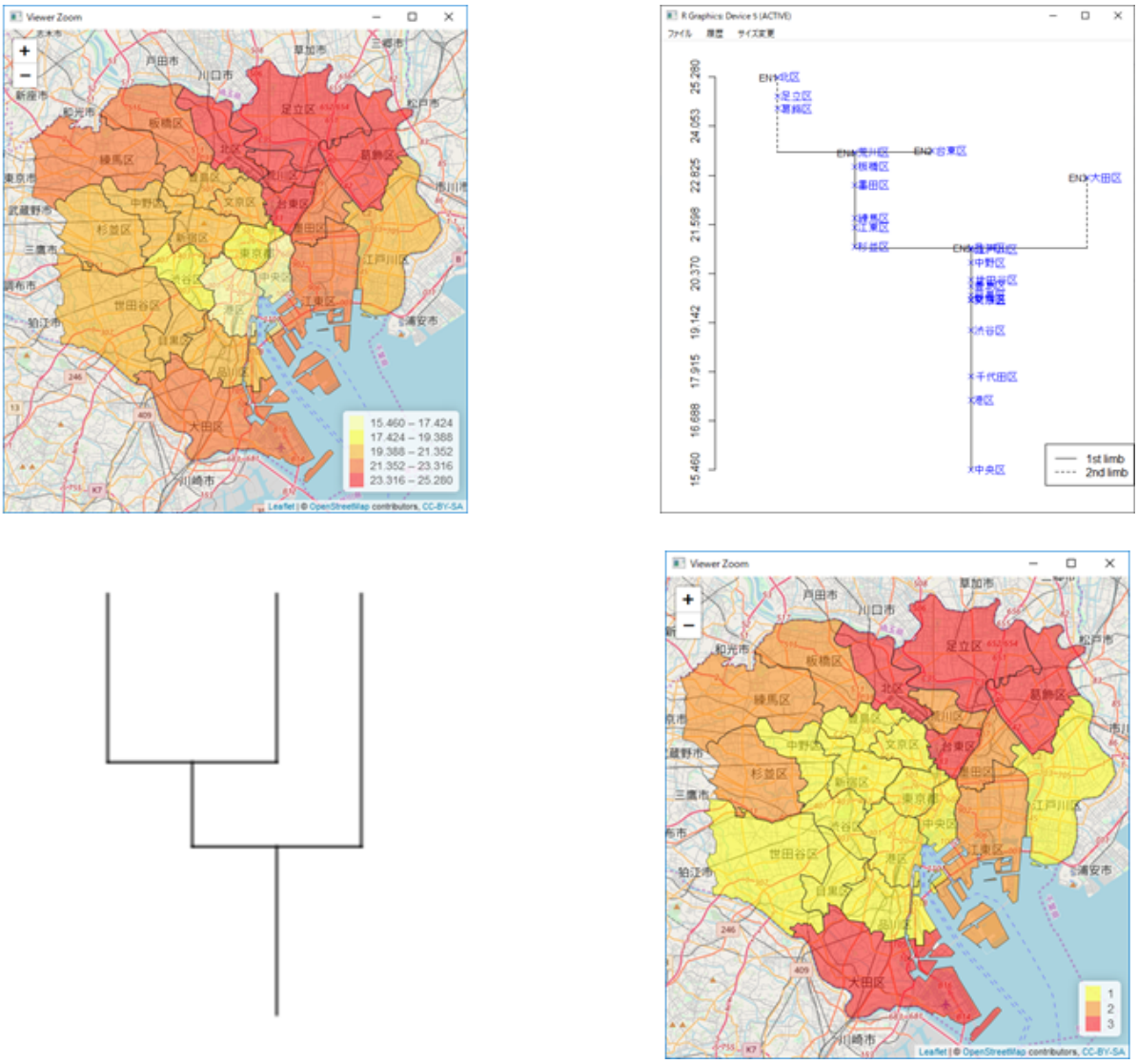


Figure 7.12: Analysis results (using 2018 data).

It is probable that the following factors were behind the transition of such a data structure. According to Suzuki *et al.* (2011) the 23 wards of Tokyo were densely populated residential areas (200 / ha or more) from the outer circumference of the Yamanote Line to the Kannana-Dori Avenue until 1995, but the range decreased from 1985 to 1995. At the same time, a large amount of hollowing out occurred in the three central wards (Chiyoda, Chuo, and Minato wards), and the population living in the city center decreased. On the other hand, since 2000 to 2005, the population growth of the three central wards and Koto ward has been increasing steadily. According to Satoh *et al.* (2011) many baby boomers have moved into suburban residential areas in the Tokyo metropolitan area since 1960. In addition, the baby boomer generation has retired and is now an elderly person. Furthermore, in recent years, there has been a tendency to rent in the city center without acquiring a home. According to Oshiro *et al.* (2009) many of the housing estates where the elderly currently live were built during the period of high economic miracle. In recent years, a number of high-rise condominiums have been built in the city center, and the population is returning to the city center. In addition, there are many child-rearing generations in the marginal areas of Tokyo's 23 wards, and the number of people aged 20 to 24 is large in the western part of Tokyo's 23 wards. Furthermore, Chiyoda Ward and Chuo Ward have tended to rejuvenate since 1995, and some of Taito Ward and Sumida Ward located in the northeastern part of the city center show signs of rejuvenation. Since 1990, the number of high-rise condominiums in central Tokyo and coastal areas has increased. In the western part of Tokyo's 23 wards, there are areas where young people expand throughout the period, and in the central part and northeastern part of the wards, the aging of the population has progressed, but in recent years there has been a tendency for rejuvenation. This tendency is especially strong in central Tokyo.

In other words, although the structure of the densely populated city center caused a temporary hollowing out, it has returned to the city center in recent years and is on a recovery trend. In addition, the population of Tokyo as a whole is increasing. Furthermore, from the perspective of demographic composition, the baby boomer generation and their children's generation tend to acquire homes without the premise of migration, and tend to have housing in the suburbs in terms of land prices, child-rearing, and commuting convenience. In addition, that generation is now elderly. In recent years, there has been little tendency to acquire homes, and many residential facilities have been built in and around the city center, and the number of younger generations is increasing in the city center by raising children there. Due to changes in living style, it is considered that the in-migration of the population has become relatively active in recent years. In addition, it is thought that the current population composition is due to the fact that many generations acquired their own homes in the suburbs when hollowing out occurred in the past. Considering this background, it is possible to infer the cause of the change in the spatial data structure due to changes in the times, as shown in Figures 7.1 to 7.12.

In this chapter, we evaluated changes in the proportion of the population aged 65 and over in the 23 wards of Tokyo from the perspective of the complexity of the spatial data structure. However, if the number of regions is too large, the shape of the dendrogram becomes very complicated, and the numbers of patterns and stages tend to be too large. Therefore, it becomes difficult to evaluate the complexity of the shape of the dendrogram. Therefore, in the next chapter, we will consider merging the peaks of the dendrogram, leaving the tree up to a shape that can explain the characteristics of the spatial data structure to some extent, and merging the leaf parts that form the rest and evaluate.

8 Merging of dendrogram peaks

This chapter defines the "merging of peaks" in the dendrogram. "Merging of peaks" aims to facilitate the evaluation of overly complex dendrograms by simplifying the dendrogram. The problem with the Echelon dendrogram is that if the number of areas to be analyzed becomes too large, the generated dendrogram will have a very complicated shape, making it difficult to understand its structure and regional relationships. When trying to evaluate a complex dendrogram, the six indicators and the number of stages in the dendrogram patterning described so far become very large, making it difficult to judge. Therefore, we consider stopping the growth until the part where the shape of the dendrogram can be sufficiently expressed, and evaluating the complexity by the shape up to that stage. As a specific example, we use data on the number of homicides (1960, 1970, 1980, 1990) in 3085 regions of each state in the United States. After calculating SMR from this data and performing Echelon analysis, an Echelon dendrogram (Figure 8.1) was created. The six indicators when this dendrogram was patterned are shown in Table 8.1, and the number of cycles in the Echelon tree was 5. Moreover, the number of stages can be found to be 433 from the value of NP. In addition, the calculation results of the four indicators of Echelon profiles (Table 8.2) and their graphs (Figure 8.2) are shown.

Table 8.1: 6 indicators of the 1960 dendrogram.

NE	NP	MF	MP	LU	LV
812	433	753	34	351	32963

Table 8.2: Dendrogram Cycle and 4 Scales (1960).

	Cycle1	Cycle2	Cycle3	Cycle4	Cycle5
Divergence	0.203	0.600	0.899	0.990	1
Scope	0.594	0.812	0.954	0.995	1
Bunching	0.478	0.520	0.166	0.018	0
Stacking	0.000	0.360	0.822	0.982	1

As is clear from these results, it is difficult to judge the complexity of the data structure because the dendrogram with a large number of regions and a complicated data structure has a large number of patterns, stages, and cycles. Scope is an index showing what percentage of the total area is included in the dendrogram at the time of each cycle. Looking at Table 8.2, in this data, the Scope value exceeds 90% as of Cycle 3. Therefore, considering that this dendrogram contains 90% of the total as of Cycle 3, it is

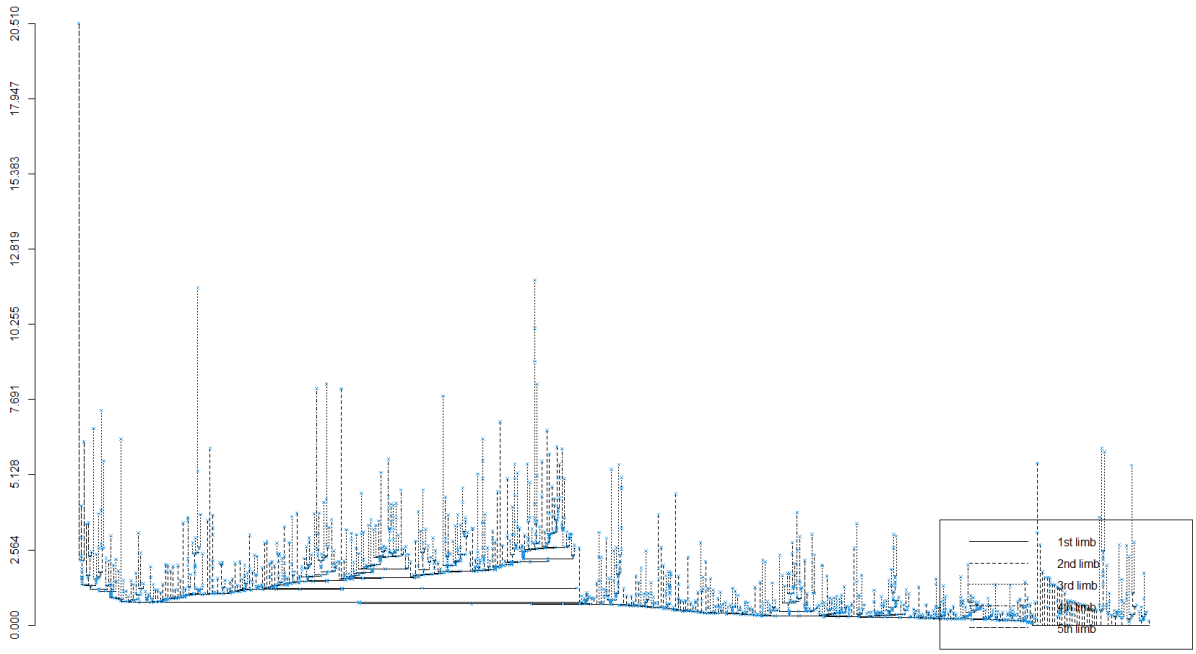


Figure 8.1: Dendrogram in SMR of 1960.

considered that the complexity of the data structure can be fully explained. As a method of generating a dendrogram up to an arbitrary cycle in this way, "merging of peaks" of the dendrogram was defined. Until the dendrogram reaches a fully explained cycle, the regions that make up the peaks with a common parent are adjacent to each other, and the peaks in the upper hierarchy are merged in order to reduce the number of peaks. By merging the peaks, not only is it easier to evaluate complex dendrograms, but it is also possible to remove outliers, so it is expected that robust evaluation will be possible. The procedure for "merging of peaks" is shown below.

1. Detect the peak of the maximum cycle.
2. Find the peak located at the highest level in step 1.
3. Of the peaks in step 2, the peaks having a common parent are placed adjacent to each other, and the Echelon analysis is performed again.
4. Repeat steps 1 to 3 until the target cycle reaches the maximum cycle.

Figure 8.3 shows the result of merging the peaks of the dendrogram of Figure 8.1 until Cycle 3. In addition, the four scales of profiles at that time are as shown in Table 8.3, and the transition is as shown in Figure 8.4. Table 8.4 shows the six indicators calculated during patterning after merging. Comparing the Table 8.1 and Table 8.4, it can be seen

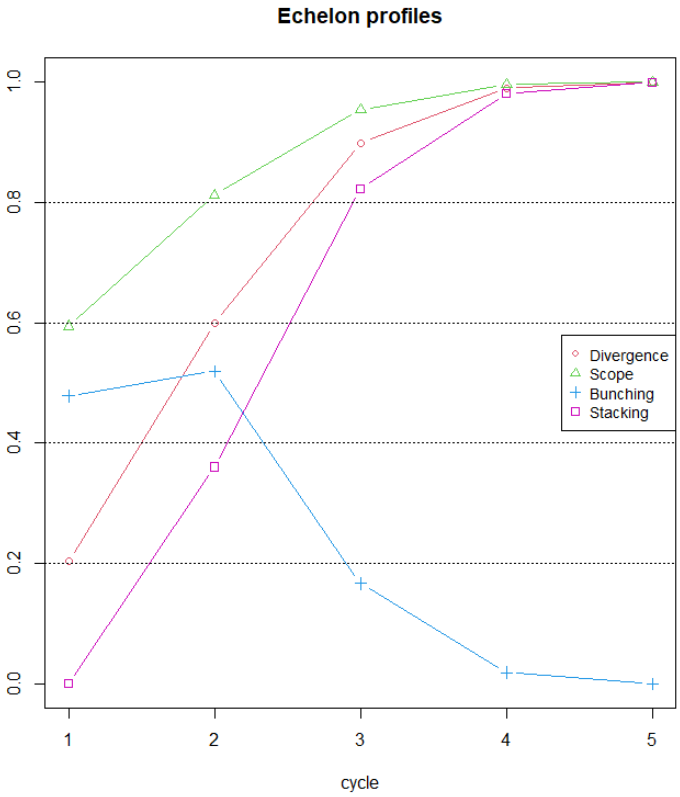


Figure 8.2: Graph in Table 8.2.

that the number of stages decreased from 433 to 378. Table 8.5 shows the results of merging peaks in the data aggregated for each year. Looking at the tables before and after the merger, it can be seen that the magnitude relationship between the number of stages has changed. Also, in this example, the merge operation did not show a very large decrease in the stages, but the larger the number of data, the greater the decrease is expected. By merging the peaks in this way, the evaluation can be performed in a state where the dendrogram can be fully explained. In addition, since the dendrogram excluding abnormal values can be evaluated, it is considered that the results show a difference between before and after the merging. Performing "merging of peaks" is useful when evaluating a data structure because it can be expressed simply while grasping the characteristics of the data structure. However, there is no index as to how much the value of Scope is enough to explain the data structure, and it is necessary for the analyst to make a judgment in consideration of the characteristics of the data.

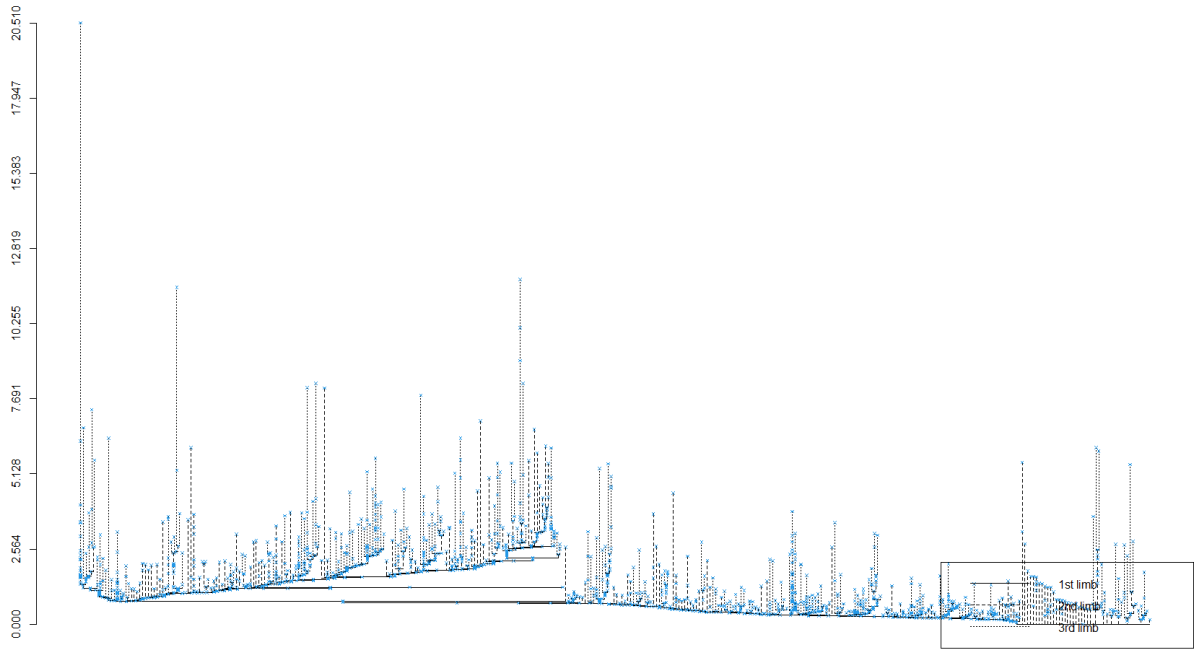


Figure 8.3: Dendrogram in SMR of 1960 after merging.

Figure 8.5 show the transition of the four scales of echelon profiles calculated each time the peaks are merged from Cycle 5 to Cycle 3 and Figure 8.6 shows how much the six indicators used in the patterning are reduced from the original dendrogram. In Figure 8.5, it can be seen that the values of the four indicators increase until the Cycle decreases, and when the Cycle decreases by one, the values also decrease significantly. Since the recalculation is performed every time the peaks are merged, it can be seen that

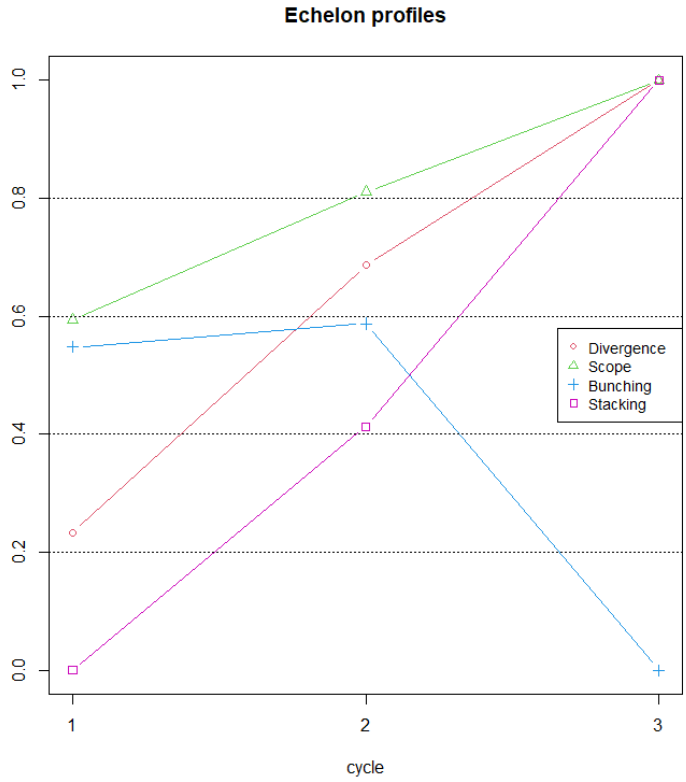


Figure 8.4: Graph in Table 8.3.

Table 8.3: Dendrogram Cycle and 4 Scales (1960) after merging.

	Cycle1	Cycle2	Cycle3
Divergence	0.233	0.686	1
Scope	0.594	0.811	1
Bunching	0.548	0.587	0
Stacking	0.000	0.413	1

Table 8.4: 6 indicators of the 1960 dendrogram after merging.

NE	NP	MF	MP	LU	LV
707	378	648	34	323	28291

Table 8.5: NE values before and after merging from 1960 to 1990.

	1960	1970	1980	1990
Before merging	433	454	439	449
After merging	378	395	391	384

the value changes each time. In addition, Figure 8.6 does not show a large change, and it can be seen that the ratio tends to decrease when the peaks are merged. The criteria for stopping merging should be determined according to the characteristics of the data, such as when the values of the metrics that characterize the data structure change significantly during the merge operation. However, this is not defined in this paper. In this paper, the standard was determined from the value of Scope, but in the future it will be required to define a clear one as needed.

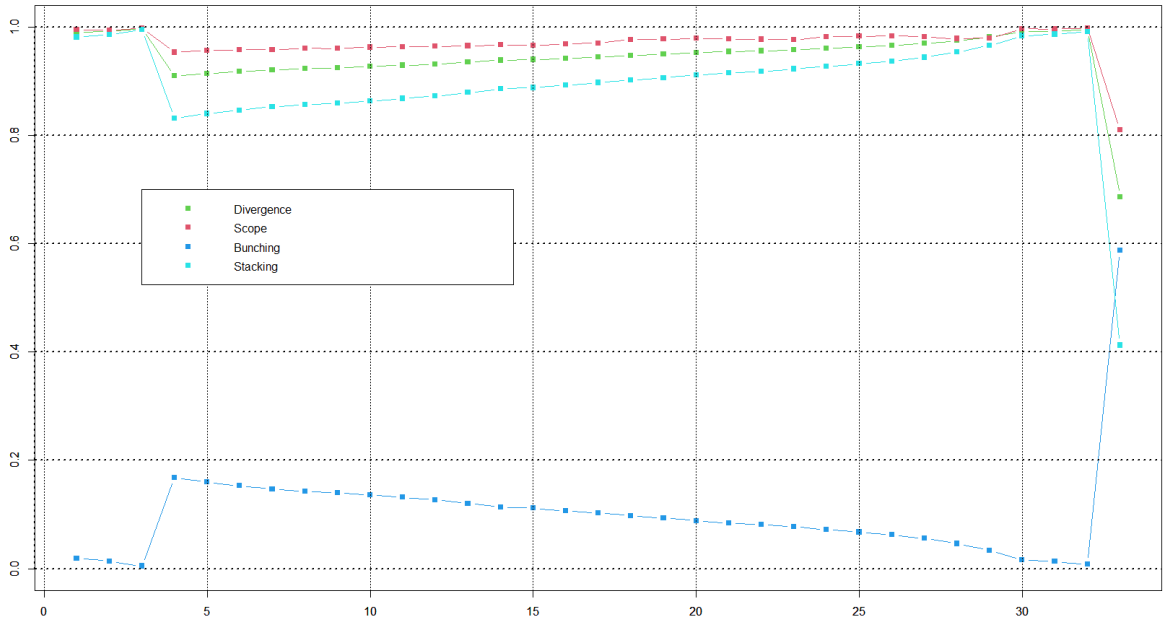


Figure 8.5: Transition of 4 scales aggregated for each merge operation.

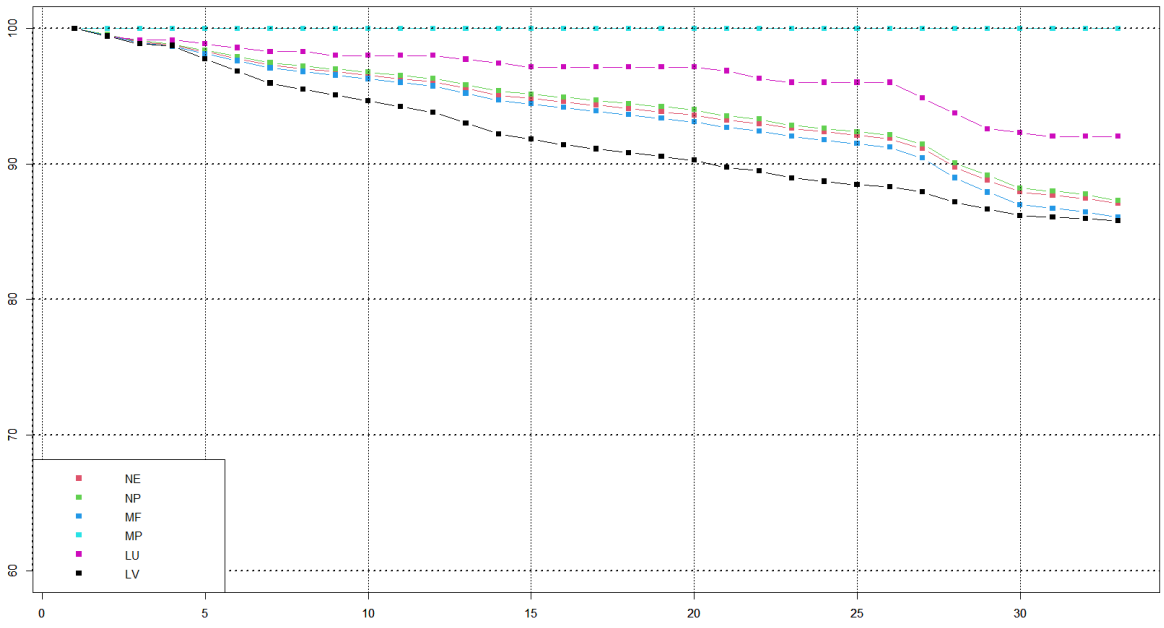


Figure 8.6: Transition of 6 indicators used for patterning aggregated for each merge operation.

9 Summary

In this paper, we introduced SMR and EBSMR, which are widely used in the field of spatial epidemiology and we described the details of Echelon analysis for useful for studying the topological structure of the surface of spatial data in a systematic and objective way. After that, the detection of spatial hotspot cluster using spatial scan statistics was described, and the scanning method for detecting hotspot cluster was described. In addition, we introduced the Web application that can perform Echelon analysis that has already been developed. We also developed a Web application that can implement these series of analyzes in a unified framework using the R package shiny. The developed software can be analyzed using the map information and data of any analysis target area prepared by the analyst. In addition, since parameter settings can be easily performed, quick recalculation and analysis can be performed with arbitrary settings. Since the interpretation of the analysis results differs depending on the characteristics and methods of the data to be handled, appropriate judgment is required.

Later in this paper, we defined indicators that can evaluate complexity in spatial data. The spatial data was visually grasped by utilizing the Echelon dendrogram generated during the Echelon analysis. In addition, by "patterning" dendrograms with various shapes, the shapes are unified, and by calculating 6 indexes including the LV defined in this paper, even in dendrograms with complicated shapes, to quantitatively evaluated the complex shapes of the dendrograms. Furthermore, by using the Cycle required when calculating the four indicators of echelon profiles and the "stage" of the dendrogram defined in this paper, we were able to compare and examine the data structure that changes with time. After that, its usefulness was confirmed using data on the proportion of the population aged 65 and over in the 23 wards of Tokyo. Finally, in order to solve the problem of the Echelon dendrogram, which has a large number of regions and becomes too complicated, we defined the "merging of peaks" of the dendrogram and evaluated it when the dendrogram could be fully explained. We also confirmed the usefulness of merging of peaks using data on the number of crimes in the United States in 3,085 regions. It is also expected that the dendrogram with the merged peaks will be used for the hotspot cluster detection. Therefore, it is necessary to newly define the concept of cluster detection and the concept of neighborhood information. In addition, there is ample room for research on the structural analysis of spatial data, and it is necessary to establish more appropriate indicators and evaluation criteria.

EcheScan

Software overview

This appendix introduces the software dedicated to Echelon analysis called EcheScan that has already been developed in Kurihara (2020). EcheScan can constructs the echelons from input data files which consists of univariate values and neighbors for each lattice and visualizes the result of a dendrogram. In addition, by inputting the information of the observed value and its expected value, the hotspot cluster based on the Poisson distribution can be detected by using the Echelon scan method. This software is built in the R shiny environment and published on the website (`https://fishi.ems.okayama-u.ac.jp/echescan`), so anyone can easily analyze it. Table 1 is a summary of information required for input and information that can be output. Next, the files required for input will be described in detail.

Table 1: Input and output files of EcheScan.

I O	File	Contents	Notes
I	Neighborhood information	Neighbor information of each lattice	
I	Univariate	Value (h) of each lattice	
I	Case & expectation	Observed(c) and expected(λ) values of each lattice	For hotspot detection based on Poisson model
O	Echelon table	Details of echelons	File format: `csv`
O	Lattices forming echelon	Lattice information within each echelon	File format: `csv`
O	Echelon dendrogram	Graphical representation of echelons	File format: `png, pdf, eps`
O	Hotspot table	Details of detected hotspots	File format: `csv`
O	Echelon dendrogram with scanning	Graphical representation of echelon scan technique	File format: `png, pdf, eps`

Neighborhood information file

The neighborhood information file provides the name and neighborhood of each area. The first column of each row is the area name, and the numbers entered after the next column are the row numbers of the adjacent areas.

Univariate file

Univariate files provide the values for each lattice in one column.

Case & Expected file

The Case & Expected Value file is a two-column file that provides observations and expectations for each region to detect hotspot clusters.

The Univariate file and Case & Expectation file must be in the same order of the regions provided in the Neighborhood information file. Also, the number of lines in these three files must be the same as the total number of areas.

example 1 : One dimensional lattice

We will introduce how to use EcheScan using the data introduced in Table 3.1. This example uses the neighborhood information file (dim1nb.txt) and the univariate file (dim1h.txt) shown in Figure 1 and Figure 2. The information in the neighborhood file in Figure 1, for example, shows that the lattice "C" (third line) is adjacent to the second line (B) and the fourth line (D). Figure 3 shows the software start screen. First, select dim1nb.txt from [Brows] in "Neighborhood Information" on the left side of the screen. If there is no problem with the file, "Univariate" is displayed and selsect dim1h.txt. Then click Run to perform the Echeron analysis and the results will be displayed on the Echeron dendrogram tab (Figure 4). The table at the top of Figure 4 provides detailed information for each hierarchy.

A	2	
B	1	3
C	2	4
D	3	5
E	4	6
F	5	7
G	6	8
H	7	9
I	8	10
J	9	11
K	10	12
L	11	13
M	12	14
N	13	15
O	14	

Figure 1: Neighborhood information file for the one dimensional lattice data.

1
2
3
4
3
4
5
4
3
2
3
2
1
2
1

Figure 2: Univariate file for the one dimensional lattice data.

This table will be described in detail. The first field is echelon number (EN). The second field is Order, which gives an integer value greater than or equal to 1; "1" means a peak, "2" means a foundation of order 1s, "3" means a foundation of order 2s, and

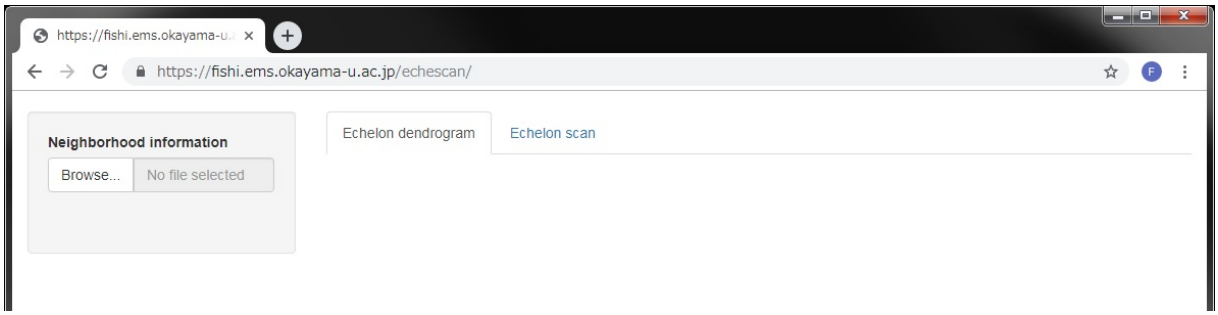


Figure 3: Screen when software start.

so on. The third field is Parent, which gives the echelon number of the parent. The forth field is Maxval, which gives the maximum value. The fifth field is Minval, which gives minimum value. The sixth field is Length, which is the length of Maxval - parent ' sMaxval. The seventh field is Cells, which gives the number of lattices. The eighth field is Progeny, which gives the number of ascendants (children) for the echelon. The nineth field is Family, which gives the number of echelons in the family. The final field is Level, which gives the number of echelons in the ancestor. These information and dendrogram of echelons can also be output as a file. The details of V ariable are shown in the papers of Myers et al. (1997) and Kurihara et al. (2000).

example 2 : Hotspot cluster detection example (Lung cancer data in New Mexico)

As a second use case, we apply echelon analysis and the echelon scan method to lung cancer data in New Mexico available on the SaTScan web site(`https://www.satscan.org/datasets/nmlung/`). The data consists of the number of cases of malignant lung cancer from 1973 to 1991 and the number of populations in 32 areas. A total of 9,254 cancer cases and a population of 25,604,291 were recorded during this period, consisting of the following category covariates: 18 age groups (group1 = ages < 5, group2 = ages5 $-$ 9, group3 = ages10 $-$ 14, ..., group17 = ages80 $-$ 84, group18 = ages85+) and gender (1 = male, 2 = female).

Introduce an example of using EcheScan to detect hotspot clusters in this data. First, as show Figure 5, 6 and 7, prepare the neighborhood information file (NMnb.txt), the univariate file (NMsmr.txt), and the case & expected value file (NMCasExp.txt) for calculating the spatial scan statistics. When you read the neighborhood information file and the univariate file, "Case & Expectation" is displayed, so load the case & expectation file (NMCasExp.txt). Then select the Echelon scan tab and select RUN to start the analysis. Figure 8 shows the execution result when setting the significance level = 0.05, the

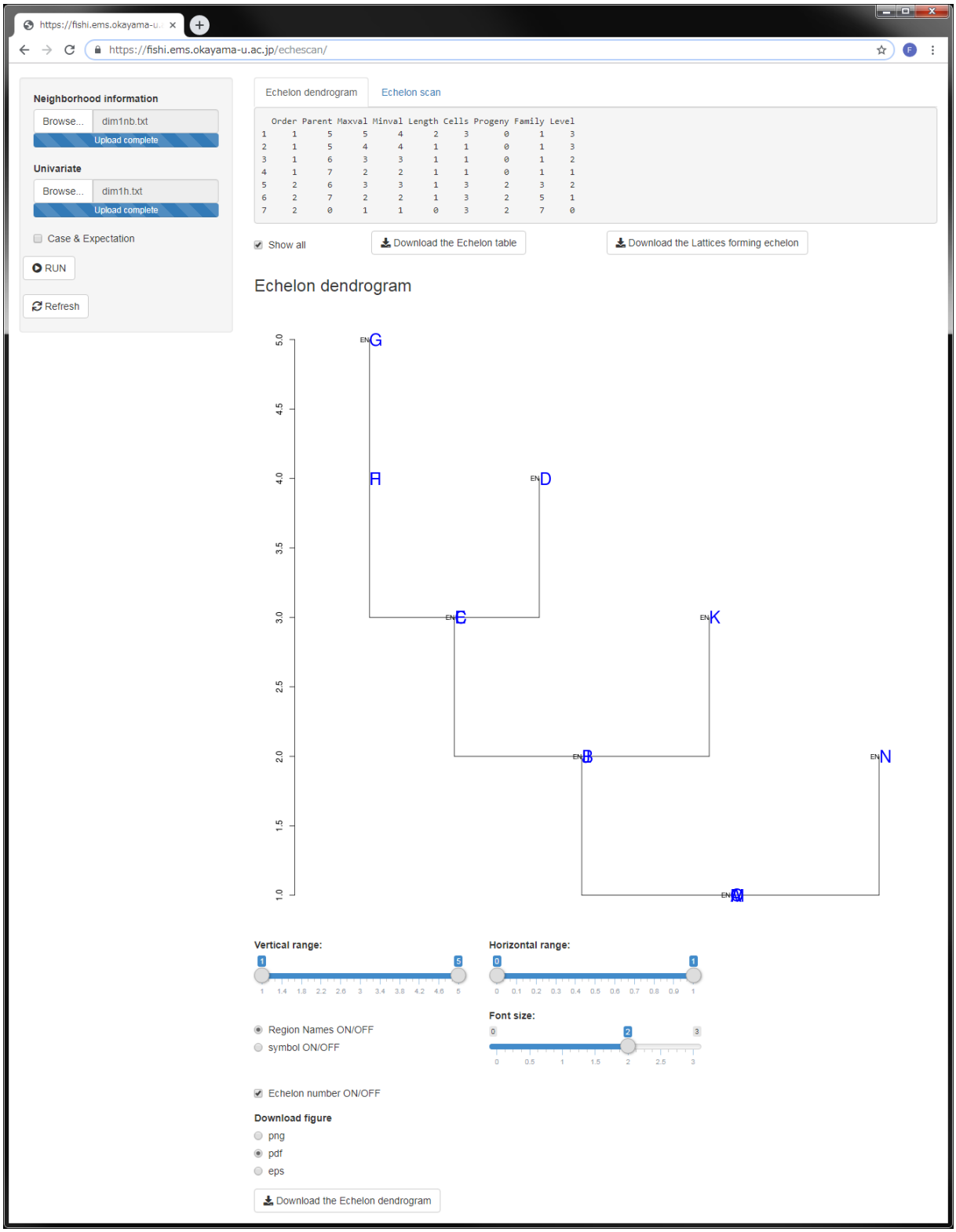


Figure 4: Execution result of Echelon analysis for the one dimensional lattice data.

maximum hotspot size = 30, and the Monte Carlo replications = 999. You can zoom in or out the dendrogram display by changing the "Vertical range:" and "Horizontal range:" settings at the bottom of the screen. Figure 9 is an enlarged view of the echelons recognized as the hotspot cluster. The $\log \lambda(Z)$ value of equation (4.8) in the detected hotspot cluster was 93.883 and the p-value was 0.001.

(1,1)	33	25	26	30	32			
(2,1)	34	9	27	28	32			
(3,1)	35	6	8	13	14	19	22	
(4,1)	36	11	18	29	31			
(5,1)	37	20	22					
(6,1)	38	3	10	14	20	22		
(7,1)	39	16	19	27				
(8,1)	40	3	13	19				
(9,1)	41	2	12	16	27			
(10,1)	42	6	14	20	24	30		
(11,1)	43	4	18	20	24	31		
(12,1)	44	9	16					
(13,1)	45	3	8	22				
(14,1)	46	3	6	10	19	27	28	30
(15,1)	47	21	25	26				
(16,1)	48	7	9	12	27			
(17,1)	49	23	25	32				
(18,1)	50	4	11	21	24	26	29	
(19,1)	51	3	7	8	14	27		
(20,1)	52	5	6	10	11	22	24	31
(21,1)	53	15	18	23	25	26	29	
(22,1)	54	3	5	6	13	20		
(23,1)	55	17	21	25				
(24,1)	56	10	11	18	20	26	30	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 5: Part of the contents of neighborhood informationfile for lung cancer data in New Mexico.

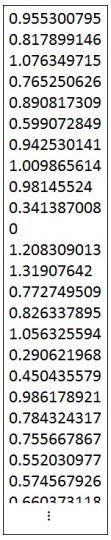


Figure 6: Part of the contents of univariate file for lung cancer data in New Mexico.

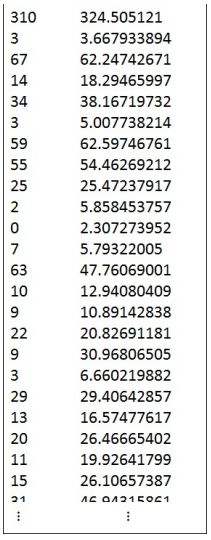


Figure 7: Part of the contents of case & expectation
file for lung cancer data in New Mexico.

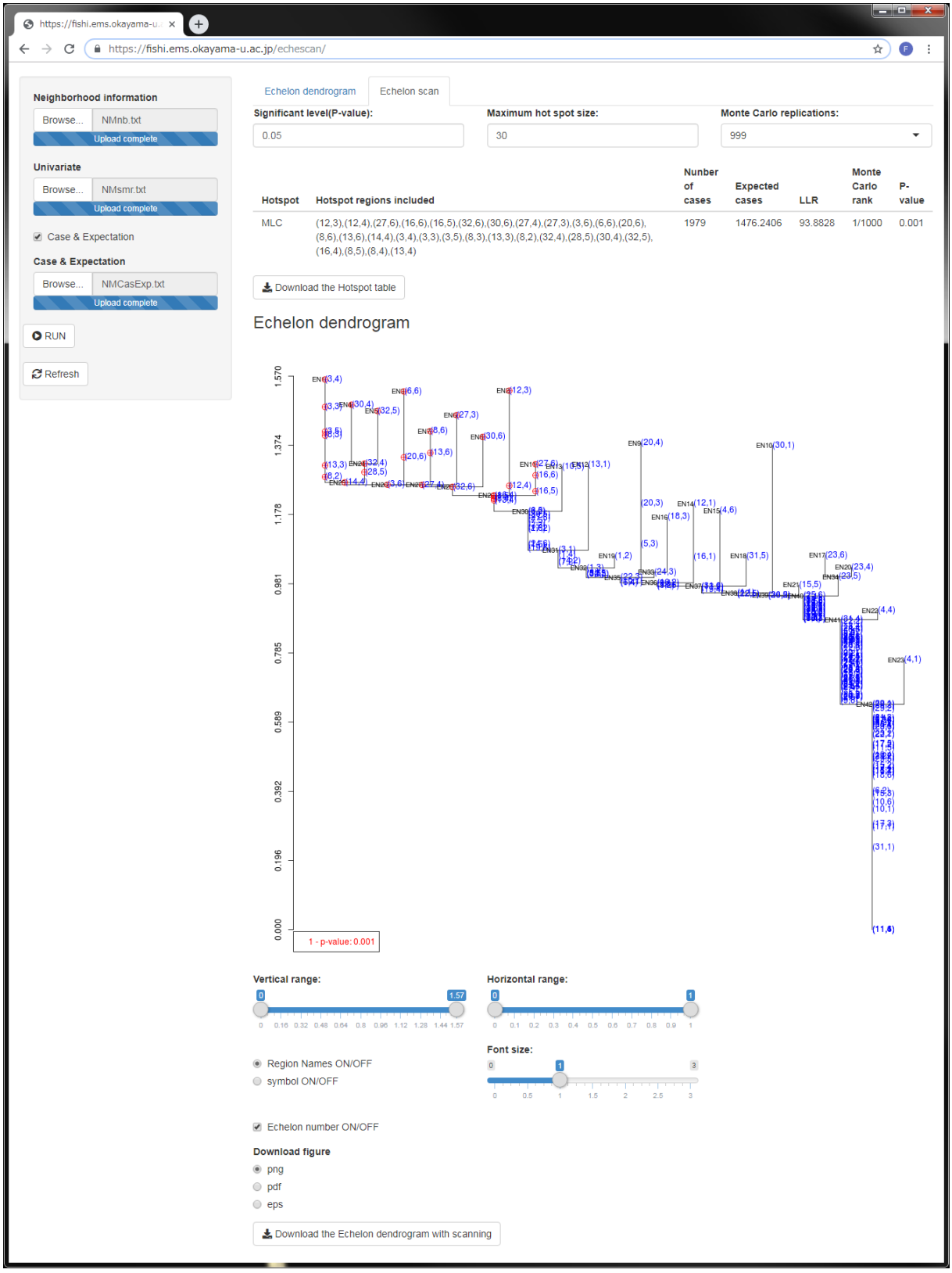


Figure 8: Execution result of hotspot cluster detection using the Echelon scan method for lung cancer data in New Mexico.

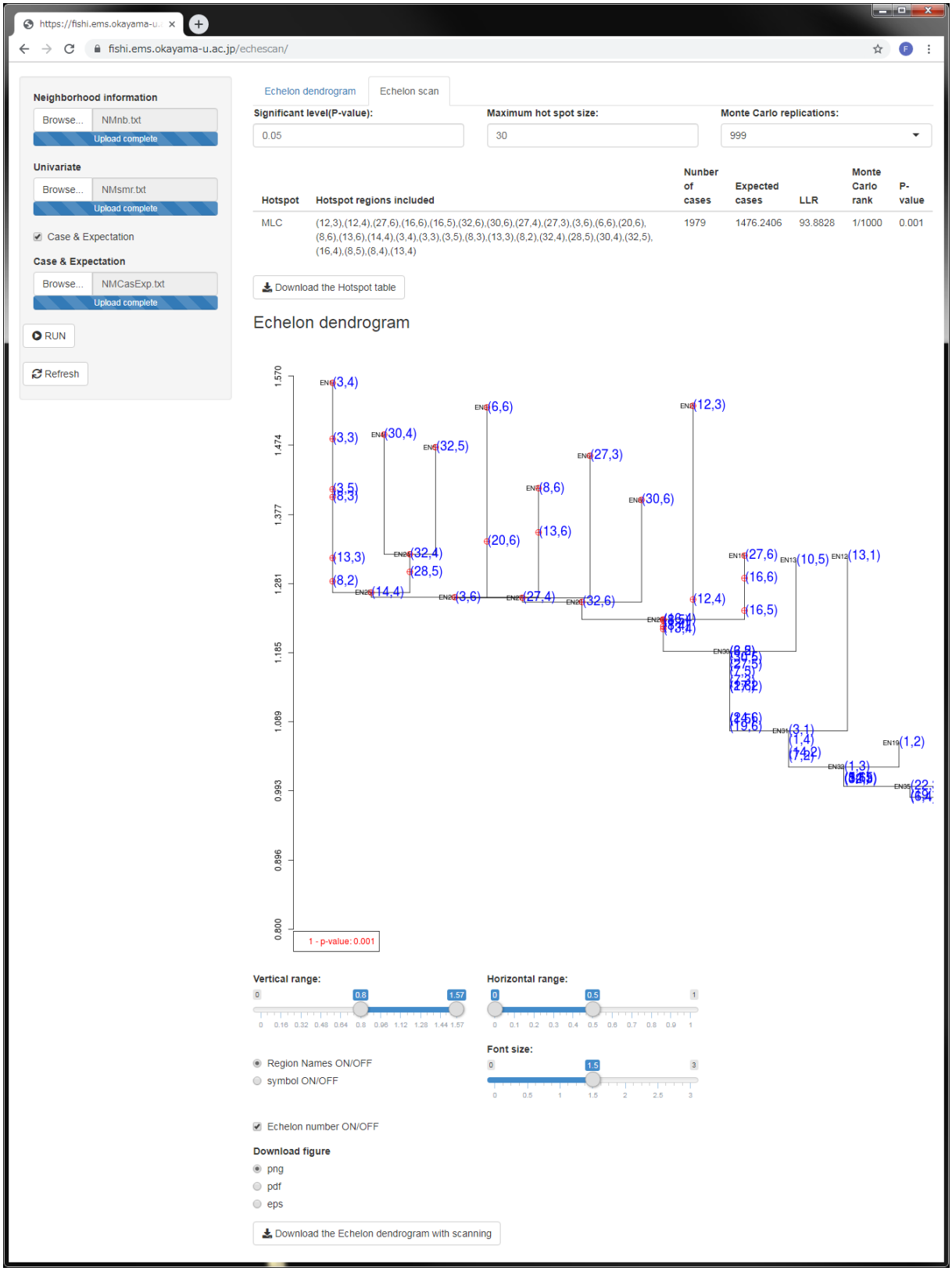


Figure 9: Enlarged view of the hierarchy of detected hotspot cluster areas.

References

Chang, W. *et al.* (2020). shiny:Web Application Framework for R. `https://cran.r-project.org/package=shiny` (accessed : January 8th, 2021)

Chen, C, Kim, A, Y., Ross, M & Wakefield, J(2018). SpatialEpi: Methods and Data for Spatial Epidemiology. `https://cran.r-project.org/web/package=SpatialEpi` (accessed : January 8th, 2021)

Cressie, N.(1993). *Statistics for Spatial Data, Revised Edition.* Wiley.

Ishioka, F., Kurihara, K., Suito, H., Horikawa, Y., & Ono, Y. (2007). Detection of hotspots for 3-dimensional spatial data and its application to environmental pollution data, *Journal of Environmental Science for Sustainable Society*, **1**(1), 15–24.

Ishioka, F. & Kurihara, K. (2012). Hotspot Detection Using Scan Method Based on Echelon Analysis, *Proceedings of the Institute of Statistical Mathematics*, **60**(1), 93–108.

Ishioka, F., Kawahara, J., Mizuta, M., Minato, S. & Kurihara, K. (2019). Evaluation of hotspot cluster detection using spatial scan statistic based on exact counting, *Japanese Journal of Statistics and Data Science*, **2**, 241–262.

Ishioka, F.(2020). echelon: The Echelon Analysis and the Detection of Spatial Clusters using Echelon Scan Method. `https://cran.r-project.org/web/package=echelon`. (accessed : January 8th, 2021)

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

Kulldorff, M., & Harvard Medical School. (2020). SaTScan v9.6.1: Software for the spatial, temporal, and space-time scan statistics. `https://www.satscan.org/` (accessed : January 8th, 2021)

Kurihara, K., Myers, W. L., & 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. (2003). The detection of hotspots based on the hierarchical spatial structure, *Bulletin of the Computational Statistics of Japan*, **15**(2), 171–183.

Kurihara, K. (2004). Classification of geospatial lattice data and their graphical Representation. *Classification, Clustering, and Data Mining Applications* (Banks, D., McMorris, F. R., Arabie, P. & Gaul, W. (Eds)), *springer*, 251–258.

Kurihara, K., Ishioka, F. & 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.

Kurihara, K. & Ishioka, F. (2007). Classification of Spatial Data Based on the Pattern of Hierarchical Structure and Its Applications, *Japanese Journal of Statistics and Data Science*, **37**(1), 113–132.

Kurihara, K., Ishioka, F. & Kajinishi, S. (2020). Spatial and temporal clustering based on the echelon scan technique and software analysis, *Japanese Journal of Statistics and Data Science*, **3**, 313–332.

Myers, W. L., Patil, G. P. & Joly, K. (1997). Echelon approach to areas of concern in synoptic regional monitoring, *Environmental and Ecological Statistics*, **4**, 1481–1496.

Oshiro, M. & Suzuki, T. (2009). District type classification by change in age structure and its dependence on housing development-A Case of Tokyo, 1970-2005- , *Journal of the City Planning Institute of Japan*, **44**(3), 727–732.

Satoh, H. & Shimizu, C. (2011). A study on Residential Relocation within the Tokyo Metropolitan area, *Journal of the City Planning Institute of Japan*, **46**(3), 559–564.

Suzuki, A., Koizumi, H & Okata, J. (2011). Housing Change on Population Recovery in Tokyo Wards from 2000 to 2006, *Journal of the City Planning Institute of Japan*, **46**(3), 739–444.

Takahashi. (2006). EB estimator for Poisson-Gamma model v2.1. `https://www.niph.go.jp/soshiki/gijutsu/download/ebpoig/index_j.html`. (accessed : January 8th, 2021)

Takahashi, K., Yokoyama, T. & Tango, T. (2007). FlexScan, `https://www.niph.go.jp/soshiki/gijutsu/index_e.html`. (accessed : January 8th, 2021)

Tango, T. (1999). Disease mapping and spatial disease clustering -Toward an appropriate interpretation and use of disease indices-, *J. Natl. Inst. Public Health*, **48**(2), 84–93.

Tango, T., Yokoyama, T. & Takahashi, K. (2007). Kuukanekigakuhenosyoutai (in Japanese), Asakurasyoten.

Tango, T. & Takahashi, K. (2012). A flexible spatial scan statistic with a restricted likelihood ratio for detecting disease clusters, *Statistics in Medicine*, **31**(30), 4207–4218.

Tango, T. & Imai, A. (2013). DMS : Disease Mapping System, `https://www.niph.go.jp/soshiki/gijutsu/download/DMS/index_j.html`. (accessed : January 8th, 2021)

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

Acknowledgements

I would like to express my sincere gratitude to Professor Koji Kurihara of Department of Environmental and Mathematical Sciences, Okayama University, who gave me the opportunity to study, for their continuous teaching, guidance and encouragement on my study. I also express my sincere appreciation to Associate professor Fumio Ishioka of Department of Environmental and Mathematical Sciences, Okayama University, and Assistant professor Na Myungjin of Faculty of Law, Okayama University, for his valuable teaching, advice and encouragement on my study.