Proposal and application of a new method for cluster detection in large-scale spatial data by the echelon scan method

MARCH 2022

Yusuke Takemura

Contents

1 Introduction

Information that describes human activities and the natural environment together with their position in space is called spatial information, and information that is processed so that it can be used for statistical analysis is called spatial data. In recent years, spatial data can be easily collected by position information terminals such as GIS (Global Information System) and GPS (Global Positioning System), and is widely used in various fields. Spatial events such as "disease mortality observed in each region" and "measurements of toxic substances in each observatory" may occur centrally in a specific area. At this time, it is called "a cluster exists".

Detection of spatial clusters is very important for understanding the current environmental conditions and future impacts. To date, as methods for evaluating the presence or absence of a cluster, for evaluating from the perspective of spatial autocorrelation (Moran 1948; Cliff and Ord 1973; Anselin 1995), and methods for testing the presence or absence of a cluster and identifying its position (Kulldorff 1997; Tango and Takahashi 2005; Ishioka et al. 2019) have been proposed. These methods have been widely used in the field of epidemiology and so on. In particular, the spatial scan statistic (Kulldorff 1997) has been widely used to detect clusters of infectious diseases such as childhood pneumonia (Andrade et al. 2004), tuberculosis (Oeltmann et al. 2008; Kammerer et al. 2013), and influenza (Manabe et al. 2016). Furthermore, Cordes and Castro (2020) detected clusters of coronavirus disease 2019 (COVID-19) in New York City.

However, the Kulldorff's method can detect only circular clusters, and it is difficult to detect clusters with arbitrary shapes. Additionally, the spatial scan statistic based on the idea of maximizing the likelihood has the problem of including low-risk regions in clusters. As a result, large clusters with unrealistic shapes may be detected. For this problem, Tango (2008) proposed the spatial scan statistic with restricted likelihood ratio to detect clusters including only the regions with high-risk. As methods based on this statistic, the restricted circular scan method (Tango 2008) and the restricted flexible scan method (Tango and Takahashi 2012) were proposed. On the other hand, even with these methods, when targeting large-scale spatial data with a number of regions ranging from thousands to tens of thousands, the amount of calculation becomes enormous and it is difficult to detect clusters with arbitrary shapes.

In general, clusters are often detected for the cumulative number of observations made in a specific period within the study area. It is also important to simultaneously detect the location and duration of clusters for the number of observations that span multiple

periods, such as the daily number of people infected with infectious disease. Such a cluster that has information on the place and period of occurrence is called a space-time cluster. Kulldorff et al. (1998) proposed a method for detecting space-time clusters on the basis of the spatial scan statistic. This method can be performed with the SaTScan^{TM} software (the latest version is 10.0.2; Kulldorff, 2022). Recent studies using Kulldorff's method include the detection of space-time clusters for COVID-19 infections (Hohl et al. 2020; Kim and Castro 2020; Martines et al. 2021).

Detection of space-time clusters has made it possible to detect when and where clusters have occured. However, Kulldorff's method can only detect a cluster comprising the same regional group that spans multiple periods. Accordingly, this method cannot capture changes in a cluster's shape over time (Patil and Taillie 2004). Furthermore, when considering changes in the shape of a cluster, an analysis using existing methods becomes difficult because of an increase in the calculation amount.

This paper is roughly composed of two parts. In the first half of this paper, we propose a new method that can detect high-risk clusters. The proposed method uses the criteria defined by Tango in the spatial scan statistic with restricted likelihood ratio to extract the upper hierarchy of the spatial hierarchical structure obtained by echelon analysis, and scans the extracted data to detect high-risk spatial clusters. We also examine the possibility of application to large-scale spatial data through simulation.

In the second half, as an application of the proposed method to spatiotemporal data, we detect space-time clusters using data of infected people with COVID-19 collected daily by each prefecture in Japan. In addition, we consider the factors that caused the detected clusters and the changes in their shapes.

2 The spatial scan statistic

The spatial scan statistic is a likelihood ratio test statistic for evaluating the presence or absence of clusters in a study area. Let us assume that a study area $\mathbf{G}$ is divided into m regions. It is also assumed that the random variables, O_i, which represent the observed number in region i, follow the Poisson distribution independently of one another. At this time, if there is no cluster in the study area, the random variable O_i with the observed value o_i can be stated as follows:

$$O_i \sim \text{Poisson}(\xi_i), \quad i = 1, 2, \ldots, m$$

where ξ_i is the expected number of cases region i. In addition, a subset of regions adjacent to each other in $\mathbf{G}$ is called *a window* and represented by $\mathbf{Z}$, and the complement of $\mathbf{Z}$ is replesented by $\mathbf{Z}^c$. Let $\mathcal{Z}$ be the universal set of $\mathbf{Z}$, then, the presence or absence of a cluster can then be given as the following hypothesis testing:

$$\begin{aligned} H_0 &: p_{\mathbf{Z}} = p_{\mathbf{Z}^c} = p, \quad {}^{\forall}\mathbf{Z} \in \mathcal{Z} \\ H_1 &: p_{\mathbf{Z}} > p_{\mathbf{Z}^c}, \qquad {}^{\exists}\mathbf{Z} \in \mathcal{Z} \end{aligned}$$

where, $p_{\mathbf{Z}}$ and $p_{\mathbf{Z}^c}$ are the probability of event occurrence in $\mathbf{Z}$ and in $\mathbf{Z}^c$, respectively. However, performing the test for each $\mathbf{Z}$ gives rise to the problem of conducting multiple testing.

Let o_i and w_i be the number of cases in region i and the number of populations, respectively. The number of cases in $\mathbf{Z}$ and $\mathbf{Z}^c$ can be expressed as $o(\mathbf{Z}) = \sum_{i \in \mathbf{Z}} o_i$ and $o(\mathbf{Z}^c) = \sum_{i \notin \mathbf{Z}} o_i$. Similarly, the number of populations in $\mathbf{Z}$ and $\mathbf{Z}^c$ can be expressed as $w(\mathbf{Z}) = \sum_{i \in \mathbf{Z}} w_i$ and $w(\mathbf{Z}^c) = \sum_{i \notin \mathbf{Z}} w_i$. Here, $o(\mathbf{G}) = o(\mathbf{Z}) + o(\mathbf{Z}^c) = \sum_{i=1}^{m} o_i$ and $w(\mathbf{G}) = w(\mathbf{Z}) + w(\mathbf{Z}^c) = \sum_{i=1}^{m} w_i$. At this time, using the Poisson distribution with mean $\xi_i = w_i p$, the likelihood ratio (LR) under null and alternative hypothesis is given as follows:

$$\begin{aligned} LR(\mathbf{Z}, p_{\mathbf{Z}}, p_{\mathbf{Z}^c}, p) &= \frac{\text{the likelihood under } H_1}{\text{the likelihood under } H_0} \\ &= \frac{\exp(-\sum_{i \in \mathbf{Z}} w_i p_{\mathbf{Z}}) \frac{\prod_{i \in \mathbf{Z}} (w_i p_{\mathbf{Z}})^{o_i}}{\prod_{i \in \mathbf{Z}} o_i!} \times \exp(-\sum_{i \notin \mathbf{Z}} w_i p_{\mathbf{Z}^c}) \frac{\prod_{i \notin \mathbf{Z}} (w_i p_{\mathbf{Z}^c})^{o_i}}{\prod_{i \notin \mathbf{Z}} o_i!}}{\exp(-\sum_{i=1}^{m} w_i p) \frac{\prod_{i=1}^{m} (w_i p)^{o_i}}{\prod_{i=1}^{m} o_i!}} \\ &= \frac{\exp(-\sum_{i \in \mathbf{Z}} w_i p_{\mathbf{Z}} - \sum_{i \notin \mathbf{Z}} w_i p_{\mathbf{Z}^c}) \times p_{\mathbf{Z}}^{o(\mathbf{Z})} \times p_{\mathbf{Z}^c}^{o(\mathbf{Z}^c)}}{\exp(-\sum_{i=1}^{m} w_i p) \times p^{o(\mathbf{G})}}. \end{aligned} \tag{2.1}$$

For this $LR(\mathbf{Z}, p_{\mathbf{Z}}, p_{\mathbf{Z}^c}, p)$, by substituting the makimum likelihood estimators $\hat{p}_{\mathbf{Z}} = \frac{o(\mathbf{Z})}{w(\mathbf{Z})}$, $\hat{p}_{\mathbf{Z}^c} = \frac{o(\mathbf{Z}^c)}{w(\mathbf{Z}^c)}$, $\hat{p} = \frac{o(\mathbf{G})}{w(\mathbf{G})}$, we obtain the following maximum likelihood function:

$$LR(\mathbf{Z}) = \frac{\left(\frac{o(\mathbf{Z})}{w(\mathbf{Z})}\right)^{o(\mathbf{Z})} \left(\frac{o(\mathbf{Z}^c)}{w(\mathbf{Z}^c)}\right)^{o(\mathbf{Z}^c)}}{\left(\frac{o(\mathbf{G})}{w(\mathbf{G})}\right)^{o(\mathbf{G})}} I\left(\frac{o(\mathbf{Z})}{w(\mathbf{Z})} > \frac{o(\mathbf{Z}^c)}{w(\mathbf{Z}^c)}\right), \tag{2.2}$$

where, since we assume $H_1 : p_{\mathbf{Z}} > p_{\mathbf{Z}^c}$, it is $I\left(\frac{o(\mathbf{Z})}{w(\mathbf{Z})} > \frac{o(\mathbf{Z}^c)}{w(\mathbf{Z}^c)}\right)$. Furthermore, $LR(\mathbf{Z})$ can be converted as follows:

$$LR(\mathbf{Z}) = \frac{\left(\frac{o(\mathbf{Z})}{w(\mathbf{Z})}\right)^{o(\mathbf{Z})} \left(\frac{o(\mathbf{Z}^c)}{w(\mathbf{Z}^c)}\right)^{o(\mathbf{Z}^c)}}{\left(\frac{o(\mathbf{G})}{w(\mathbf{G})}\right)^{o(\mathbf{Z})} \left(\frac{o(\mathbf{G})}{w(\mathbf{G})}\right)^{o(\mathbf{Z}^c)}} I\left(\frac{o(\mathbf{Z})}{w(\mathbf{Z})} > \frac{o(\mathbf{Z}^c)}{w(\mathbf{Z}^c)}\right). \tag{2.3}$$

Let $\xi(\mathbf{Z})$ and $\xi(\mathbf{Z}^c)$ be the expected number of cases inside and outside $\mathbf{Z}$, respectively. Using $\xi_i = w_i \cdot \frac{o(\mathbf{G})}{w(\mathbf{G})}$, these are given by following equations:

$$\xi(\mathbf{Z}) = \sum_{i \in \mathbf{Z}} w_i \cdot \frac{o(\mathbf{G})}{w(\mathbf{G})} = w(\mathbf{Z}) \cdot \frac{o(\mathbf{G})}{w(\mathbf{G})},$$

$$\xi(\mathbf{Z}^c) = \sum_{i \notin \mathbf{Z}} w_i \cdot \frac{o(\mathbf{G})}{w(\mathbf{G})} = w(\mathbf{Z}^c) \cdot \frac{o(\mathbf{G})}{w(\mathbf{G})}.$$

Substituting $\xi(\mathbf{Z})$ and $\xi(\mathbf{Z}^c)$, $LR(\mathbf{Z})$ can be represented by the following equation:

$$\begin{aligned} LR(\mathbf{Z}) &= \frac{\left(\frac{o(\mathbf{Z})}{w(\mathbf{Z})}\right)^{o(\mathbf{Z})} \left(\frac{o(\mathbf{Z}^c)}{w(\mathbf{Z}^c)}\right)^{o(\mathbf{Z}^c)}}{\left(\frac{\xi(\mathbf{Z})}{w(\mathbf{Z})}\right)^{o(\mathbf{Z})} \left(\frac{\xi(\mathbf{Z}^c)}{w(\mathbf{Z}^c)}\right)^{o(\mathbf{Z}^c)}} I\left(\frac{o(\mathbf{Z})}{\xi(\mathbf{Z})} > \frac{o(\mathbf{Z}^c)}{\xi(\mathbf{Z}^c)}\right) \\ &= \left(\frac{o(\mathbf{Z})}{\xi(\mathbf{Z})}\right)^{o(\mathbf{Z})} \left(\frac{o(\mathbf{Z}^c)}{\xi(\mathbf{Z}^c)}\right)^{o(\mathbf{Z}^c)} I\left(\frac{o(\mathbf{Z})}{\xi(\mathbf{Z})} > \frac{o(\mathbf{G}) - o(\mathbf{Z})}{o(\mathbf{G}) - \xi(\mathbf{Z})}\right) \quad (\because \xi(\mathbf{G}) = o(\mathbf{G})) \\ &= \left(\frac{o(\mathbf{Z})}{\xi(\mathbf{Z})}\right)^{o(\mathbf{Z})} \left(\frac{o(\mathbf{Z}^c)}{\xi(\mathbf{Z}^c)}\right)^{o(\mathbf{Z}^c)} I\left(o(\mathbf{Z}) > \xi(\mathbf{Z})\right). \end{aligned} \tag{2.4}$$

Therefore, the spatial scan statistic is given by the following equation:

$$\lambda_K(\mathbf{Z}) = \begin{cases} \left(\frac{o(\mathbf{Z})}{\xi(\mathbf{Z})}\right)^{o(\mathbf{Z})} \left(\frac{o(\mathbf{Z}^c)}{\xi(\mathbf{Z}^c)}\right)^{o(\mathbf{Z}^c)}, & (o(\mathbf{Z}) > \xi(\mathbf{Z})) \\ 1. & (\text{otherwise}) \end{cases} \tag{2.5}$$

Window $\mathbf{Z}$ that maximizes $\lambda_K(\mathbf{Z})$ is defined as the most likely cluster (MLC). Typically, $\log \lambda_K(\mathbf{Z})$ is used to simplify the calculation. The significance of the MLC is evaluated using the Monte Carlo method.

3 The method for detecting spatial clusters

3.1 The circular scan method

In detecting spatial clusters, how to find the window $\mathbf{Z}$ that maximizes the statistic $\lambda_K(\mathbf{Z})$ is important. In general, the number of combinations of regions included in $\mathbf{Z}$ is enormous, and it is not realistic to find all of them. Therefore, it is necessary to find $\mathbf{Z}$ efficiently. As the method for scanning $\mathbf{Z}$, Kulldorff (1997) proposed the circular scan method. In this method, as shown in Fig. 3.1, expand the circle centered on the representative point of region i until it reaches the maximum spatial window size (MSWS) set in advance. The MSWS is a parameter related to the size of the cluster determined by the analyst and expressed as K. As the MSWS, the maximum number of regions or populations in the cluster are used. In this chapter, the MSWS is the maximum number of regions included in a cluster. At this time, the regions included inside the expanding circle are sequentially taken into $\mathbf{Z}$. That is, the circular scan method obtains a window $\mathbf{Z}_{ik}$ containing region i itself and consisting of k regions in order from i. Therefore, the universal set of $\mathbf{Z}_{ik}$ is given by the following equation:

$$\mathcal{Z}_1 = \{\mathbf{Z}_{ik} \mid 1 \leq i \leq m, 1 \leq k \leq K\}, \tag{3.1}$$

here, when there are multiple regions with the same distance from the region i, if the number of regions in $\mathbf{Z}_{ik}$ does not exceed K, they are simultaneously included in $\mathbf{Z}_{ik}$. The circular scan method has high detection accuracy when the true cluster is circular-shaped. On the other hand, it is not suitable for detecting non-circular clusters such as linear and ring-shaped.

3.2 The flexible scan method

For the proplem of the circular scan method, which makes it difficult to detect non-circular clusters, Tango and Takahashi (2005) proposed the flexible scan method to detect

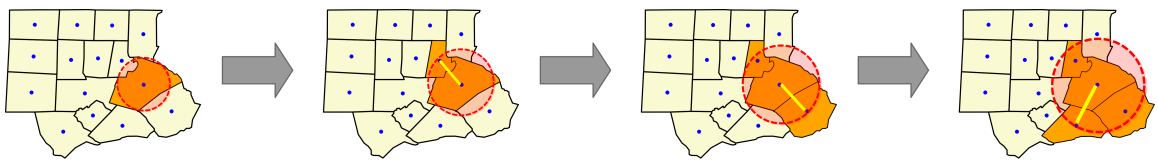


Figure 3.1: Scanning process of the circular scan method

clusters with arbitrary shapes. In this method, first, we consider a set $\mathbf{Z}_{iK}$ consisting of K regions in order from i, centering on i. At this time, if there are u_{ik} subsets in $\mathbf{Z}_{iK}$ consisting of k regions adjacent to each other including i, and let $\mathbf{Z}_{ik(u)}$ be the u-th window, then the universal set of $\mathbf{Z}_{ik(u)}$ is given by the following equation:

$$\mathcal{Z}_2 = \{\mathbf{Z}_{ik(u)} \mid 1 \leq i \leq m, 1 \leq k \leq K, 1 \leq u \leq u_{ik}\}. \tag{3.2}$$

A feature of this method is that it uses not only the coordinates of the representative points but also the adjacent information between regions, which is not used in the circular scan method. Therefore, it can scan all windows including i itself within a certain range and detect clusters with shapes that are difficult to detect by the circular scan method. However, the problem with this method is that the number of windows obtained by scanning is very large, and the analysis time tends to be long. For this reason, in the software FleXScan v3.1.2 (Takahashi et al. 2010) that implements the flexible scan method, it is recommended to set the value of the MSWS with $K \leq 20$. Hence, it is not suitable for large-scale spatial data with a large number of regions.

4 The spatial scan statistic with restricted likelihood ratio

When detecting clusters using data such as disease mortality and crime occurrence, it is desirable that the region i included in the window $\mathbf{Z}$ is a region with a high risk of satisfying $o_i > \xi_i$. However, since Eq. (2.5) is calculated based on $\mathbf{Z}$, which is a set of regions i, unrealistic results sometimes occur, such as detecting $\mathbf{Z}$ including region i where $o_i < \xi_i$. For such a problem, Tango (2008) proposed the spatial scan statistic with a restricted likelihood ratio given by

$$\lambda_T(\mathbf{Z}) = \begin{cases} \left(\dfrac{o(\mathbf{Z})}{\xi(\mathbf{Z})}\right)^{o(\mathbf{Z})} \left(\dfrac{o(\mathbf{Z}^c)}{\xi(\mathbf{Z}^c)}\right)^{o(\mathbf{Z}^c)}, & (o(\mathbf{Z}) > \xi(\mathbf{Z}), p_i < \alpha, {}^{\forall} i \in \mathbf{Z}) \\ 1, & (\text{otherwise}) \end{cases} \tag{4.1}$$

where p_i is the one-tailed p value of the test for null hypothesis given by the $mid-p$ value

$$p_i = \Pr\{O_i \geq o_i + 1 \mid O_i \sim \text{Pois}(\xi_i)\} + \frac{1}{2}\Pr\{O_i = o_i \mid O_i \sim \text{Pois}(\xi_i)\} \tag{4.2}$$

and α is the prespecified significance level for the individual region. For the significance level is 0.05, Tango (2008) defined the setting of α as follows:

1. $\alpha = 0.10 - 0.20$ to detect small clusters with a sharp increase in risk;
2. $\alpha = 0.20 - 0.30$ to detect small to mid-sized clusters with a moderate increase in risk;
3. $\alpha = 0.30 - 0.40$ to detect larger clusters with a slight increase in risk.

As a guide, Tango (2008) recommended $\alpha = 0.20$. Tango's statistic considers each region's risk rate, thereby enabling including only the regions that satisfy $o_i > \xi_i$ into the MLC.

We show a simple example of the difference in analysis results between Kulldorff's statistic and Tango's statistic. We assumed that the study area $\mathbf{G}$ shown in Fig. 4.1 has $o(\mathbf{G}) = 980$ and $\xi(\mathbf{G}) = 980$, respectively. Furthermore, consider the regions a, b, and c in $\mathbf{G}$, and assume that they have the values shown in Table 4.1. Here, from the value of $\theta = o/\xi$, a and b are high-risk regions ($\theta_a = 1.91$, $\theta_b = 1.49$), and c is a low-risk region ($\theta_c = 0.95$). In addition, it is assumed that a and b are not directly adjacent to each other, and these three regions are connected by c (see Fig. 4.1). At this time, consider detecting a cluster from the four subsets $\{a\}$, $\{b\}$, $\{c\}$ and $\{a, b, c\}$. For the set of regions

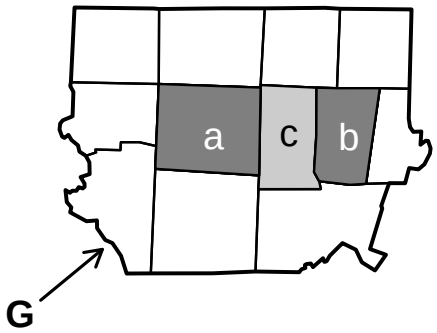


Figure 4.1: The study area **G** and the location of regions a, b and c

$\mathbf{Z} = \{a, b, c\}$ and its outside, o and ξ are calculated by the following equations:

$$
\begin{aligned}
o(\mathbf{Z}) &= 153 + 208 + 57 = 418, \\
\xi(\mathbf{Z}) &= 80 + 140 + 60 = 280, \\
o(\mathbf{Z}^c) &= o(\mathbf{G}) - o(\mathbf{Z}) = 980 - 418 = 562, \\
\xi(\mathbf{Z}^c) &= \xi(\mathbf{G}) - \xi(\mathbf{Z}) = 980 - 280 = 700.
\end{aligned}
$$

Therefore, $\log \lambda_K(\mathbf{Z}) = 44.09$, and $\{a, b, c\}$ has the largest statistic among them. However, since $\{a, b, c\}$ includes c, it is not suitable when considering a cluster consisting only of high-risk regions. On the other hand, when Tango's statistic is applied with $\alpha = 0.20$, c is not included in the cluster because $p_c < \alpha$. Hence, $\{a\}$, which has the next largest statistic, becomes a cluster.

Tango's statistic can be applied to existing methods based on Kulldorff's statistic. As a method applied to the existing method, the restricted circular scan method (Tango 2008) and the restricted flexible scan method (Tango and Takahashi 2012) have been proposed. These methods can be performed with the FleXScan software (the latest version is 3.1.2; Takahashi et al. 2013) and the rflexscan package (Otani and Takahashi 2019), which is the package of statistical analysis software R.

Table 4.1: Values of regions a, b and c

region	o	ξ	$\theta = o/\xi$	$\log \lambda_K$	p_i
a	153	80	1.91	29.25	2.10×10^{-13}
b	208	140	1.49	17.18	4.02×10^{-8}
c	57	60	0.95	0.081	0.64

5 The new method based on the echelon scan method

5.1 The echelon scan method

The echelon scan method (Ishioka et al. 2007; Ishioka et al. 2019) searches for a cluster using the hierarchical structure of the spatial data obtained by conducting echelon analysis (Myers et al. 1997; Kurihara 2004; Kurihara et al. 2020). Echelon analysis is a method for systematically and objectively visualizing the topological structure of spatial data by dividing the spatial position based on the height on the surface for the univariate value of each region. Figure 5.1 shows the flow of echelon analysis; the structure of the spatial data obtained by echelon analysis is represented by a graph called the echelon dendrogram.

As an example, suppose the 5×5 grid data shown in Fig. 5.2. Figure 5.2a shows the attribute value of each region (the attribute value of the region in the first row and column A is 2), and Fig. 5.2b shows the loation ID for each region. We defined the spatial adjacency of each region as four neighborhoods (up, down, left and right). Table 5.1 shows the neighboring information defined for each region. The echelon dendrogram shown in Fig. 5.3 is created for this grid data. The vertical axis of the dendrogram represents the attribute value of the data, and the symbols in the dendrogram denote the position of each region on the dendrogram. At this time, the echelon dendrogram in Fig. 5.3 consists of seven parts called echelons. When each echelon is expressed as $\mathrm{En}(h)(h = 1, 2, \ldots, 7)$, each region belongs to one of the echelons. For example, $\mathrm{En}(1) = \{15, 14, 9\}$. $\mathrm{En}(h)$, which does not have an echelon higher than itself, is called a peak. Therefore, in Fig. 5.3, En(1), En(2), En(3) and En(4) are peaks.

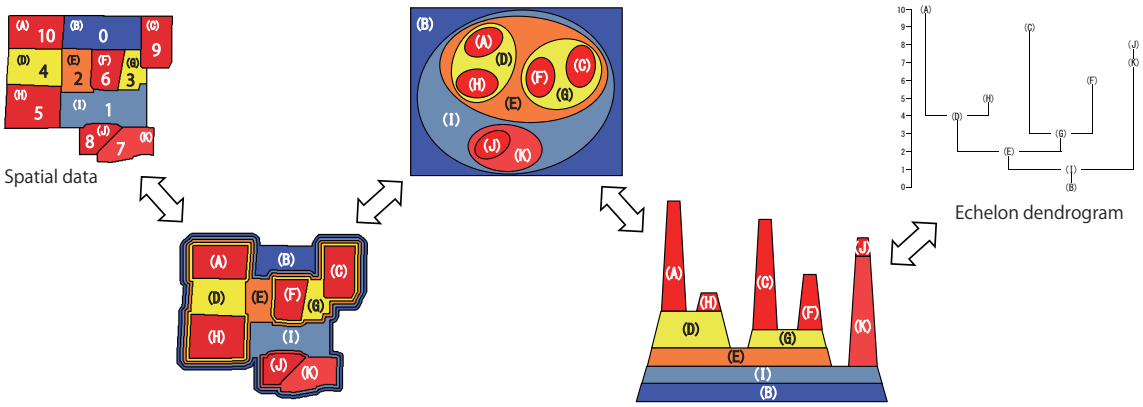


Figure 5.1: Flow in the echelon dendrogram created using echelon analysis

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

(a) Attribute value of each region

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

(b) Location ID for each region

Figure 5.2: 5×5 grid data

Table 5.1: Neighboring information of each region

Location	Neighbors	Location	Neighbors
1	2, 6	14	9, 13, 15, 19
2	1, 3, 7	15	10, 14, 20
3	2, 4, 8	16	11, 17, 21
4	3, 5, 9	17	12, 16, 18, 22
5	4, 10	18	13, 17, 19, 23
6	1, 7, 11	19	14, 18, 20, 24
7	2, 6, 8, 12	20	15, 19, 25
8	3, 7, 9, 13	21	16, 22
9	4, 8, 10, 14	22	17, 21, 23
10	5, 9, 15	23	18, 22, 24
11	6, 12, 16	24	19, 23, 25
12	7, 11, 13, 17	25	20, 24
13	8, 12, 14, 18		

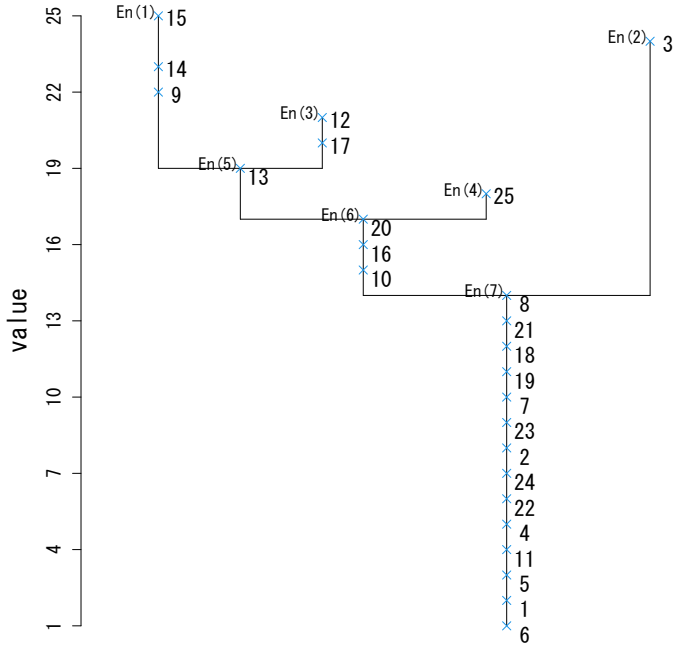


Figure 5.3: Echelon dendrogram for the grid data

The echelon scan method scans from the region included in the peak. Specifically, first, the region with ID = 15 having the maximum attribute value in En(1) is scanned as $\mathbf{Z} = \{15\}$. Next, the region with ID = 14 belonging to the same echelon is taken into $\mathbf{Z}$, and $\mathbf{Z} = \{15, 14\}$. Similarly, the region with ID = 9 is taken into $\mathbf{Z}$, and $\mathbf{Z} = \{15, 14, 9\}$. After that, the region with ID = 13 belonging to En(5) under En(1) is scanned and taken into $\mathbf{Z}$. On the other hand, En(3) = $\{12, 17\}$ exists above En(5). In this case, the regions in En(3) is also taken into $\mathbf{Z}$. That is, $\mathbf{Z} = \{15, 14, 9, 13, 12, 17\}$. The regions are scanned in this way, and the scan is performed from all peaks until the number of regions in $\mathbf{Z}$ reaches K. The general algorithm of the echelon scan method is shown in Algorithm 1. Algorithm 1 gives the window $\mathbf{Z}$, which is a candidate for MLC, and its $\log \lambda_K(\mathbf{Z})$ as the output value (O) as the result of performing the echelon scan method for the input value (I) shown in Table 5.2. Here, $LLR(Z)$ and $HV(Z)$ used in Algorithm 1 mean the value of $\log \lambda_K(\mathbf{Z})$ and the value of variable used as the MSWS.

When the echelon dendrogram has D peaks, let $\mathbf{Z}_{k(d)}$ be the window including of k regions obtained by scanning from the d-th peak. The universal set of $\mathbf{Z}_{k(d)}$ is given by follow equation:

$$\mathcal{Z}_3 = \{\mathbf{Z}_{k(d)} \mid 1 \leq k \leq K, 1 \leq d \leq D\}. \tag{5.1}$$

The echelon scan method expresses data in a hierarchical structure and preferentially scans from the region that constitutes peaks of it. Therefore, the echelon scan method

Table 5.2: Input and output values of Algorithm 1

I/O	Name	Elements
I	NP	Number of peaks
I	NE	Number of echelons
I	$NR(i)$	Number of regions included in the i-th echelon
I	$CH(EN(i))$	All regions included in the upper echelons of the i-th echelon
I	$ZE(i,j)$	The j-th region from the top of the i-th echelon
I	$MAXHV$	Value of the MSWS
O	$MAXZ$	Window $\mathbf{Z}$ for the candidate of MLC
O	$MAXLLR$	$\log \lambda_K(\mathbf{Z})$ for the candidate of MLC

Algorithm 1: Algorithm of the echelon scan method

Ensure: Find window $\mathbf{Z}$ and maximum LLR

```
MAXZ ⇐ φ
MAXLLR ⇐ −∞
i ⇐ 1
while i ≤ NE do
  j ⇐ 1
  if i ≤ NP then
    Z ⇐ ZE(i, j)
  end if
  if i > NP then
    Z ⇐ CH(EN(i)) ∪ ZE(i, j)
  end if
  while j ≤ NR(i) and HV(Z) ≤ MAXHV do
    if LLR(Z) > MAXLLR then
      MAXZ ⇐ Z
      MAXLLR ⇐ LLR(Z)
    end if
    j ⇐ j + 1
    if ZE(i, j) ≠ φ then
      Z ⇐ Z ∪ ZE(i, j)
    end if
  end while
  i ⇐ i + 1
end while
```

can reduce the calculation cost as compared with the circular scan method and the flexible scan method (Ishioka and Kurihara 2012; Ishioka et al., 2019).

5.2 The adjusted echelon scan method

The echelon scan method also scans the regions included in lower echelons of the echelon dendrogram. However, lower echelons include regions where $o_i < \xi_i$, and regions that should actually be detected as a cluster are generally included in upper echelons. Therefore, we propose the adjusted echelon scan method (AESM) as an improvement technique of Echelon scan method for detecting high-risk clusters. In the AESM, the upper hierarchies of the spatial data are extracted using p_i and Tango's α, and the echelon scan method is applied to the extracted data. Specifically, the steps below are followed in this process.

Step 1. Extract the data of region i that satisfies $p_i < \alpha$ from the analysis data.

Step 2. Apply echelon analysis to the extracted data to create an echelon dendrogram.

Step 3. The region included in the upper echelon of the dendrogram is taken into $\mathbf{Z}$ in order, and $\mathbf{Z}$, which maximizes $\log \lambda_K(\mathbf{Z})$, is the MLC.

As an example, we compare clusters detected by the echelon scan method and the AESM using data from sudden infant death syndrome (SIDS) observed between 1974 and 1984 in 100 counties in North Carolina, United States (Cressie and Chan 1989). Let w_i and o_i be the number of infants born and the number of deaths due to SIDS within the period in region i $(i = 1, 2, ..., 100)$, respectively. At this time, the expected number ξ_i of observations is calculated by the following equation:

$$\xi_i = w_i \frac{\sum_{r=1}^{100} o_r}{\sum_{r=1}^{100} w_r}. \quad i = 1, 2, \ldots, 100 \tag{5.2}$$

Furthermore, as an index indicating risk of region i, θ_i was calculated by the following equation:

$$\theta_i = \frac{o_i}{\xi_i}. \quad i = 1, 2, \ldots, 100 \tag{5.3}$$

Figure 5.4 shows the choropleth map created based on θ_i. The color of the choropleth map in the figure shows the value of θ_i. In addition, the ID of each region is shown on the map. From the choropleth map, it can be seen that high-risk regions are distributed in the south and northeast.

Figure 5.5 shows the result of applying the echelon scan method with the MSWS as $K = 50$, which corresponds to half of the total number of regions. The window $\mathbf{Z}$ that

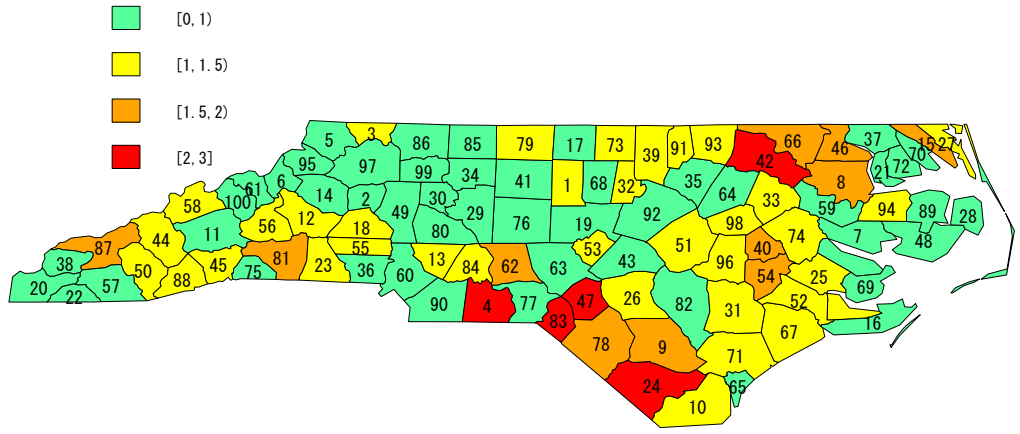


Figure 5.4: Coropleth map based on θ_i of SIDS data

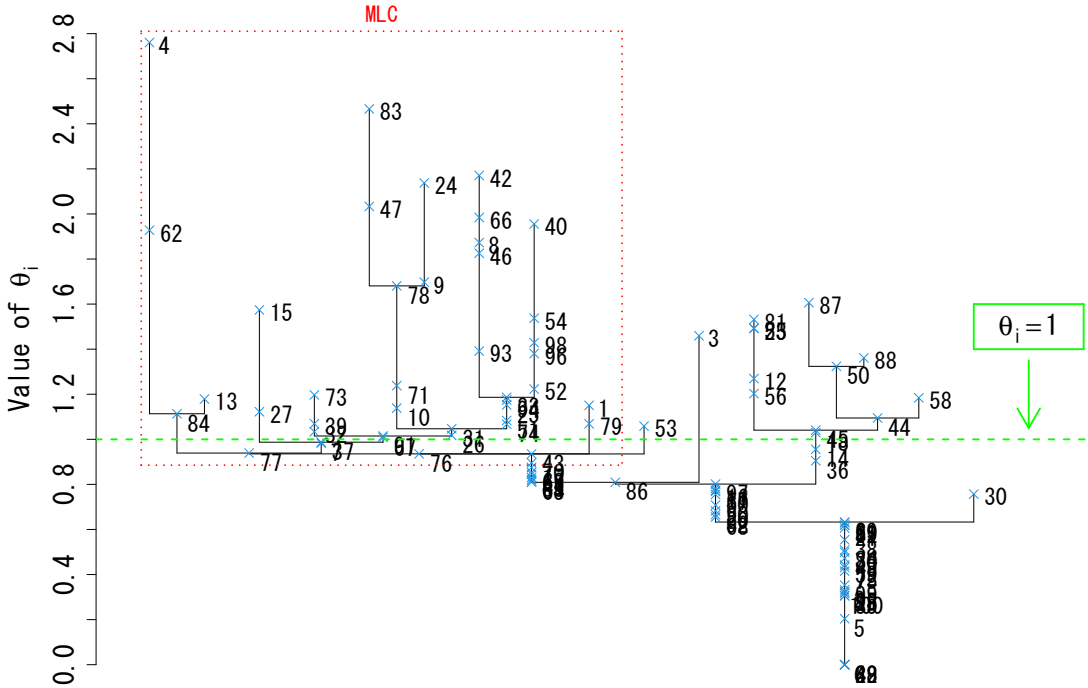


Figure 5.5: Echelon dendrogram based on θ_i of SIDS data

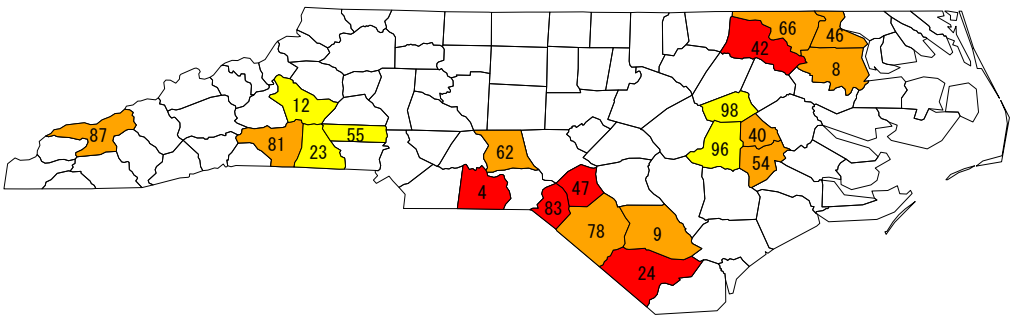


Figure 5.6: Extracted data in the AESM with $\alpha = 0.20$

maximizes $\log \lambda_K(\mathbf{Z})$ was obtained when scanning up to the part surrounded by the dotted line on the echelon dendrogram. As a result, this $\mathbf{Z}$ was MLC. MLC was formed from as many as 41 regions, including 4 low-risk regions (ID $= 7, 37, 76, 77$) with $\theta_i < 1$.

Next, cluster detection is performed using the AESM. We show the result when $\alpha = 0.20$ recommended by Tango (2008) was set. The map showing the regions that satisfy $p_i < \alpha$ from the data is shown in Fig. 5.6. Furthermore, Fig. 5.7 is the echelon dendrogram created based on the extracted regions. As is clear from Fig. 5.7, the echelon dendrogram of AESM is created only from high-risk regions. MLC and Secondary cluster obtained by scanning this dendrogram are shown by the dotted line on the dendrogram in Fig. 5.7. Here, secondary cluster is a window in which the value of $\log \lambda_K(\mathbf{Z})$ is the second highest after MLC under the condition that the regions do not overlap with MLC. Unlike the echelon scan method, the AESM detected small clusters.

Figure 5.8 and Fig. 5.9 show the maps that visualize the clusters detected by the echelon scan method and the AESM, respectively. In addition, the number of regions in each cluster, $\theta(\mathbf{Z}) = o(\mathbf{Z})/\xi(\mathbf{Z})$, which is the value of risk, and the values of $\log \lambda_K(\mathbf{Z})$ are shown in Table 5.3. From Fig. 5.8, MLC of the echelon scan method has a complicated shape, and includes the low-risk regions that do not fit the feeling. Furthermore, from Table 5.3, the value of $\log \lambda_K(\mathbf{Z})$ is as high as 43.29, but $\theta(\mathbf{Z})$ is as low as 1.28.

The echelon scan method uses a hierarchical structure of data, which makes it possible to find window $\mathbf{Z}$ with high likelihood by scanning. However, when the lower echelons are scanned, a small number of low-risk regions may concatenate regions included in the upper echelons, and as a result, a cluster having a distorted shape shown in Fig. 5.8 may be detected. On the other hand, the location of the clusters detected by the AESM shown

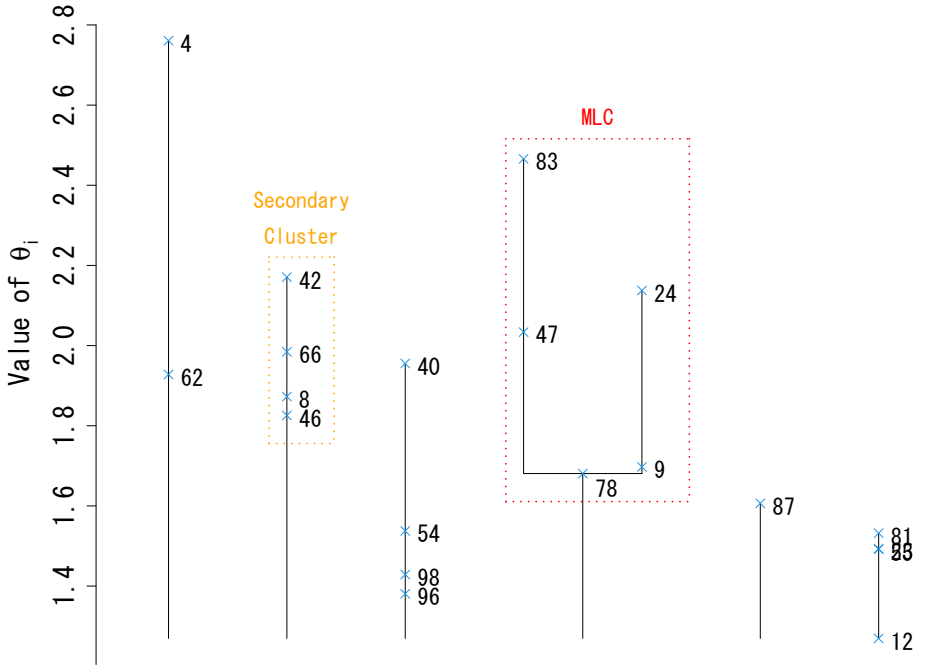


Figure 5.7: Echelon dendrogram in the AESM

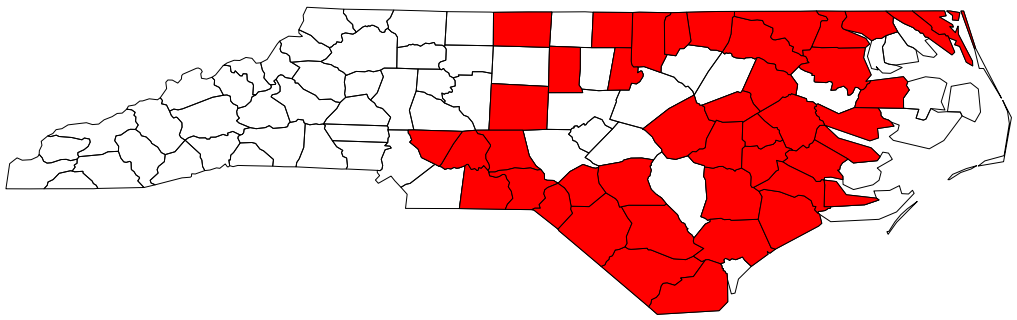


Figure 5.8: Cluster detected by the echelon scan method

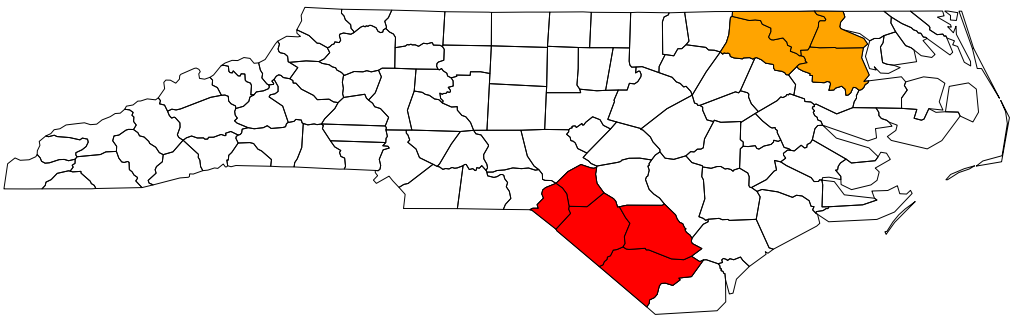


Figure 5.9: Clusters detected by the AESM

Table 5.3: Results of application to SIDS data

Methods	Cluster type	Number of regions	$\theta(\mathbf{Z})$	$\log \lambda_K(\mathbf{Z})$
Echelon scan	MLC	41	1.28	43.29
AESM	MLC	5	1.91	25.38
	Secondary cluster	4	2.02	14.34

in Fig. 5.9 is consistent with the set of high-risk regions in the choropleth map in Fig. 5.4. Table 5.3 also shows that the AESM detected high-risk clusters compared to the echelon scan method. From this, the AESM can accurately detect clusters consisting only of high-risk regions that are closer to the actual feeling.

5.3 Calculation cost of the AESM

Methods for detecting clusters based on Tango's statistic, such as the restricted flexible scan method, solve the problem of detecting non-local clusters by incorporating only the regions satisfying $p_i < \alpha$ into $\mathbf{Z}$. However, for large-scale spatial data consisting of thousands to tens of thousands of regions, even the restricted flexible scan method becomes impractical in terms of calculation cost. Therefore, in cluster detection for large-scale spatial data, it is important to detect non-local clusters having arbitrary shapes and to reduce the amount of calculation.

One of the advantages of the AESM is the reduction of computational cost by simplifying the echelon dendrogram. Comparing the echelon scan method dendrogram in

Fig. 5.5 with the AESM dendrogram in Fig. 5.7, it can be seen that the AESM has a simpler structure.

Here, we actually compare the calculation costs of the four methods of the echelon scan method, the restricted circular scan method, the restricted flexible scan method and the AESM. For the data used for verification, $L \times L$ grid data ($L = 10, 40, 70, 100$) was created, and one true cluster $\boldsymbol{R}$, which is circular-shaped, was set for each of the data. Let s be the number of regions included in $\boldsymbol{R}$, and the verification was performed in the five cases of $s = 10, 20, 30, 40, 50$. The analysis was performed 100 times each using the four methods, and the average analysis time (seconds) $\pm$ standard deviation is shown in Table 5.4. We used R fot the analysis. The echelon scan method used the echelon package (Ishioka 2020), the restricted circular scan method and the restricted flexible scan method used the rflexscan package. We created and used the new function for applying the AESM. The specifications of the PC that we used are Windows 10 Intel (R), Core (TM) i7, CPU 3.40GHz, RAM 16GB. The system.time function, which is R function, was used to measure the analysis time. The results in Table 5.4 are purely measurements of the time required for a single scan and are not Monte Carlo tested.

From Table 5.4, the analysis time of each method tends to increase as the analysis target area increases. In particular, when $s \geq 30$, the restricted flexible scan method has an extremely long analysis time compared to other methods under the same conditions. On the other hand, with the AESM, analysis can be performed in about 3 seconds even when $L = 100$ and $s = 50$, which can reduce the analysis time to about 1/10 compared to other methods. Since Monte Carlo simulation is required to test the significance of clusters, the time required for a single scan should be as short as possible. From the above, we consider that the AESM is more effective for detection of high-risk clusters for large-scale spatial data than the existing method.

Table 5.4: Comparison of analysis time. "—" indicates that the analysis time of each analysis exceeds 1000 seconds.

		L			
		10	40	70	100
Echelon scan					
	10	0.012 ± 0.0076	0.40 ± 0.078	5.05 ± 1.98	32.67 ± 12.25
	20	0.011 ± 0.0090	0.41 ± 0.092	5.78 ± 2.47	28.31 ± 14.12
s	30	0.011 ± 0.0082	0.41 ± 0.086	6.16 ± 1.81	29.55 ± 19.50
	40	0.010 ± 0.0082	0.44 ± 0.090	4.96 ± 1.73	39.98 ± 24.10
	50	0.011 ± 0.0080	0.51 ± 0.092	4.99 ± 2.14	57.43 ± 19.60
Restricted circular scan					
	10	0.0018 ± 0.0052	0.52 ± 0.045	5.65 ± 0.13	30.67 ± 0.64
	20	0.0023 ± 0.0053	0.49 ± 0.028	5.81 ± 0.65	29.22 ± 0.82
s	30	0.0028 ± 0.0067	0.48 ± 0.017	5.71 ± 0.14	29.58 ± 1.05
	40	0.0022 ± 0.0056	0.48 ± 0.010	5.85 ± 0.28	29.40 ± 0.98
	50	0.0041 ± 0.0074	0.48 ± 0.016	6.11 ± 0.32	29.77 ± 0.86
Restricted flexible scan					
	10	0.0023 ± 0.0058	0.51 ± 0.035	5.63 ± 0.13	31.71 ± 3.29
	20	0.12 ± 0.011	1.21 ± 0.78	7.75 ± 5.17	30.97 ± 7.62
s	30	14.47 ± 0.055	513.95 ± 848.19	—	—
	40	—	—	—	—
	50	—	—	—	—
AESM					
	10	0.025 ± 0.0082	0.37 ± 0.018	1.27 ± 0.041	2.96 ± 0.061
	20	0.027 ± 0.0081	0.37 ± 0.019	1.25 ± 0.038	3.01 ± 0.072
s	30	0.026 ± 0.0081	0.37 ± 0.028	1.25 ± 0.051	2.99 ± 0.075
	40	0.029 ± 0.0072	0.36 ± 0.027	1.25 ± 0.047	3.01 ± 0.079
	50	0.029 ± 0.0073	0.35 ± 0.024	1.25 ± 0.042	3.03 ± 0.083

6 Simulation for comparison of detection accuracy

6.1 Data generation

We perfom simulations using two types of area data in order to compare the cluster detection accuracy of the existing method and the AESM. As the first area data, the 10 $\times$ 10 grid data shown in Fig. 6.1 is used. The numbers in the figure are the location IDs of each region. In recent years, a lot of mesh data, which collected for each region divided into grids on the map, has been released. Furthermore, as a feature of the grid data, the shape of each region is the same and the number of adjacent regions is almost constant. Therefore, it is possible to fairly compare the detection accuracy for each method.

Next, Table 5.4 shows that the AESM is an effective method from the viewpoint of analysis time for large-scale spatial data as compared with other methods. In addition, previous studies have not shown detection accuracy in large-scale spatial data. Therefore, we compare the existing method and the AESM using the data by county in the United States in Fig. 6.2.

In this paper, we generate the observation data for simulation according to Tango (2008). We assume that one true cluster $\boldsymbol{R}$ with s regions exists in the study area $\mathbf{G}$ with m regions. The hypothesis test for the risk θ_i of the region i can be stated as follows:

$$\begin{aligned} H_0 : \theta_i &= 1, \quad {}^{\forall} i \in \mathbf{G} \\ H_1 : \theta_i &> 1, \quad {}^{\forall} i \in \boldsymbol{R} \end{aligned}$$

where θ_i is given by the following equation:

$$\theta_i = \frac{o_i}{\xi_i}, \quad i = 1, 2, \ldots, m \tag{6.1}$$

and ξ_i is calculated by the following equation:

$$\xi_i = w_i \frac{\sum_{r=1}^{m} o_r}{\sum_{r=1}^{m} w_r}. \quad i = 1, 2, \ldots, m \tag{6.2}$$

Under the alternative hypothesis, we generate a random sample $(o_1, o_2, \ldots, o_m)$ of size o from the multinomial distribution with parameters $(q_1, q_2, \ldots, q_m)$ where

$$q_i = \frac{\pi_i w_i}{\sum_{r=1}^{m} \pi_r w_r}, \quad i = 1, 2, \ldots, m \tag{6.3}$$

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Figure 6.1: 10×10 grid data

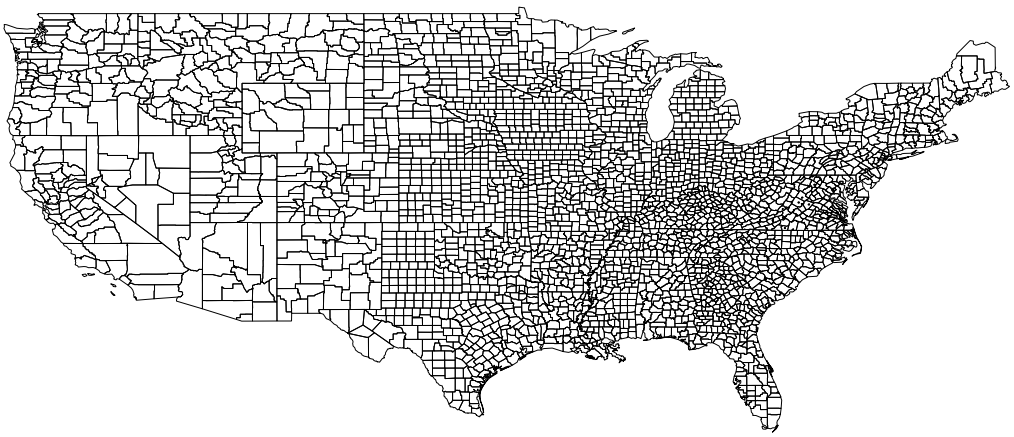


Figure 6.2: The county data in the United States

where

$$\pi_i = \begin{cases} \theta_i, & {}^\forall i \in \boldsymbol{R} \\ 1. & \text{otherwise.} \end{cases}$$

Under these conditions, we generate N simulation data sets and compare the detection accuracy of each method.

6.2 Evaluation index

We use Sensitivity and PPV (Positive Prediction Value) proposed by Huang et al. (2007) as evaluation indexes for the detection accuracy of each method. Each index is defined by the following equation:

$$\text{Sensitivity} = \frac{\text{Length}(\mathbf{Z}_{\text{MLC}} \cap \boldsymbol{R})}{s}, \tag{6.4}$$

$$\text{PPV} = \frac{\text{Length}(\mathbf{Z}_{\text{MLC}} \cap \boldsymbol{R})}{\text{Length}(\mathbf{Z}_{\text{MLC}})}, \tag{6.5}$$

where $\mathbf{Z}_{\text{MLC}}$ is a window that becomes MLC, and Length$(\cdot)$ means the number of regions. Both indexes take a value from 0 to 1, and the closer the value is to 1, the true cluster $\boldsymbol{R}$ can be detected more accurately. In this paper, in order to visually express these indexes, we propose a new scatter plot shown in Fig. 6.3. In Fig. 6.3, Sensitivity is taken on the horizontal axis and PPV is taken on the vertical axis, and the points where $(x, y) = (\text{Sensitivity}, \text{PPV})$ are plotted. Depending on the result of the simulation, points may be plotted at the same coordinates. Therefore, the number of data contained in each point is expressed by the size and color shading of each point. If many points are plotted on or near the intersection of the straight lines of Sensitivity $= 1$ and PPV $= 1$, it shows that the detection accuracy is high. In addition, let S_j and P_j be the values of Sensitivity and PPV in the j-th simulation data set, respectively, and we define index of cluster detection accuracy (ICDA) given by the following equation:

$$\text{ICDA} = \frac{1}{N} \sum_{j=1}^{N} \frac{2 \times S_j \times P_j}{S_j + P_j} \tag{6.6}$$

6.3 Analysis of grid data

6.3.1 Circular-shaped cluster

Based on Tango and Takahashi (2012), for 10×10 grid data, we assume the circular-shaped true cluster $\boldsymbol{R}_1$ (number of cluster regions $s = 20$) shown in Fig. 6.4a, and compare

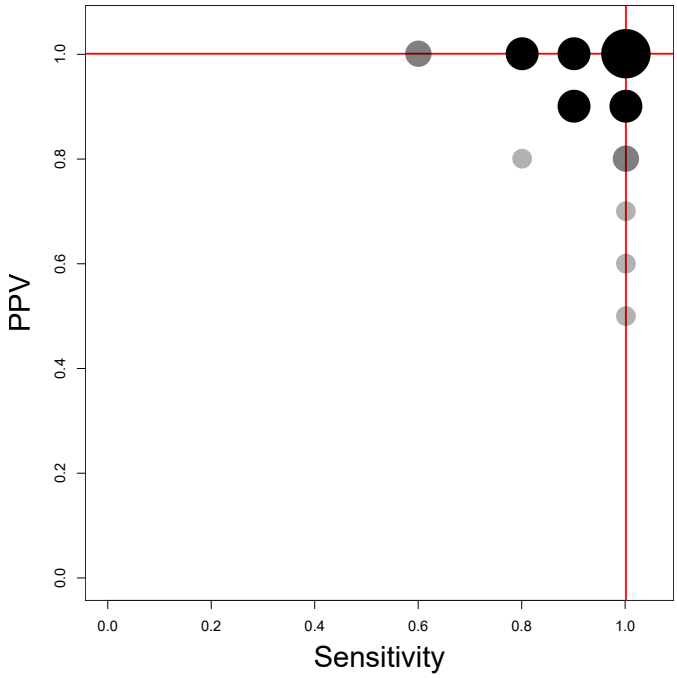
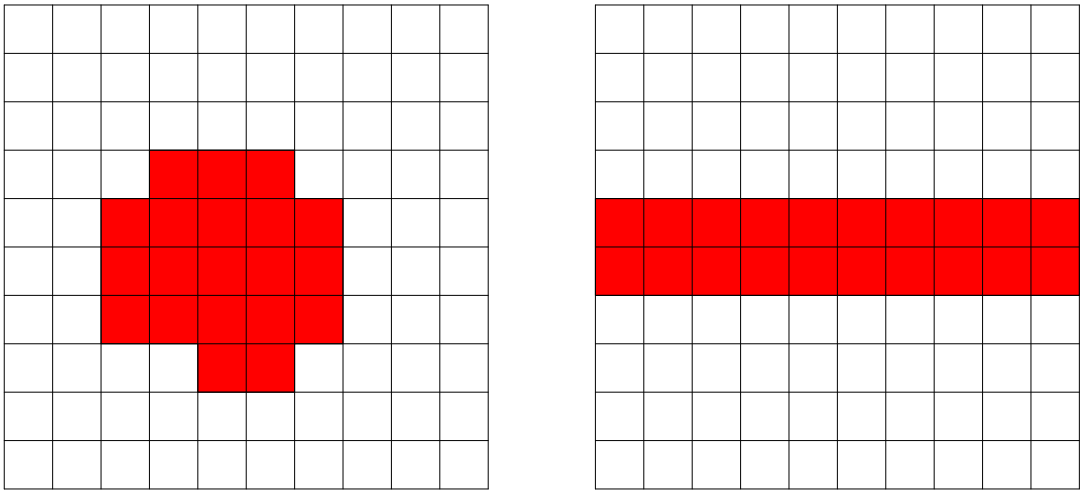


Figure 6.3: Example of a scatter plot of Sensitivity and PPV

detection accuracy of the echelon scan method, the restricted circular scan method, the restricted flexible scan method and the AESM. Under conditions of $E[\theta_{i\in \boldsymbol{R}_1}] = 1.7$ and $E[\theta_{i\notin \boldsymbol{R}_1}] = 1.0$, $N = 1000$ sets of simulation data were generated according to the procedure described in Sect. 6.1. We set to $o(\mathbf{G}) = 2708$ based on the mortality rate of common diseases such as cancer. Figure 6.5 and Table 6.1 show the analysis results when the MSWS of each method is half of the total number of regions $m/2 = 50$, and $\alpha = 0.20$. Figure 6.5 shows that the distribution of points varies in each method. Since the PPV tends to be low in the echelon scan method, we consider that many regions other than $\boldsymbol{R}_1$ were actually taken into MLC. On the other hand, in the restricted circular scan method, PPV = 1 in many cases, and Sensitivity tends to be low. Therefore, we consider that this method could only partially detect $\boldsymbol{R}_1$. In addition, since there are few points near $(\text{Sensitivity}, \text{PPV}) = (1, 1)$, we get the impression that the detection accuracy is low overall. Next, the restricted flexible scan method and the AESM had the same detection accuracy. The tendency of these methods to decrease Sensitivity is common to the restricted circular scan method. However, since many points are distributed near $(\text{Sensitivity}, \text{PPV}) = (1, 1)$, we consider that the detection accuracy of two methods is high overall.

Figure 6.6 shows the ICDA of each method at $\alpha = 0.10, 0.20, 0.30$ when the value of $E[\theta_{i\in \boldsymbol{R}_1}]$ is changed from 1.1 to 2.0 in 0.1 increments. When $\alpha = 0.10$, the ICDA of the echelon scan method is the highest at any risk. However, by setting the value of α high,



(a) $\boldsymbol{R}_1$ **(b)** $\boldsymbol{R}_2$

Figure 6.4: Assumed true clusters $\boldsymbol{R}_1$, $\boldsymbol{R}_2$

Table 6.1: Comparison of ICDA of each method ($K = m/2$, $\alpha = 0.20$, $E[\theta_{i \in \boldsymbol{R}}] = 1.7$)

Methods	$\boldsymbol{R}_1$	$\boldsymbol{R}_2$
Echelon scan	0.861	0.853
Restricted circular scan	0.470	0.691
Restricted flexible scan	0.844	0.815
AESM	0.844	0.870

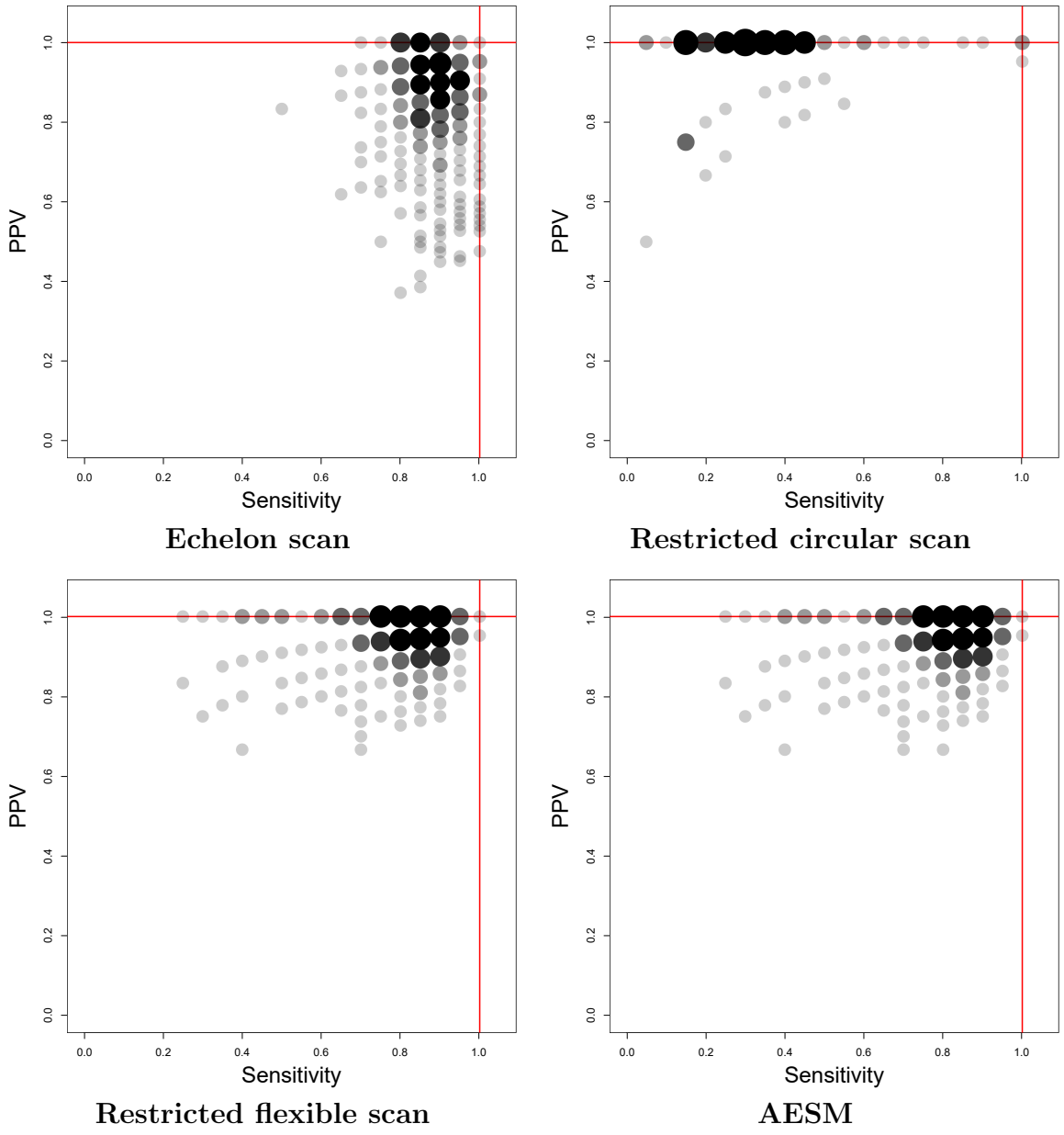


Figure 6.5: Comparison of the detection accuracy for each risk in each method assuming $\boldsymbol{R}_1$

the ICDA of the restricted flexible scan method and the proposed method improved, and at $\alpha = 0.30$, the ICDA was equivalent to that of the echelon scan method when $E[\theta_{i\in \boldsymbol{R}_1}] \geq 1.3$. On the other hand, the restricted circular scan method has a lower ICDA than other methods, even though it assumed a circular-shaped cluster.

6.3.2 Linear-shaped cluster

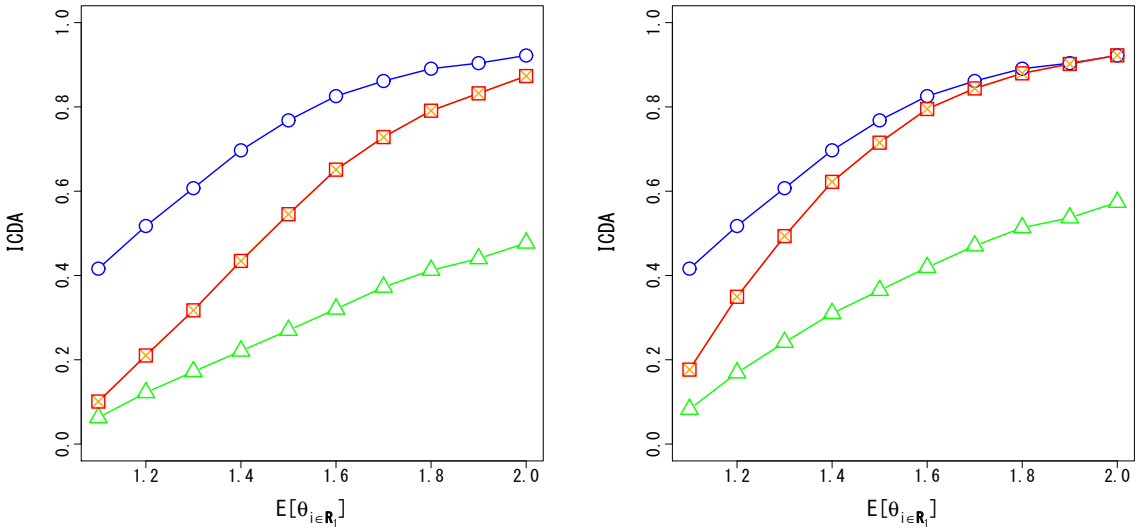
In the analysis of actual data, clusters may occur along roads, railroad tracks or coastlines. In such a case, a long linear-shaped cluster is formed in one direction. However, it may be difficult to detect by the restricted circular scan method or the restricted flexible scan method due to the nature of scan method. Therefore, we assume the linear true cluster $\boldsymbol{R}_2$ shown in Fig. 6.4b, and compare the detection accuracy of each method. Since we consider that the restricted circular scan method and the restricted flexible scan method can detect the linear-shaped cluster by dividing it into multiple clusters, here, we compare detection accuracy including significant secondary clusters. The risks in $\boldsymbol{R}_2$, N and $o(\mathbf{G})$, were set as in the case of the circular-shaped cluster ($\boldsymbol{R}_1$). Figure 6.7 and Table 6.1 show the analysis results when the MSWS of each method is half of the total number of regions $m/2 = 50$, and $\alpha = 0.20$. As for the detection accuracy of the echelon scan method, PPV was lower than that of other methods as in the case of assuming $\boldsymbol{R}_1$, and Sensitivity also tended to be slightly lower. On the other hand, the restricted circular scan method tends to have many points distributed around Sensitivity = 0.5, we consider that linear-shaped cluster cannot be sufficiently detected. The restricted flexible scan method and AESM have similar distributions, however, the result shows that Sensitivity of AESM is slightly higher.

Figure 6.8 shows the ICDA of each method at $\alpha = 0.10, 0.20, 0.30$ when the value of $E[\theta_{i\in \boldsymbol{R}_2}]$ is changed from 1.1 to 2.0 in 0.1 increments. The ICDA of the echelon scan method is not much different from the case of assuming $\boldsymbol{R}_1$. In the restricted circular scan method, the restricted flexible scan method and the AESM, the ICDA changed depending on the value of α. In particular, the AESM showed higher ICDA than that of the Echelon scan method at $E[\theta_{i\in \boldsymbol{R}_2}] \geq 1.4$ and $\alpha = 0.30$.

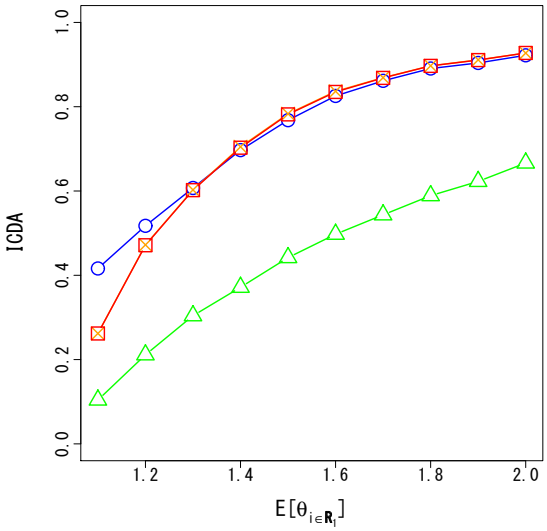
6.4 Analysis of data by county in the United States

6.4.1 Circular-shaped cluster

In this section, we verify the effectiveness of AESM for large-scale spatial data. First, for the data by county in United States ($m = 3085$), we assume the circular-shaped cluster $\boldsymbol{R}_3$ shown in Fig. 6.9a, and compare detection accuracy of the existing method and the AESM. For the number of regions that consist $\boldsymbol{R}_3$, we set to $s = 100$ because we can



$\alpha = 0.10$ $\alpha = 0.20$



$\alpha = 0.30$

Figure 6.6: Comparison of ICDA for each risk in each method assuming $\boldsymbol{R}_1$. "◯" represents the echelon scan method, "△" represents the restricted circular scan method, "×" represents the restricted flexible scan method and "□" represents the AESM.

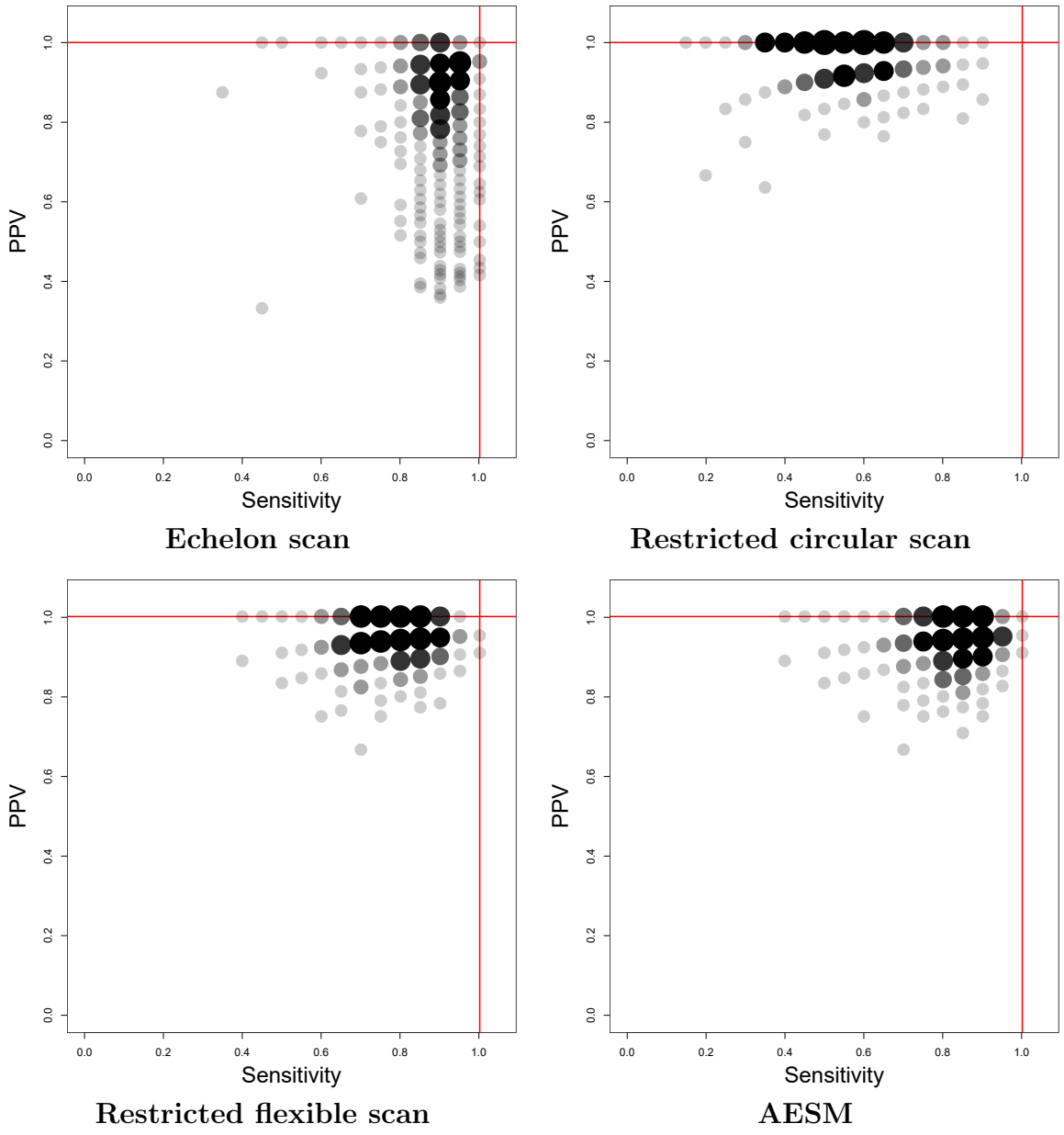
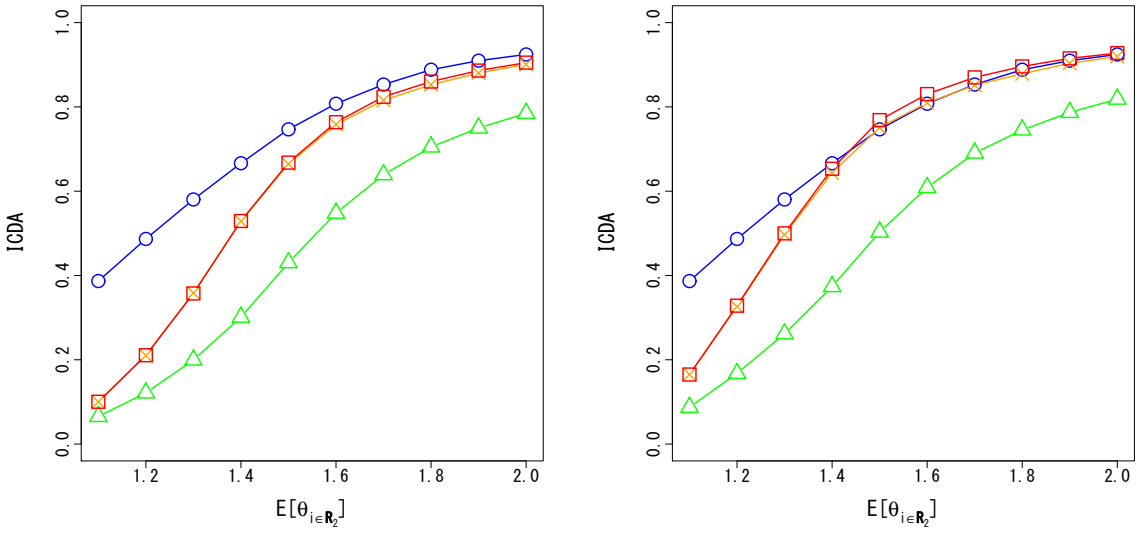
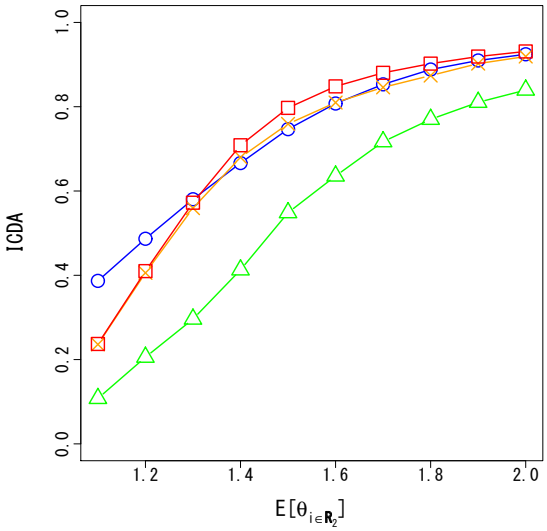


Figure 6.7: Comparison of the detection accuracy for each risk in each method assuming $\boldsymbol{R}_2$



$\alpha = 0.10$ $\alpha = 0.20$



$\alpha = 0.30$

Figure 6.8: Comparison of ICDA for each risk in each method assuming $\boldsymbol{R}_2$. "◯" represents the echelon scan method, "△" represents the restricted circular scan method, "×" represents the restricted flexible scan method and "□" represents the AESM.

Table 6.2: Comparison of ICDA of each method ($K = m/2$, $\alpha = 0.20$, $E[\theta_{i\in\boldsymbol{R}}] = 1.5$)

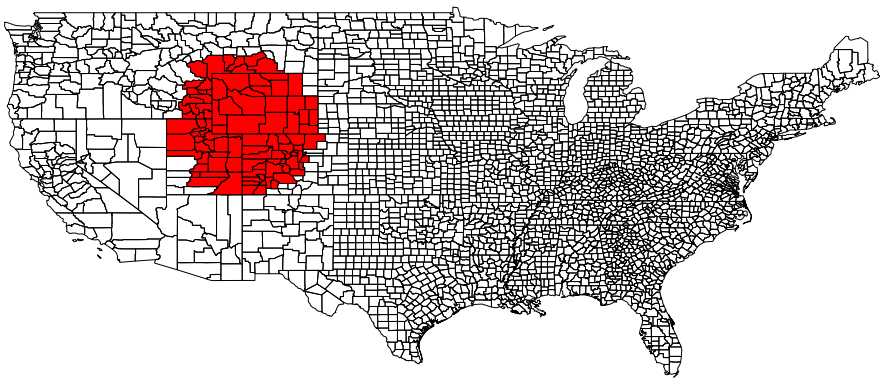
Methods	$\boldsymbol{R}_3$	$\boldsymbol{R}_4$
Echelon scan	0.879	0.891
Restricted circular scan	0.235	0.810
AESM	0.880	0.892

calculate Sensitivity intuitively. Since it was difficult to analyze with the restricted flexible scan method due to the problem of analysis time, here we compare the detection accuracy of the three types of scanning methods: the echelon scan method, the restricted circular scan method and the AESM. Under conditions of $E[\theta_{i\in\boldsymbol{R}_3}] = 1.5$ and $E[\theta_{i\notin\boldsymbol{R}_5}] = 1.0$, $N = 1000$ sets of simulation data were generated according to the procedure described in Sect. 6.1. $o(\mathbf{G})$ was set to $o(\mathbf{G}) = 642427$ based on the mortality rate of common diseases such as cancer, as in Sect. 6.3. Figure 6.10 and Table 6.2 show the analysis results when the MSWS of each method is half of the total number of regions $m/2 = 1542$, and $\alpha = 0.20$. In the echelon scan method and AESM, Fig. 6.10 shows that both Sensitivity and PPV are 0.8 or higher for many data, and we consider that detection accuracy of them is high. In addition, Fig. 6.10 shows that the echelon scan method tends to lower PPV in some data and the AESM improves it. In contrast, In the restricted circular scan method, since Sensitivity was extremely low, we consider that it could not detect sufficiently even setting a circular-shaped cluster.

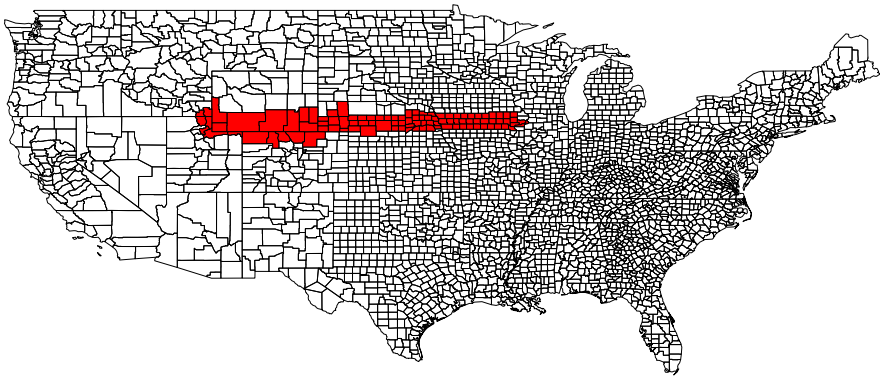
Figure 6.11 shows the ICDA of each method at $\alpha = 0.10, 0.20, 0.30$ when the value of $E[\theta_{i\in\boldsymbol{R}_3}]$ is changed from 1.1 to 2.0 in 0.1 increments. The ICDA of the echelon scan method and the AESM does not show a large difference when $E[\theta_{i\in\boldsymbol{R}_3}] \geq 1.4$, however, the ICDA of the echelon scan method drops sharply when $E[\theta_{i\in\boldsymbol{R}_3}] \leq 1.3$. On the other hand, because the AESM maintain ICDA ≥ 0.50 even at low risk at $\alpha = 0.20$, we consider that AESM has higher detection accuracy than other methods. The restricted circular scan method raises the ICDA by setting α high, however, Fig. 6.11 shows that it is lower than other methods.

6.4.2 Linear-shaped cluster

Assuming the linear-shaped true cluster $\boldsymbol{R}_4$ shown in Fig. 6.9b, we compare the detection accuracy of each method. In addition, since $\boldsymbol{R}_4$ is a linear-shaped cluster, we compare detection accuracy including significant secondary clusters. The settings in the analysis are the same as in the case of $\boldsymbol{R}_3$. Figure 6.12 and Table 6.2 show the analysis results when the MSWS of each method is half of the total number of regions $m/2 = 1542$, and $\alpha = 0.20$.



(a) $\boldsymbol{R}_3$



(b) $\boldsymbol{R}_4$

Figure 6.9: Assumed true clusters $\boldsymbol{R}_3$, $\boldsymbol{R}_4$

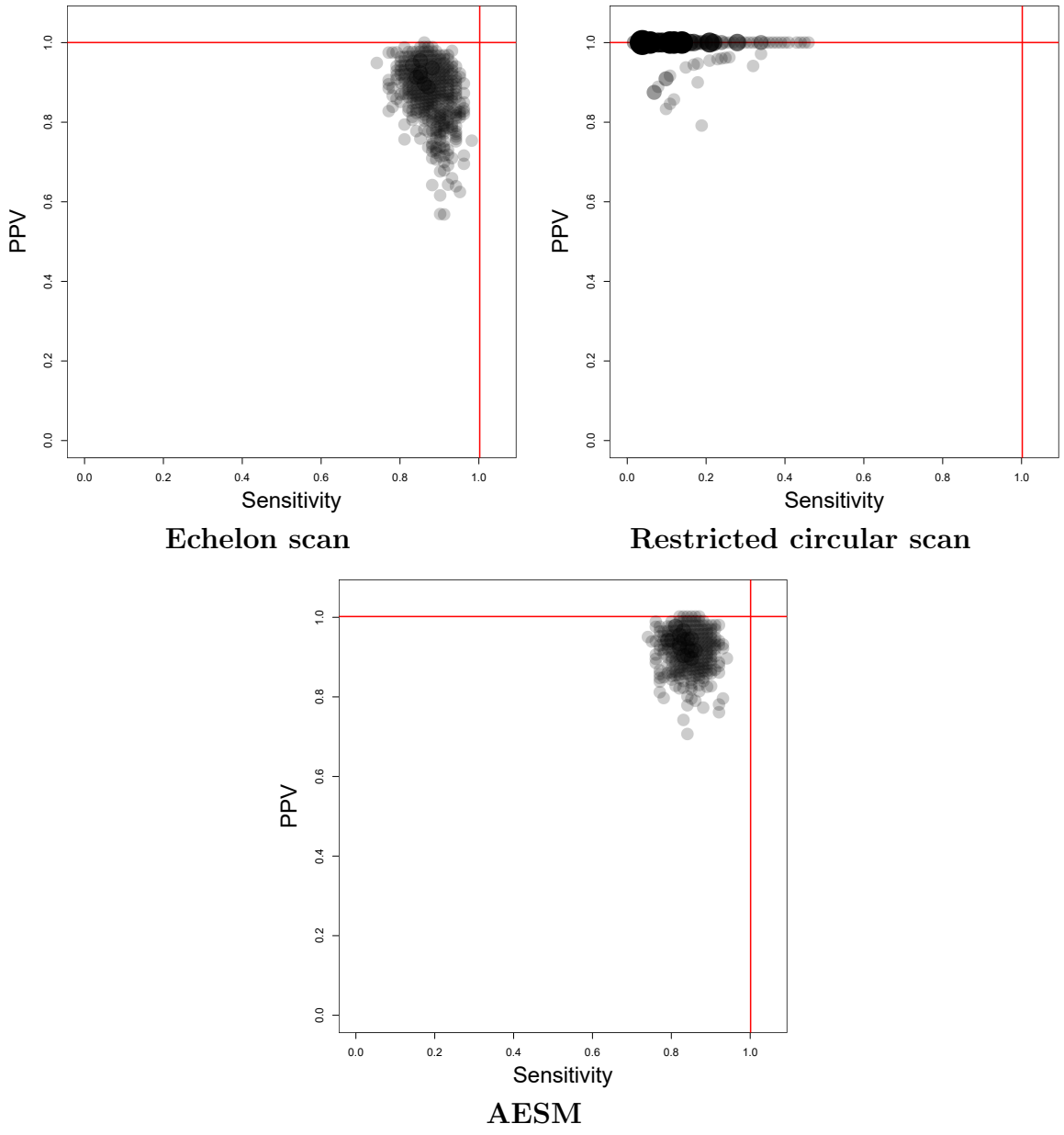
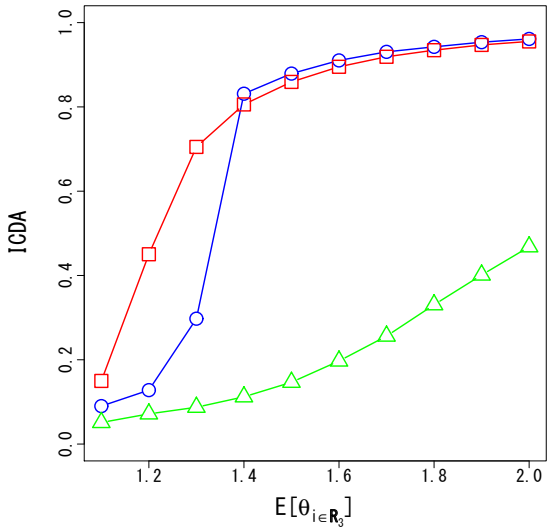
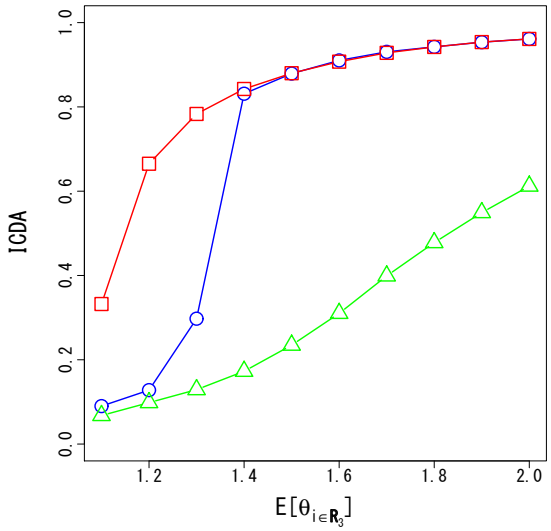


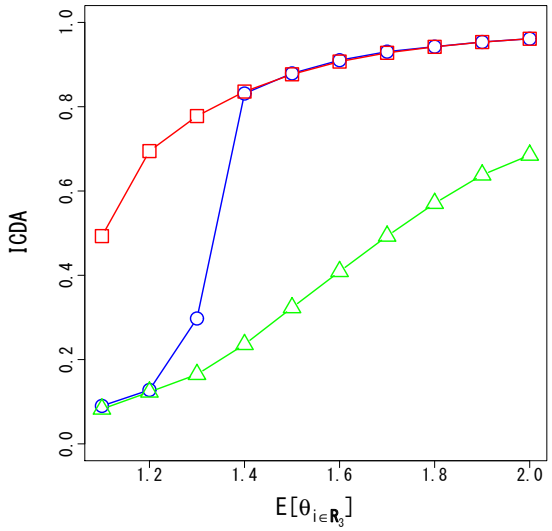
Figure 6.10: Comparison of the detection accuracy for each risk in each method assuming $\boldsymbol{R}_3$



$\alpha = 0.10$



$\alpha = 0.20$



$\alpha = 0.30$

Figure 6.11: Comparison of ICDA for each risk in each method assuming $\boldsymbol{R}_3$. "◯" represents the echelon scan method, "△" represents the restricted circular scan method and "□" represents the AESM.

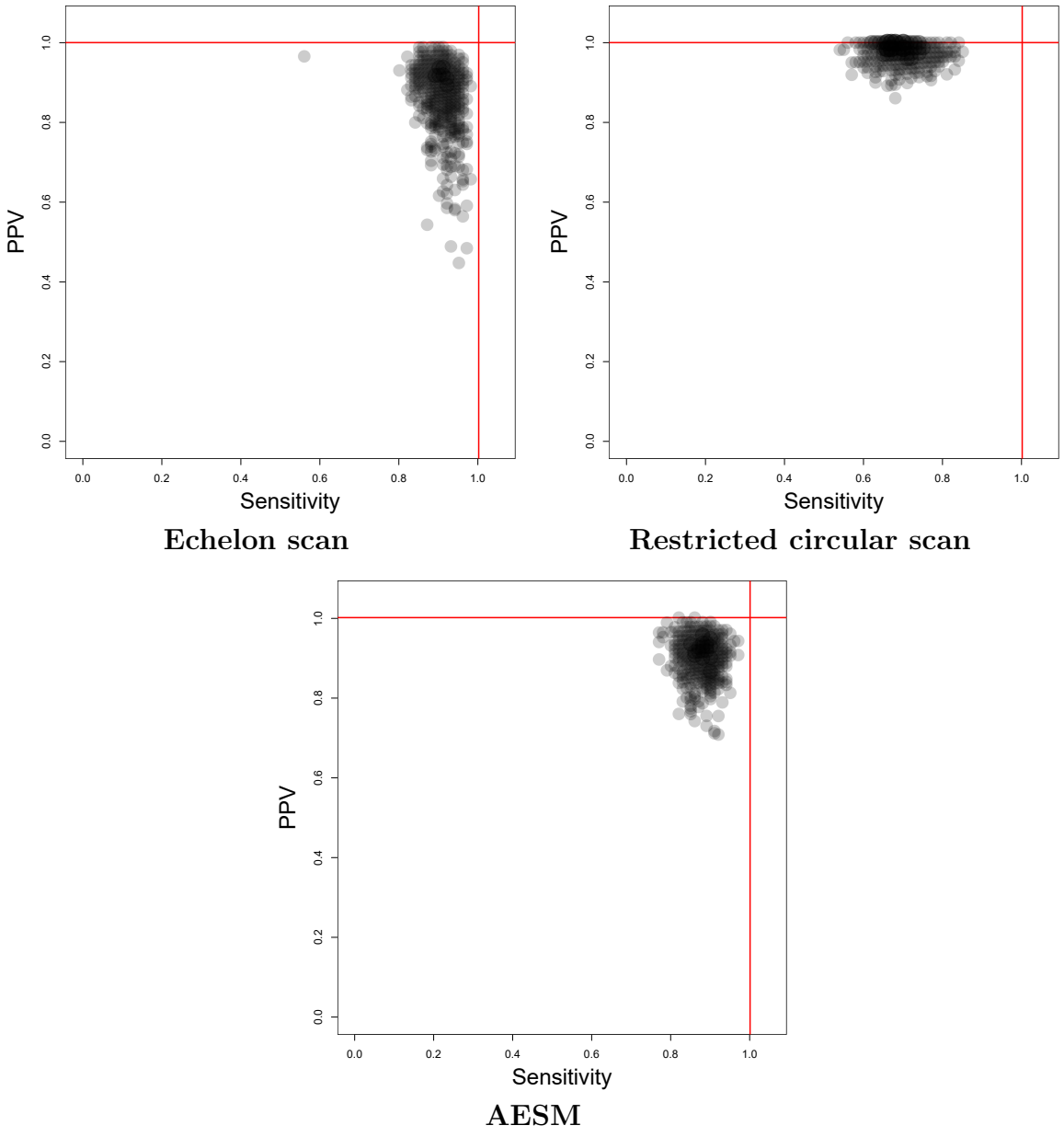


Figure 6.12: Comparison of the detection accuracy for each risk in each method assuming $\boldsymbol{R}_4$

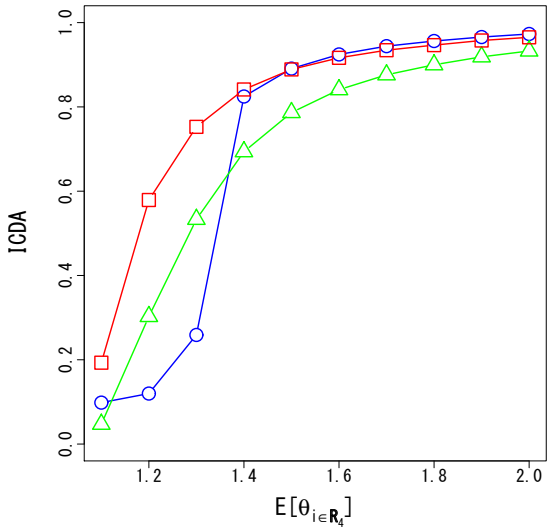
Figure 6.12 shows that the restricted circular scan method has high detection accuracy. In particular, PPV of it tends to be higher than other methods. However, since Sensitivity ≤ 0.8 in all data, we consider that the true cluster was not sufficiently detected. In contrast, since the echelon scan method tends to have high Sensitivity and low PPV, we consider that many regions where were not included in $\boldsymbol{R}_4$ were detected as clusters. In many data, both of Sensitivity and PPV of the AESM was 0.8 or higher. Therefore, we consider that the AESM has higher detection accuracy than other methods.

Figure 6.13 shows the ICDA of each method at $\alpha = 0.10, 0.20, 0.30$ when the value of $E[\theta_{i\in\boldsymbol{R}_4}]$ is changed from 1.1 to 2.0 in 0.1 increments. From Fig. 6.13, the echelon scan method and the AESM have no difference in the ICDA when $E[\theta_{i\in\boldsymbol{R}_4}] \geq 1.4$ for each α. However, the AESM when $E[\theta_{i\in\boldsymbol{R}_4}] \leq 1.3$ has higher values than other methods. Furthermore, the restricted circular scan method also tends to have higher ICDA at low risk than the echelon scan method.

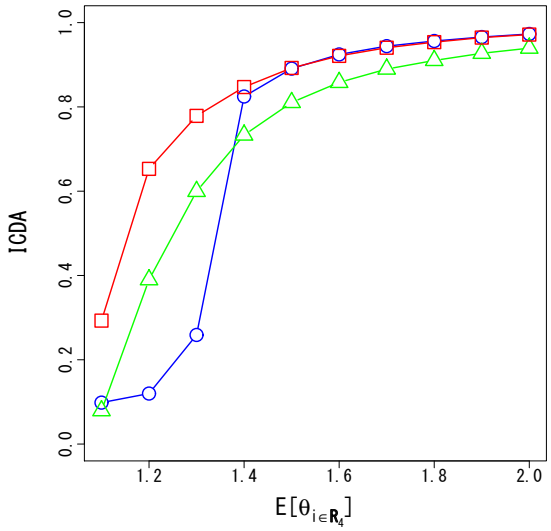
6.5 Discussion

First, we consider the detection accuracy of the echelon scan method and the AESM. First, we consider the detection accuracy of the echelon scan method and the AESM. In the simulation using the grid data performed in Sect. 6.3, the result showed the echelon scan method has a higher ICDA than the AESM. In particular, the difference is large when $\alpha = 0.10$. According to Tango's guide for α, $\alpha = 0.10$ detects "small clusters with a sharp increase in risk". Therefore, in the simulation conducted this time, we consider that small clusters were detected compared to the true cluster to be a factor that lowers the ICDA of the AESM. In all simulation results, the ICDA of the AESM shows the equivalent value to that of the echelon scan method as α is set higher, and we consider that the size of the detection clusters depending on α affects the ICDA. In addition, in the grid data, when $E[\theta_{i\in\boldsymbol{R}_4}] \leq 1.2$, the risk of regions set as the true cluster $\boldsymbol{R}$ cannot be said to be high in the first place. As a result, Sensitivity became extremely low because the AESM detected only a part of regions forming $\boldsymbol{R}$. From this, we considered that the difference from the echelon scan method was large in the low-risk true cluster.

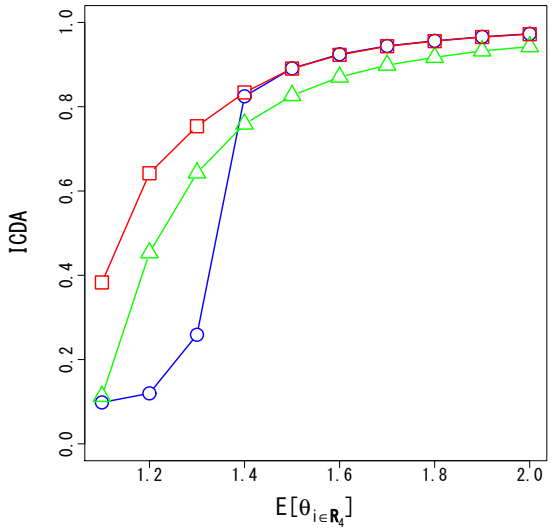
On the other hand, according to the data by county in United States, the ICDA of the echelon scan method decreased sharply at low risk, and the number of regions included in clusters detected was exceeded 1000 regions. Therefore, we considered that the denominator of PPV became larger and ICDA became extremely low compared to the case of grid data. Although the AESM could not detect true clusters sufficiently at low risk, we consider that the ratio of including unexpected regions is smaller than that of the echelon scan method. For this reason, we consider that the AESM has high detection accuracy. From the above, the AESM can detect high-risk clusters because it does not



$\alpha = 0.10$



$\alpha = 0.20$



$\alpha = 0.30$

Figure 6.13: Comparison of ICDA for each risk in each method assuming $\boldsymbol{R}_4$. "◯" represents the echelon scan method, "△" represents the restricted circular scan method and "□" represents the AESM.

include low-risk regions compared to the echelon scan method, although the size of the clusters detected differs depending on α. In particular, we consider that the detection accuracy could be improved for large-scale spatial data.

Next, we consider that the detection accuracy of the restricted circular scan method, the restricted flexible scan method and the AESM. In the simulation of the grid data in Sect. 6.3, ICDA of the restricted flexible scan method and the AESM became higher values than the restricted circular scan method. However, in the simulation of the data by United States county in Sect. 6.4, the restricted flexible scan method could not analyze and obtain results. We consider that this is because the amount of calculation becomes enormous for large-scale spatial data due to the feature of scanning all windows that can be constructed within a certain range. Because the AESM has a detection accuracy equal to or higher than that of the restricted flexible scan method, we consider that it is an effective method in that it can be applied to large-scale spatial data.

The restricted circular scan method had a lower ICDA than the AESM in all simulations. The reason for this is that in the restricted circular scan method, the order to be scanned is predetermined by the distance from the region where scanning is started, and the scanning is stopped when the region i where $p_i < \alpha$ appears in the scanning process. Therefore, we consider that it was difficult to sufficiently detect the true cluster as compared with other methods. In fact, in the circular clusters compared only by MLC, the Sensitivity of the restricted circular scan method was extremely low. From the above, we consider that the AESM is an effective method in that it has higher detection accuracy than the existing method using Tango's statistic and can be applied to large-scale spatial data.

In this paper, we assumed the observation data of death due to general diseases, and set the observed value $o(\mathbf{G})$ in the study area to perform the simulation. Regarding the fluctuation of the detection accuracy in this $o(\mathbf{G})$ setting, Tango and Takahashi (2012) showed that the fluctuation is not large except when assuming a disease that occurs extremely rarely. Therefore, even if different observation values $o(\mathbf{G})$ are set, we consider that the fluctuation of the analysis results shown in this paper is small.

7 Detection of space-time clusters

7.1 The cylindrical scan method

When conducting an analysis of infectious diseases, it is important to know where and when the cluster occurred. Such a cluster having information on the position and period of occurrence is called a spatiotemporal cluster. As a method for detecting space-time clusters, Kulldorff et al. (1998) proposed the cylindrical scan method based on the spatial scan statistic. We assume that observation data for each time point exists for each region in the study area consisting of m regions, and let $l(i,t)\,(i=1,2,\ldots,m;t=1,2,\ldots,T)$ be the region i at time point t. At this time, if there is no cluster in the study area at any time, the random variable $O_{i,t}$ with the observed value $o_{i,t}$ in region i at time point t can be stated as follows:

$$O_{i,t} \sim \text{Poisson}(\xi_{i,t}), \quad i=1,2,\ldots,m;t=1,2,\ldots,T$$

where $\xi_{i,t}$ is the expected number of cases in region i at time point t.

In such spatiotemporal data, assume a circular window $\mathbf{Z}_{ik}$ consisting of k regions centered on $l(i,t)$ on the plane of the study area at time point t. The method of scanning this window $\mathbf{Z}_{ik}$ is the same as the circular scan method described in Sect. 3.1. For window $\mathbf{Z}_{ik}$ on a plane, we consider a cylindrical window $\mathbf{W}$ from time point s_1 to time point s_2, where $1 \leq s_1 \leq s_2 \leq T$, and the univarsal set $\mathcal{W}$ of $\mathbf{W}$ is defined by Eq. (3.1) as follows:

$$\mathcal{W} = \mathcal{Z}_1 \times \mathcal{T}, \tag{7.1}$$

where $\mathcal{T}$ is defined as follows:

$$\mathcal{T} = \{[s_1,\, s_2] \mid 1 \leq s_1 \leq s_2 \leq T\}. \tag{7.2}$$

At this time, the maximum value of the width $s_1 - s_2 + 1$ of the interval $[s_1,\, s_2]$ is called the maximum temporal window size (MTWS). The window $\hat{\mathbf{W}}$ that maximizes the spatial scan statistic shown in Eq. (2.5) is MLC, and the significance of it is evaluated using the Monte Carlo method.

Figure 7.1 shows an image of the cylindrical scan method. In the cylindrical scan method, by scanning while changing the radius and height of the window, it is possible to concurrently detect the location and time interval of the space-time cluster. However, since this method applies a cylinder with a precise circular surface, only clusters with the

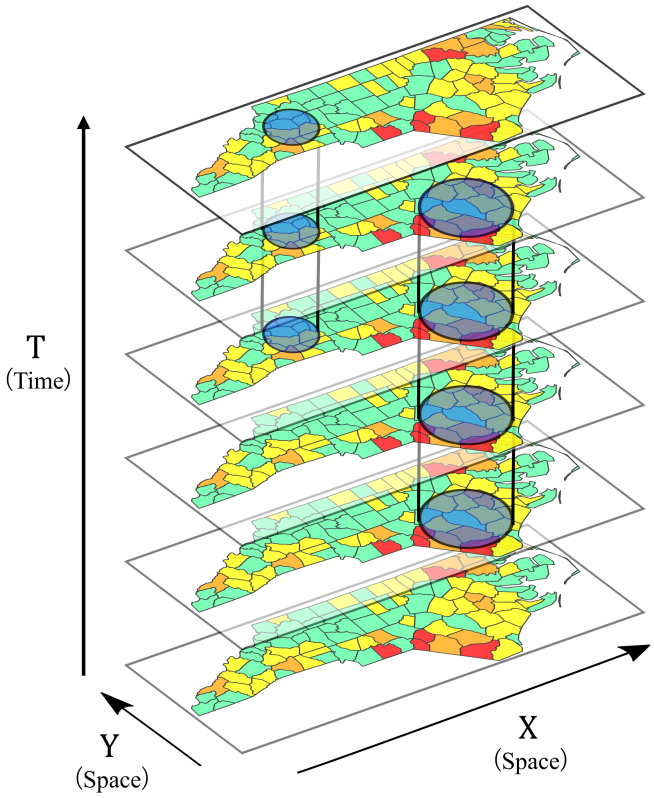


Figure 7.1: Image of the cylindrical scan method

same regional group are detectable. Accordingly, detection becomes difficult when the cluster's shape changes over time (Patil and Taillie 2004). Figure 7.2 shows example of clusters that are difficult to detect with the cylindrical scan method. In Fig. 7.2, the red regions show the location of the regions included in the true space-time cluster. In real data, the true cluster may change over time. For example, when the number of regions included in a true cluster increases and its scale expands or when it is divided into multiple clusters and they move. However, the regions with blue dots detected by the cylindrical scan method does not change over time. Therefore, because it cannot capture changes in these clusters, the true cluster is partially detected or the regions different from the true cluster are mistakenly detected.

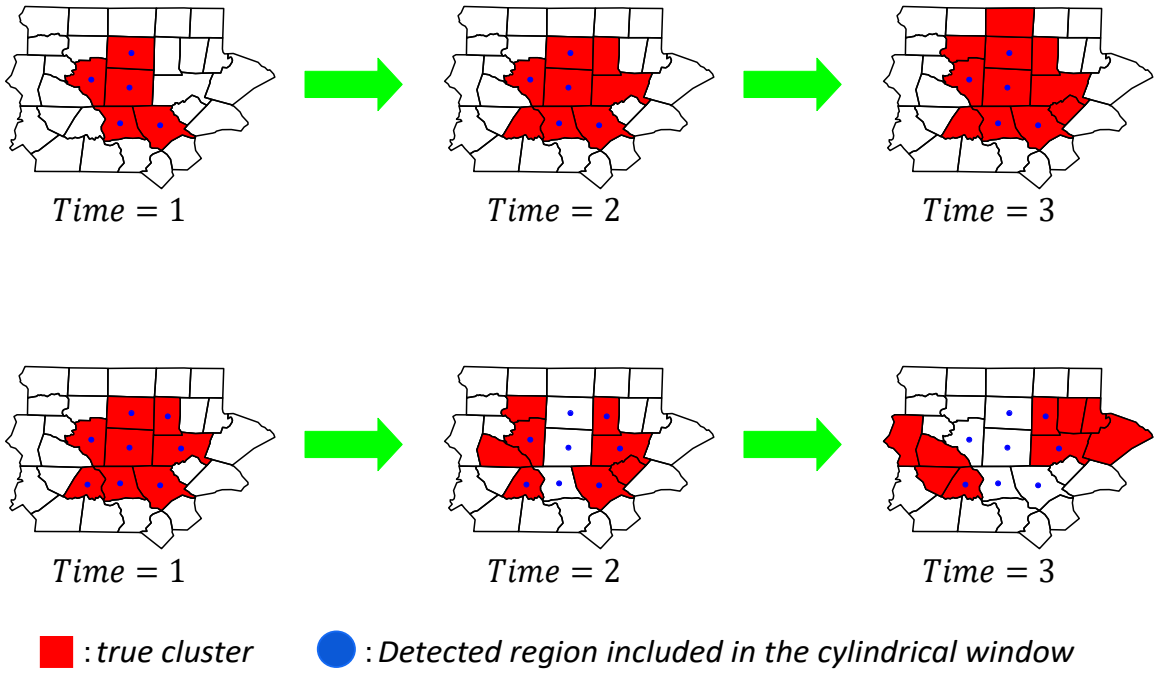


Figure 7.2: Example of expanding cluster (upper) and dividing cluster (lower)

7.2 Space-time cluster detection using the AESM

In the case of infectious diseases, the disease may spread to the surrounding the initial cluster. Therefore, it is important to capture changes in the cluster's shape over time to identify the nature of the infection's spread and the factors involved therein. Echelon analysis can create a dendrogram using the neighboring information of each value (region), even in spatiotemporal data. As an example, 5×5 grid data at three different time points are shown in Fig. 7.3. Here, the attribute value of each region in the grid data is in the area (the attribute value of the region in row A and the first column at $t = 1$ is 11). Figure 7.3d shows the location ID for each region. These data can be considered the spatial data of 75 regions (25 regions $\times$ 3 time points). When each region is denoted by $l(i,t)$ $(i = 1, 2, \ldots, 25; t = 1, 2, 3)$, the simplest example defining neighbors $NB(l(i,t))$ of $l(i,t)$ is given by:

$$NB(l(i,t)) = \begin{cases} \{l(k,t) \mid \text{region } i \text{ and } k \text{ are neighbors}\} \cup l(i,t+1), & t = 1 \\ \{l(k,t) \mid \text{region } i \text{ and } k \text{ are neighbors}\} \cup l(i,t+1) \cup l(i,t-1), & t = 2 \\ \{l(k,t) \mid \text{region } i \text{ and } k \text{ are neighbors}\} \cup l(i,t-1), & t = 3 \end{cases} \tag{7.3}$$

where $l(k,t)$ $(k = 1, 2, \ldots, 25; k \neq i)$ is the region adjacent to $l(i,t)$ at time point t. Figure 7.4 shows the echelon dendrogram for the data when the spatial adjacency at the given time point is defined as four neighborhoods (up, down, left and right). The dendrogram's

vertical axis represents the attribute value of the data, and the symbols in the dendrogram denote the position of each region on the dendrogram (where "C4(3)" refers to the region in row C and the fourth column at $t = 3$). It is possible to detect space-time clusters by scanning based on the structure of this dendrogram. Accordingly, it can capture changes over time of the cluster, such as expansion, contraction and movement.

Echelon analysis makes it possible to represent the spatiotemporal data as a two-dimensional echelon dendrogram. However, when data is collected over a long period, the scale of the data will range from thousands to tens of thousands of values, even if the number of regions within the scanned space is small. Hence, the number of calculations required when the echelon scan method is applied becomes vast, dramatically increasing the analysis time. This paper applied the AESM to the spatiotemporal data to detect clusters. First, $p_{i,t}$, which is given to each region at time point t, is defined as follows:

$$p_{i,t} = \Pr\{O_{i,t} \geq o_{i,t} + 1 \mid O_{i,t} \sim \mathrm{Pois}(\xi_{i,t})\} + \frac{1}{2}\Pr\{O_{i,t} = o_{i,t} \mid O_{i,t} \sim \mathrm{Pois}(\xi_{i,t})\}, \quad (7.4)$$

Figure 7.5 shows the application of the AESM to spatiotemporal data. By extracting the regions that satisfy $p_{i,t} < \alpha$, it is possible to detect clusters comprising only high-risk regions accurately. Additionally, since the region to be scanned is reduced, the calculation cost is inhibited, even for large-scale data. The upper left Fig. 7.5 represents spatiotemporal data, where red-colored regions are high-risk and satisfy $p_{i,t} < \alpha$ and blue-colored regions do not satisfy $p_{i,t} < \alpha$. In Step 1, only the red-colored regions are extracted from the original data. A dendrogram is created from the extracted data by echelon analysis in Step 2. Finally, in Step 3, the cluster is detected by scanning from the upper hierarchy of the dendrogram.

	1	2	3	4	5
A	11	41	22	58	7
B	2	72	59	68	63
C	53	4	9	15	45
D	50	26	33	5	65
E	42	3	24	25	49

(a) $t = 1$

	1	2	3	4	5
A	73	40	61	39	34
B	14	21	13	20	19
C	69	30	51	32	36
D	37	12	18	31	56
E	23	54	27	48	47

(b) $t = 2$

	1	2	3	4	5
A	71	70	46	60	67
B	52	17	66	1	35
C	74	55	57	75	28
D	64	8	38	29	10
E	6	43	16	44	62

(c) $t = 3$

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

(d) Location ID for each region

Figure 7.3: Sample of spatiotemporal data.

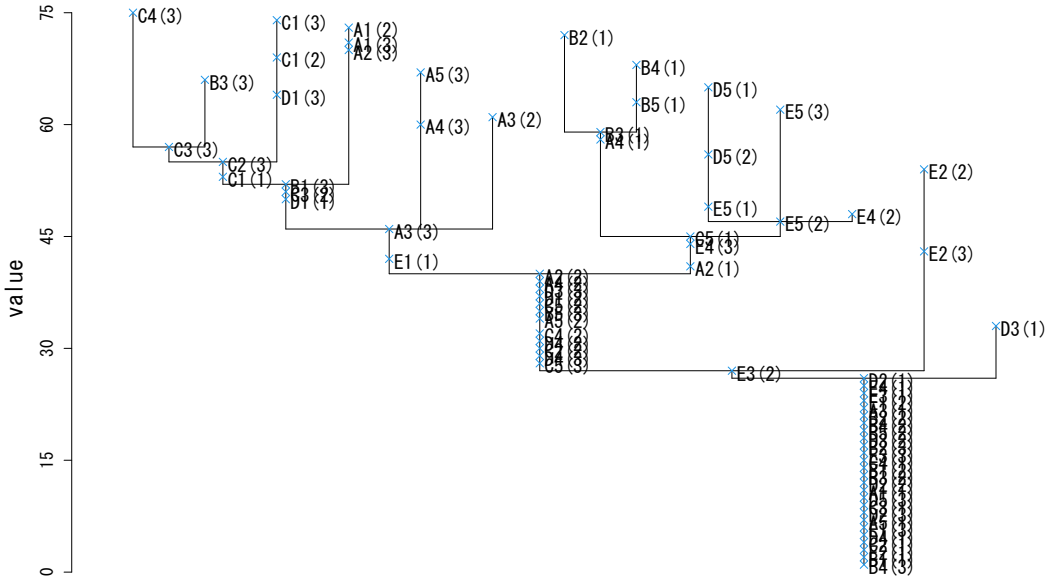


Figure 7.4: Echelon dendrogram for the sample data

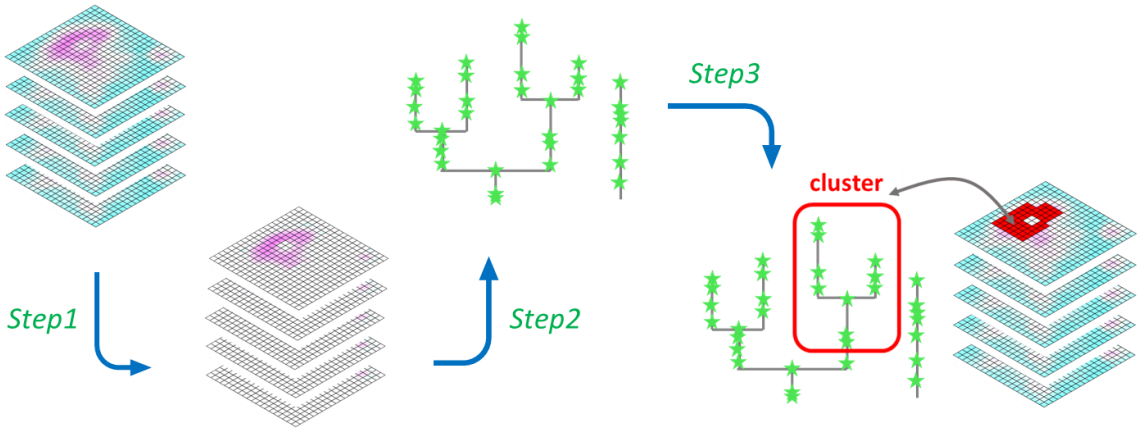


Figure 7.5: Flow of space-time cluster detection using the AESM

8 Real data analysis

8.1 Data on COVID-19-infected people in Japan

Coronavirus disease 2019 (COVID-19) is caused by a novel coronavirus known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The virus has spread worldwide since it was first reported in Wuhan, Hubei Province, China, in December 2019. In Japan, the number of infected people has seen a sustained increase since the first confirmed case of COVID-19 in January 2020. The country's infection status is reported in various media, and information is actively disclosed in each prefecture. As such, interest in COVID-19 is very high. Studies on COVID-19 have been advanced globally in various fields. Furthermore, research on space-time cluster detection is being conducted. However, research on space-time cluster detection has not been conducted much in Japan. Therefore, we consider that it is epidemiologically and sociologically important to capture the temporal changes of clusters that occur in Japan.

We obtained the dataset created by ESRI Japan Co., Ltd (2021) based on the status of test-positive individuals in each prefecture (domestic cases, excluding airport quarantine and charter flight cases) announced by the Ministry of Health, Labor, and Welfare. This dataset is available on a dedicated ESRI Japan Co., Ltd. website (`https://coronavirus-esrijapan-ej.hub.arcgis.com/`). We used the number of people newly infected per day, aggregated for 326 days from March 11, 2020, to January 30, 2021. However, since these numbers were calculated based on the difference from the cumulative number of infected people reported on a preceding day, the number of newly infected people may have a negative value if there was a data correction at the time. There were 22 such cases; we replaced these numbers with 0. Figure 8.1 features a graph showing the number of newly infected people in Japan and the moving average for this number over the preceding seven days during the study period. As of January 30, 2021, the total number of infected people was 384,014, and the number of infected people per day had the highest value, at 7,863 on January 08, 2021.

8.2 Space-time clusters based on population

We applied both the cylindrical scan method and the AESM to the data regarding COVID-19-infected people in Japan described in Sect. 8.1 to detect space-time clusters based on population. We collected the data of residents in each prefecture as the population data. We first used the SaTScan^{TM} software to apply the cylindrical scan method.

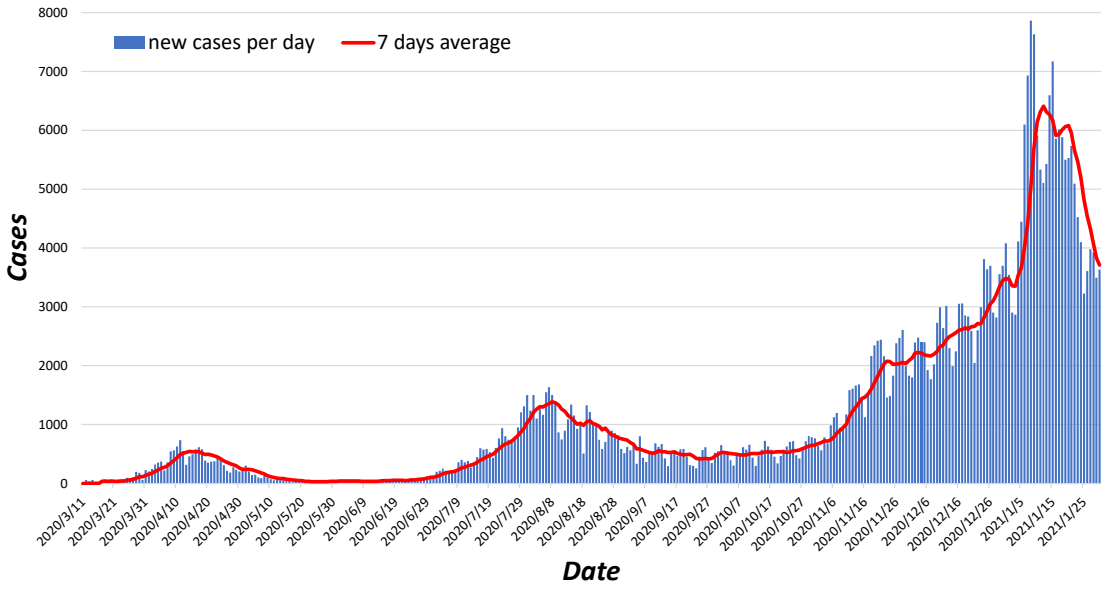


Figure 8.1: Number of daily COVID-19 cases from March 11, 2020, to January 30, 2021.

The setting of each method is described as follows. In the cylindrical scan method, we restricted MSWS to include 20% or less of the population. This setting is necessary because about 10% of the population in Japan is concentrated in Tokyo, so if it were set to 10% or less, Tokyo might not be detected. The second reason is that the population of each district is about 10% to 20% of the total population, which made it easy to interpret the results. In addition, we restricted MTWS for the cylindrical scan method to 180 days or fewer. In this study, we aimed to capture the shape change of clusters by detecting long-term clusters with the AESM. Therefore, to allow comparison with the AESM results, we felt it was necessary to detect long-term clusters with the cylindrical scan method. This guided our selection of the settings described above.

For the AESM, we restricted MSWS to include 20% or less of the population and set the criterion α at 0.01. Tango's index was shown based on a simulation of data consisting of about 100 regions regarding the setting of α. However, in the case of large-scale data such as spatiotemporal data, the number of regions included in the detected clusters may be larger than expected even if α is set at 0.05. This is because, unlike existing methods, the AESM has no restrictions on the cluster's shape that can be detected. Therefore, in analyzing this study, we determined that it was necessary to set the value of α to be more restrictive than the values of Tango's index and set α to 0.01.

We used the standardized morbidity ratio (SMR) as the attribute value for each prefecture ($i = 1, 2, \ldots, 47$) at a time $t\,(= 1, 2, \ldots, 326)$ for the echelon analysis. Let $o_{i,t}$ and $\xi_{i,t}$ be the number of cases and the expected number of cases in each prefecture at time

t, respectively. We calculated SMR using the following formula:

$$\theta_{i,t} = \frac{o_{i,t}}{\xi_{i,t}}. \tag{8.1}$$

As the simplest expected number of cases, without considering covariates, such as age and gender, we defined $\xi_{i,t}$ as follows:

$$\xi_{i,t} = w_{i,t} \times \frac{\sum_{i=1}^{47} o_{i,t}}{\sum_{i=1}^{47} w_{i,t}}, \tag{8.2}$$

where $w_{i,t}$ is the population of region i at time t. We used the estimated population published monthly by each prefecture for the population in the study area. Furthermore, as the neighboring information for each area, we used the data regarding adjacent prefectures that determines the eligible area for coupons distributed by the regional tourism support project implemented by the Japanese government. We obtained this data from the following URL (`https://goto.jata-net.or.jp/coupon/area.html`). Besides the geographical adjacencies, this information includes adjacencies between prefectures that can be traveled by a sea route as a day trip. This data was included because that Okinawa does not have geographically adjacent prefectures. Figure 8.2 shows the geographical location of each prefecture in Japan, and Table 8.1 provides the numbers of the areas adjacent to each prefecture. We considered that $\theta_{i,t}$ of region i at time t was affected by $\theta_{i,t-1}$ of the previous day, and $\theta_{i,t+1}$ of the next day was affected by $\theta_{i,t}$; We considered region i at time t adjacent to the same region on the previous and subsequent days as temporal adjacency information. Thus, when each prefecture is denoted by $l(i,t)\,(i = 1, 2, \ldots, 47; t = 1, 2, \ldots, 326)$, the neighboring information, $NB(l(i,t))$, is defined as follows:

$$NB(l(i,t)) = \begin{cases} \{l(k,t) \mid \text{region } i \text{ and } k \text{ are neighbor}\} \cup l(i,t+1), & t = 1 \\ \{l(k,t) \mid \text{region } i \text{ and } k \text{ are neighbor}\} \cup l(i,t+1) \cup l(i,t-1), & 1 < t < 326 \\ \{l(k,t) \mid \text{region } i \text{ and } k \text{ are neighbor}\} \cup l(i,t-1), & t = 326 \end{cases} \tag{8.3}$$

where $l(k,t)\,(k = 1, 2, \ldots, 47; k \neq i)$ is the prefecture adjacent to $l(i,t)$ at time point t.

Table 8.1: Neighboring information of each prefecture

No.	Location	Neighbors
1	Hokkaido	2
2	Aomori	1, 3, 5
3	Iwate	2, 4, 5
4	Miyagi	3, 5, 6, 7

(continued)

No.	Location	Neighbors
5	Akita	2, 3, 4, 6
6	Yamagata	4, 5, 7, 15
7	Fukushima	4, 6, 8, 9, 10, 15
8	Ibaraki	7, 9, 11, 12
9	Tochigi	7, 8, 10, 11
10	Gunma	7, 9, 11, 15, 20
11	Saitama	8, 9, 10, 12, 13, 19, 20
12	Chiba	8, 11, 13, 14
13	Tokyo	11, 12, 14, 19, 22
14	Kanagawa	12, 13, 19, 22
15	Niigata	6, 7, 10, 16, 20
16	Toyama	15, 17, 20, 21
17	Ishikawa	16, 18, 21
18	Fukui	17, 21, 25, 26
19	Yamanashi	11, 13, 14, 20, 22
20	Nagano	10, 11, 15, 16, 19, 21, 22, 23
21	Gifu	16, 17, 18, 20, 23, 24, 25
22	Shizuoka	13, 14, 19, 20, 23
23	Aichi	20, 21, 22, 24
24	Mie	21, 23, 25, 26, 29, 30
25	Shiga	18, 21, 24, 26
26	Kyoto	18, 24, 25, 27, 28, 29
27	Osaka	26, 28, 29, 30
28	Hyogo	26, 27, 31, 33, 36, 37
29	Nara	24, 26, 27, 30
30	Wakayama	24, 27, 29, 36
31	Tottori	28, 32, 33, 34
32	Shimane	31, 34, 35
33	Okayama	28, 31, 34, 37
34	Hiroshima	31, 32, 33, 35, 38
35	Yamaguchi	32, 34, 38, 40, 44
36	Tokushima	28, 30, 37, 38, 39
37	Kagawa	28, 33, 36, 38
38	Ehime	34, 35, 36, 37, 39, 44
39	Kochi	36, 38
40	Fukuoka	35, 41, 42, 43, 44

(continued)

No.	Location	Neighbors
41	Saga	40, 42
42	Nagasaki	40, 41, 43
43	Kumamoto	40, 42, 44, 45, 46
44	Oita	35, 38, 40, 43, 45
45	Miyazaki	43, 44, 46
46	Kagoshima	43, 45, 47
47	Okinawa	46

The analytica results using the cylindrical scan method with the above settings are shown in Table 8.2 and Fig. 8.3a, and the results from the AESM are shown in Table 8.3 and Fig. 8.3b. Figure 8.3 shows the five clusters with the highest $\log \lambda_K(\mathbf{Z})$ values among the clusters; these were judged to be significant at $p = 0.001$, based on the results of 999 Monte Carlo simulations for each method. Each region's SMR height included in the clusters was expressed using a color gradient; darker colors indicate higher values. The seventh column in Tables 8.2 and 8.3 lists the relative risk (RR), which is calculated as follows:

$$RR = \frac{o(\mathbf{Z})/\xi(\mathbf{Z})}{o(\mathbf{Z}^c)/\xi(\mathbf{Z}^c)}. \tag{8.4}$$

Figure 8.4 visualizes each prefecture; the numbered areas in the figure are the prefectures that were included as a cluster, even if only for one day, in either method.

When the cylindrical scan method was applied, Tokyo and Kanagawa were detected as MLC, and Osaka, Hokkaido, Okinawa, and Fukuoka were detected as secondary clusters. Table 8.2 and Fig. 8.3a show that clusters (excluding Cluster 5) were detected for an extended period, and the MLC was a cluster that lasted approximately 5 months. In

Table 8.2: Details of the clusters detected using the cylindrical scan method

	Location	Time frame	$\log \lambda_K(\mathbf{Z})$	$o(\mathbf{Z})$	$\xi(\mathbf{Z})$	RR
MLC	Tokyo Kanagawa	8/4−1/30	27742.61	123208	63641.36	2.38
Cluster 2	Osaka	7/15−12/19	5296.81	24722	12064.27	2.12
Cluster 3	Hokkaido	10/23−12/10	2748.90	8153	3171.65	2.60
Cluster 4	Okinawa	7/29−11/8	1951.70	3288	873.41	3.79
Cluster 5	Fukuoka	1/18	775.30	1071	238.99	4.49

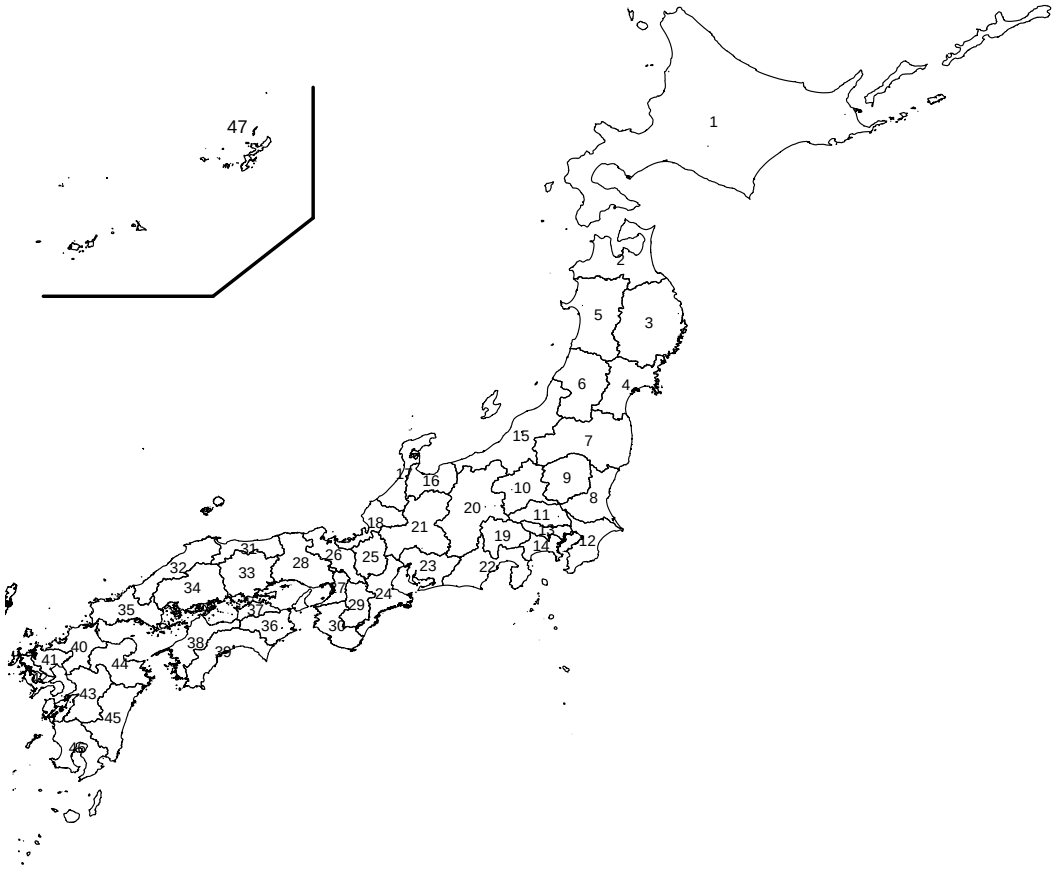
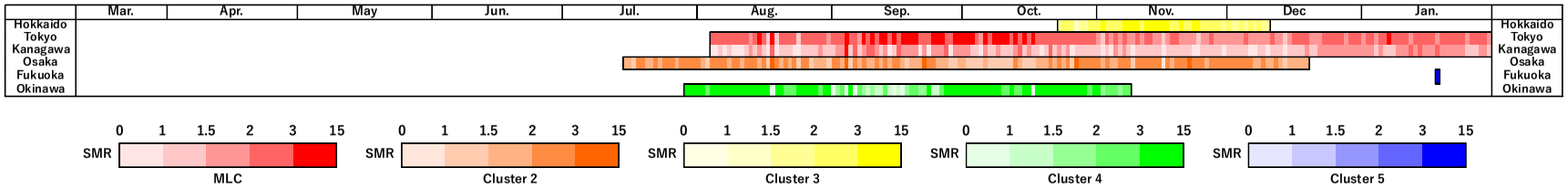
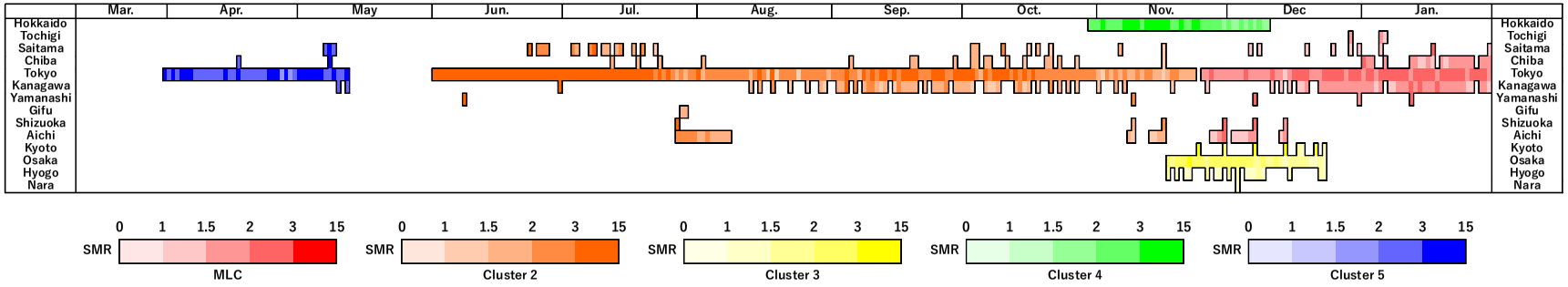


Figure 8.2: Geographical location of each prefecture in Japan — Okinawa (No.47), shown in the upper left of the figure, is actually located in the southwestern part of Japan.



(a) Cylindrical scan method

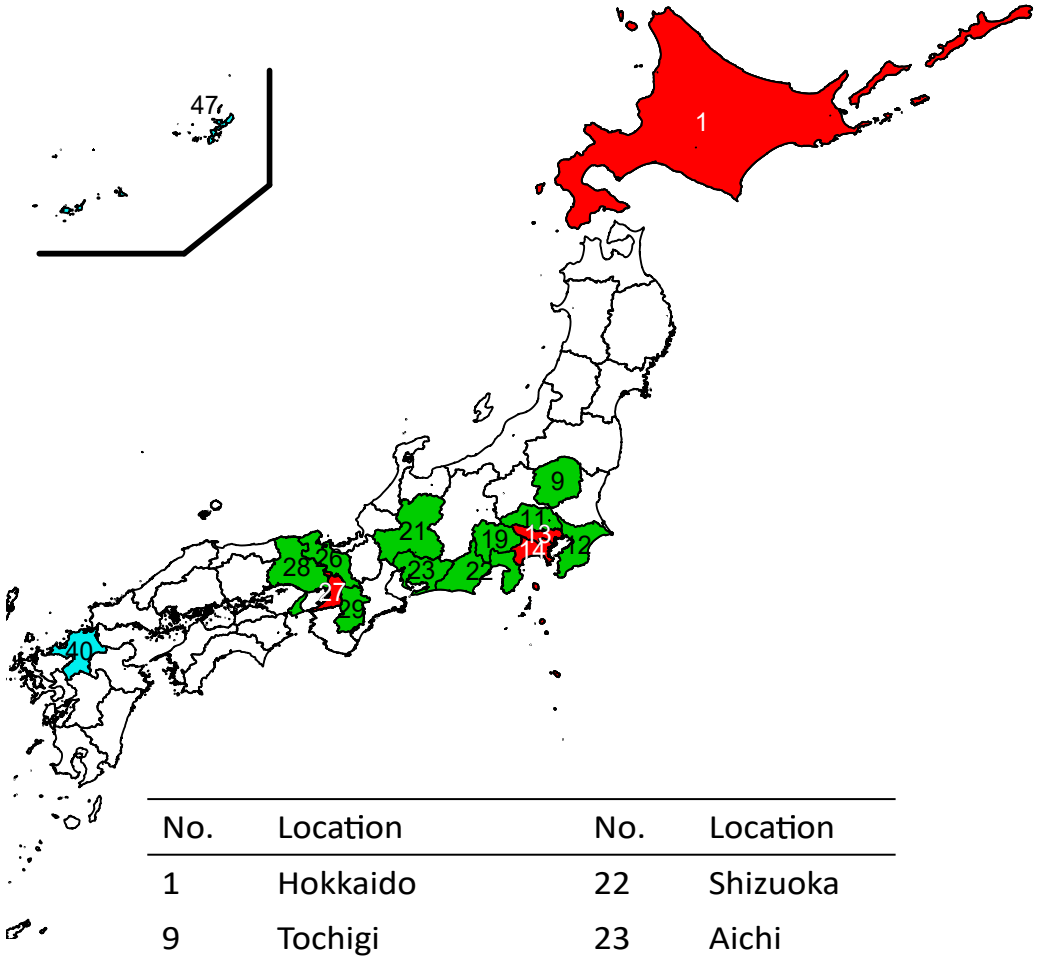


(b) AESM

Figure 8.3: Population-based space-time clusters detected using the cylindrical scan method and the AESM — Although the cylindrical scan method does not use SMR for analysis, in this paper, we performed visualization using SMR to confirm whether prefectures included in clusters have a high-risk. The colored parts in the figure show the prefectures and periods included in the cluster detected by each method, and the high and low of SMR are shown by shades of color.

Table 8.3: Details of the clusters detected using the AESM

	Location	Time frame	$\log \lambda_K(\mathbf{Z})$	$o(\mathbf{Z})$	$\xi(\mathbf{Z})$	RR
MLC	Tochigi Saitama Chiba Tokyo Kanagawa Yamanashi Shizuoka Aichi	11/25−1/30	22387.54	100244	50951.99	2.31
Cluster 2	Saitama Chiba Tokyo Kanagawa Yamanashi Gifu Shizuoka Aichi	6/1−11/23	12257.88	41911	18037.33	2.49
Cluster 3	Kyoto Osaka Hyogo Nara	11/17−12/23	2911.54	15722	8105.49	1.99
Cluster 4	Hokkaido	10/30−12/10	2724.22	7819	2986.19	2.65
Cluster 5	Saitama Chiba Tokyo Kanagawa	3/31−5/12	2033.73	4766	1607.84	2.99



No.	Location	No.	Location
1	Hokkaido	22	Shizuoka
9	Tochigi	23	Aichi
11	Saitama	26	Kyoto
12	Chiba	27	Osaka
13	Tokyo	28	Hyogo
14	Kanagawa	29	Nara
19	Yamanashi	40	Fukuoka
21	Gifu	47	Okinawa

Figure 8.4: Geographical location of prefectures detected as population-based clusters — Red-colored prefectures were detected by both the cylindrical scan method and the AESM. Light blue and green-colored prefectures were detected only by the cylindrical method and only by the AESM, respectively.

Cluster 4, which was detected for Okinawa, $RR = 3.79$ (see Table 8.2), indicating that it was a high-risk cluster, but, as seen in Fig. 8.3a, there was also a day when $\theta_{i,t} < 1$ within the cluster period.

Next, when considering the results of the AESM, besides Tokyo and Kanagawa, prefectures around Tokyo, such as Chiba and Saitama, were also detected as MLC and Cluster 2. From Fig. 5.7b, the Tokyo vicinity was repeatedly included in the clusters for short durations, and the expansion and contraction of the clusters could be observed. Furthermore, an additional cluster was detected in the early part of the target period, from March 31 to May 12, which had not been detected by the cylindrical scan method.

8.3 Space-time clusters based on number of PCR tests

The spread of infectious diseases such as COVID-19 may be centered within areas where people are actively moving. Attempts to slow the spread of infection include conducting sufficient tests on individuals suspected of being infected, such as the close contacts of those who have already been identified as infected. However, in some circumstances, potentially infected individuals could not be sufficiently tested in regions where the number of observed cases was large compared to the number of tests that could be performed. Thus, we assumed that these potentially undetected infected people would impact the development and expansion of the clusters. To detect space-time clusters resulting from such risks, we also conducted an analysis using the number of polymerase chain reaction (PCR) tests performed per day in each prefecture rather than using the population in each prefecture. According to the Johns Hopkins Coronavirus Resource Center, a PCR test is a viral test that aims to identify the presence of a virus's genetic material, as well as evidence of an active viral infection, using an oral or nasal swab or a saliva test. We obtained data on the number of PCR tests performed in each prefecture from the website noted in Sect. 8.1. The number of PCR tests performed per day was calculated using the difference between the cumulative number of tests performed up to the current and the preceding day. However, there were days when some prefectures did not report the cumulative number of tests. In such cases, the number of tests per day was set to 0. In this paper, the number of tests per day was calculated by dividing the increased number if the cumulative number of tests was updated by the required update period. For example, if a prefecture showed zero new tests for 11 days, and there was an increase in the cumulative number of tests of 3,564 on day 12, then by calculating $3{,}564/12 = 297$, the number of new tests on each day during this period was set to 297. This process yielded 281 cases in which the number of newly infected persons per day

was larger than the number of new tests. Accordingly, we processed these data as missing values. The AESM can be applied to the data even with the missing values because the regions with missing values are not used when creating the Echelon dendrogram. The analysis settings were the same as in Sect. 8.2, and $\xi_{i,t}$ was calculated with $w_{i,t}$ as the number of PCR tests performed in region i at time t.

The results of the AESM are shown in Table 8.4 and Fig. 8.5. The numbered prefectures shown in Fig. 8.6 are the newly detected locations as clusters in this analysis. In Cluster 2, 17 prefectures were detected as clusters, demonstrating that infections were widespread during this period. Additionally, Fig. 8.5 shows that Ibaraki was continuously detected for an extended period in both the MLC and Cluster 2, and its SMR was higher than that of other prefectures during this period. Clusters 3 and 5 were detected as clusters at the start of the target period, and an expansion centered on Tokyo was observed. Furthermore, Cluster 3 was detected as a high-risk cluster with $RR = 4.34$.

Table 8.4: Details of the clusters detected using the AESM

	Location	Time frame	$\log \lambda_K(\mathbf{Z})$	$o(\mathbf{Z})$	$\xi(\mathbf{Z})$	RR
MLC	Ibaraki Tochigi Gunma Saitama Chiba Tokyo Kanagawa Yamanashi Shizuoka	12/22−1/29	13308.03	83852	47578.69	1.97

(continued)

	Location	Time frame	$\log \lambda_K(\mathbf{Z})$	$o(\mathbf{Z})$	$\xi(\mathbf{Z})$	RR
Cluster 2	Ibaraki Tochigi Gunma Saitama Chiba Tokyo Kanagawa Yamanashi Shizuoka Aichi Mie Kyoto Osaka Hyogo Nara Wakayama Tokushima	11/10−12/21	3581.24	28432	16710.80	1.75
Cluster 3	Saitama Chiba Tokyo Kanagawa	4/15−5/7	2069.27	2970	687.80	4.34
Cluster 4	Shizuoka Aichi	12/2−12/10	1145.93	1902	498.52	3.83
Cluster 5	Ibaraki Tochigi Saitama Chiba Tokyo Kanagawa Shizuoka	3/23−4/11	940.47	2213	746.89	2.97

8.4 Discussion

We began by considering the results of detecting space-time clusters based on population. Human movement is one of factors that impact the spread of COVID-19 infections.

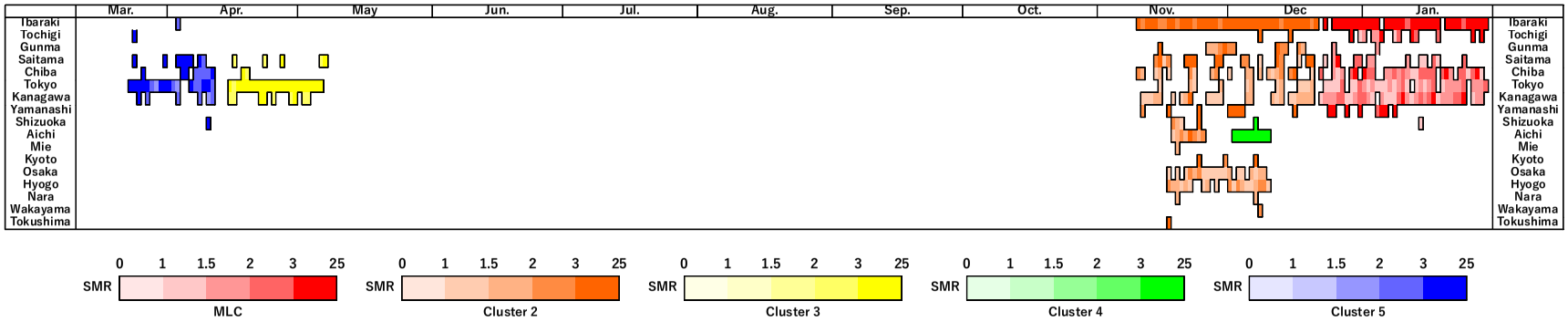
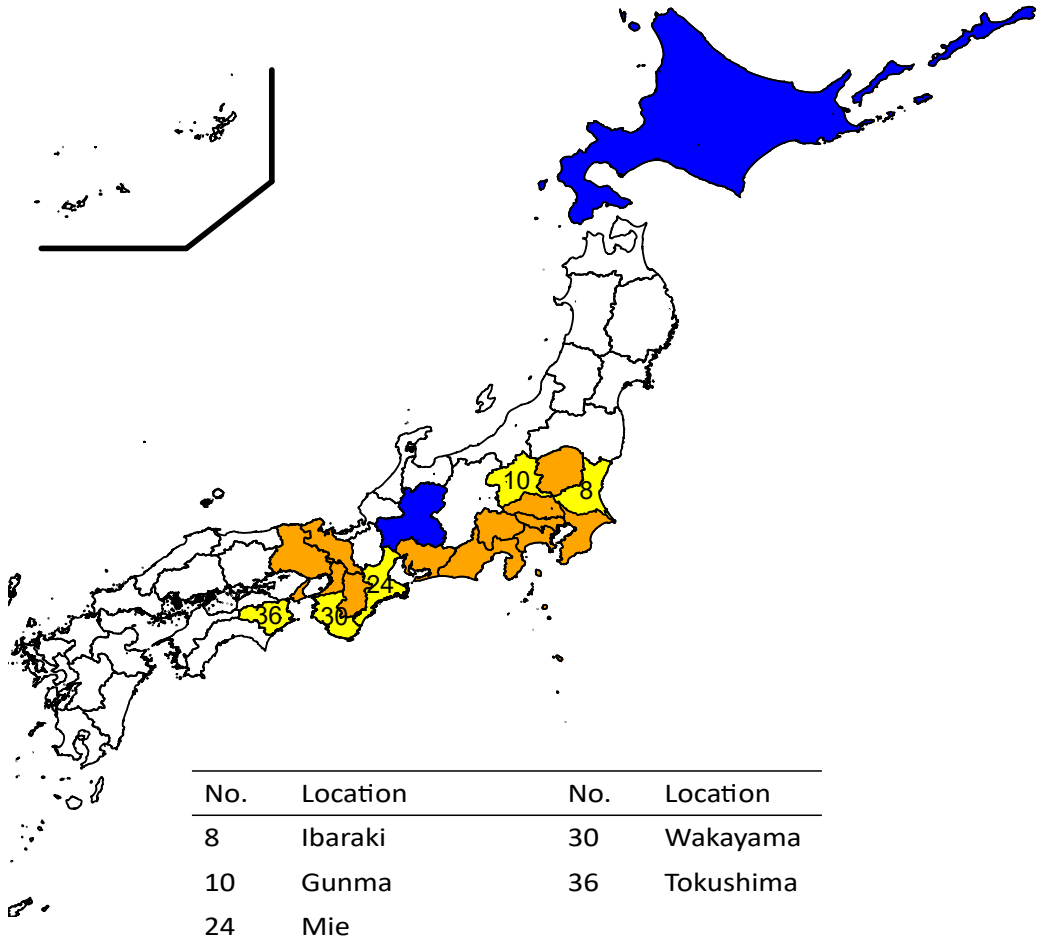


Figure 8.5: Space-time clusters based on the number of PCR tests



No.	Location	No.	Location
8	Ibaraki	30	Wakayama
10	Gunma	36	Tokushima
24	Mie		

Figure 8.6: Geographical location of prefectures detected by the AESM as clusters based on the number of PCR tests — Orange-colored prefectures were detected in both analyses based on population and the number of PCR tests. Blue and yellow-colored prefectures were detected only in analysis based on population and only in analysis based on the number of PCR tests, respectively.

We considered how this aspect how this influenced the generation of clusters. Tokyo, Kanagawa, Osaka, and Fukuoka, which were detected as clusters by the cylindrical scan method, have large populations and are prefectures where many people move for business purposes. Hokkaido and Okinawa are prefectures that many people visit for tourism. Specifically, we considered that the number of tourists had increased compared in late July when Okinawa began to be detected as a cluster; the summer holiday had begun in Japan, and the government's tourism support measures had been implemented. In contrast, the AESM did not detect Okinawa as one of the top five clusters. Considering Cluster 4, as shown in Fig. 8.3a, detected in Okinawa, the SMR exhibited a low value on some days during the detected periods, presumably because multiple clusters that occurred in a short time had been detected as a single cluster. Thus, when the AESM was applied the short-term clusters had a lower $\log \lambda_K(\mathbf{Z})$ than the long-term cluster, and, consequently, they were undetected as a high-ranking cluster.

The cylindrical scan method and the AESM identified the cluster in the Tokyo metropolitan area as the MLC. The AESM detected similar areas in Clusters 2 and 5. Approximately 10% of the Japanese population lives in Tokyo; thus, many people enter and leave the surrounding areas when commuting to work and school. Based on the spread of infection in Tokyo, the surrounding area was also detected as a cluster. Thus, we assume the cluster expansion and contraction would be reflected in the areas surrounding Tokyo. Figure 8.1 shows the number of infected people rapidly increasing in late December 2020. Figure 8.3b shows that the MLC expanded in these areas for the same period. In Japan, many people return home during the New Year holidays or attend events such as Christmas parties with their friends and family. However, during this period, we assume that most people restricted their travel to distant areas due to the influence of COVID-19. As a result, we considered that the movement of people increased in the area around Tokyo, compared to other areas, and this spread infection. Figure 8.3b shows that Cluster 2 included Tokyo in late June, and Fig. 8.7 shows the number of newly infected people in Japan and Tokyo. The number of infected people was small nationwide; however, the proportion for Tokyo was very high during this period. We assume that Tokyo was detected as Cluster 2 because the risk was relatively high compared to other prefectures.

Hokkaido (late October to early December) and Osaka (mid-November to mid-December) were also detected as clusters by the cylindrical scan method and the AESM. Figure 8.8 and Fig. 8.9 are graphs showing the number of newly infected people in Hokkaido and Osaka from October 1, 2020, to December 31, 2020, and the moving average for this number over the preceding seven days. In fact, Fig. 8.8 and Fig. 8.9 show that the number of infected people is increasing rapidly when these areas were detected as clusters. Initially, Tokyo was not included in the tourism support project conducted by the Japanese government, which started in July, however it was included from October 1. Therefore,

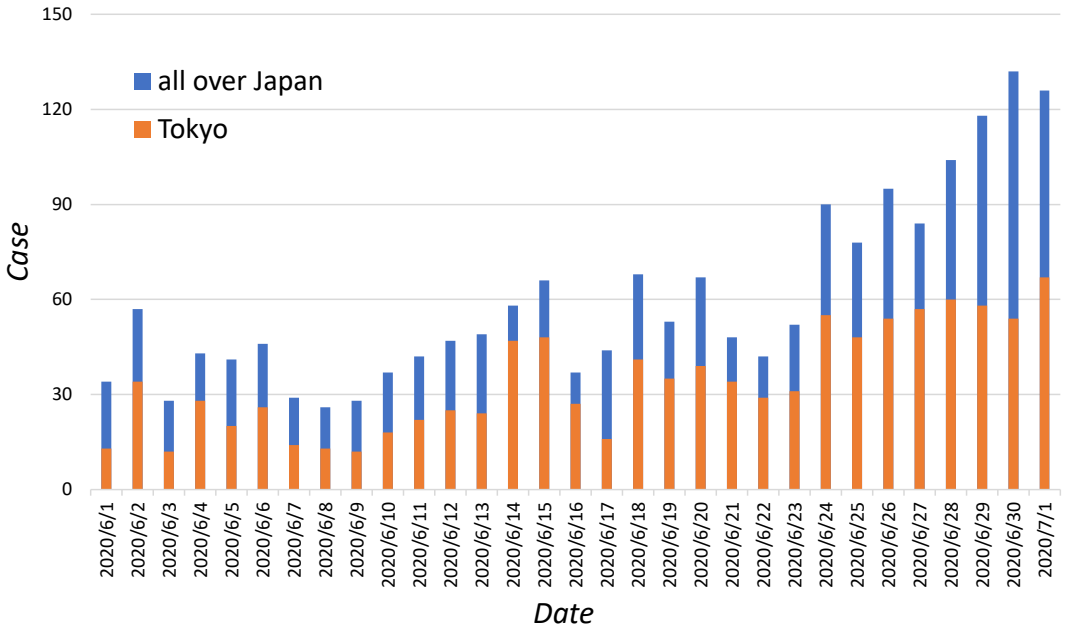


Figure 8.7: Number of daily cases throughout Japan and in Tokyo from June 1 to July 1, 2020.

the number of tourists has increased nationwide since October, and it can be considered that more people visited Hokkaido where is a popular tourist destination. In Osaka, in addition to tourists, the movement of people due to commuting to business or school can be also considered to be a factor in the generation of clusters. Since November 24, the government's tourism support project has been suspended in some areas in Hokkaido and Osaka due to the spread of the infection. In Hokkaido, no clusters have been detected since December 11, about two weeks after November 24. Because the incubation period of COVID-19 is said to be up to 2 weeks, it is suggested that the increase in infected people was suppressed by the decrease in tourists due to the suspension of project.

Next, we considered the space-time clusters based on the number of PCR tests. Figure 8.5 shows that the clusters detected based on these tests lasted for approximately one month. It is also shown that Cluster 2 expanded to an extremely wide area, including the regions surrounding Osaka and Tokyo. Figure 8.1 shows that the number of infected people increased during the period close to November when Cluster 2 was detected. We assume that this occurred because the number of prefectures in which the ratio of infected persons to the number of PCR tests performed was high had increased during this period. Ibaraki in particular, exhibited a high SMR value. Figure 8.10 shows the positive rate of the PCR testing in Ibaraki during the period when the MLC and Cluster 2 were detected, which reflected high values, e.g., $60\% - 70\%$. On April 15, 2021, the Subcommittee on Novel Coronavirus Disease Control, which is an organization of the Japanese government,

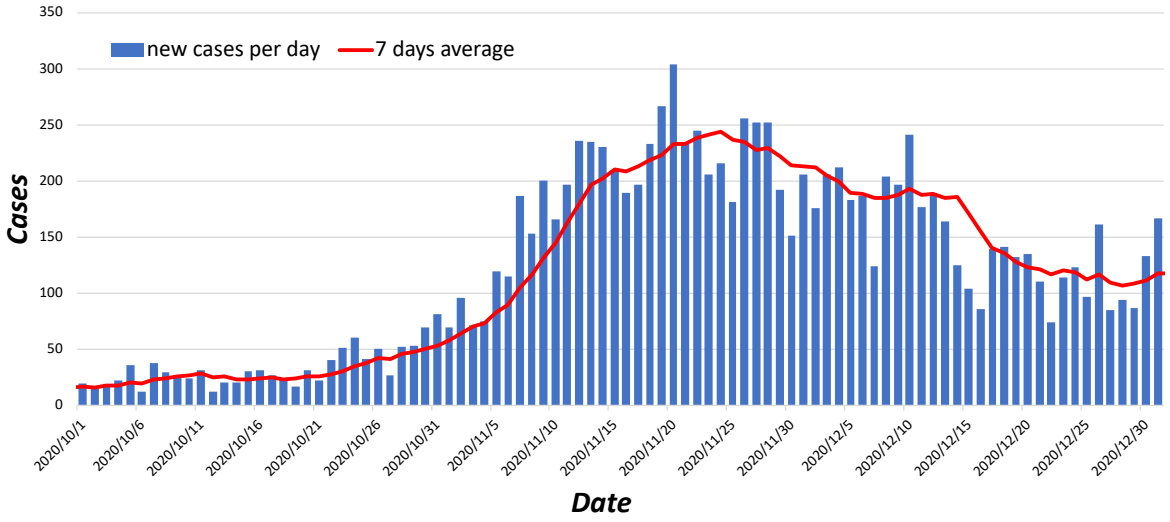


Figure 8.8: Number of daily COVID-19 cases in Hokkaido from October 1 to December 31, 2020.

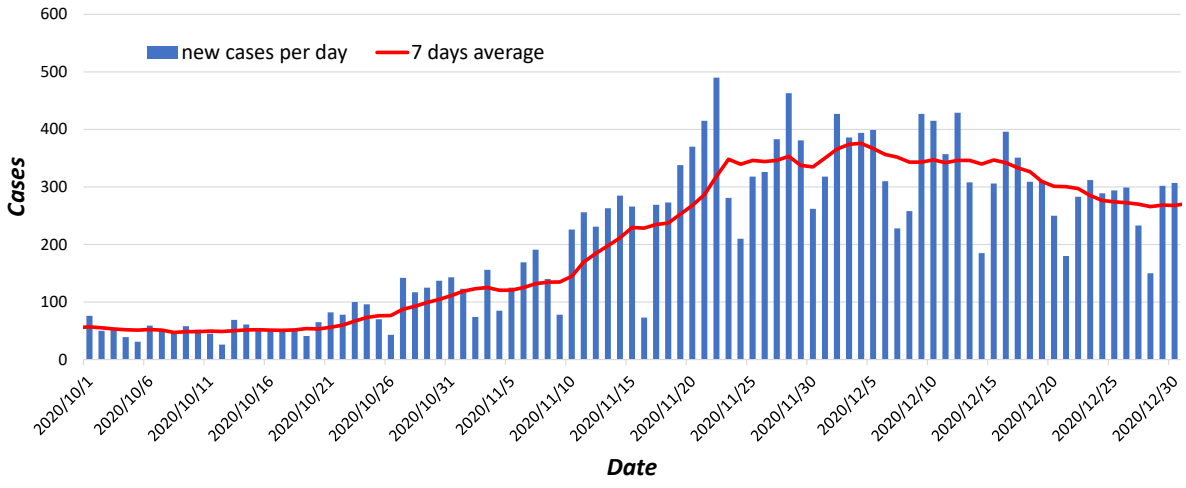


Figure 8.9: Number of daily COVID-19 cases in Osaka from October 1 to December 31, 2020.

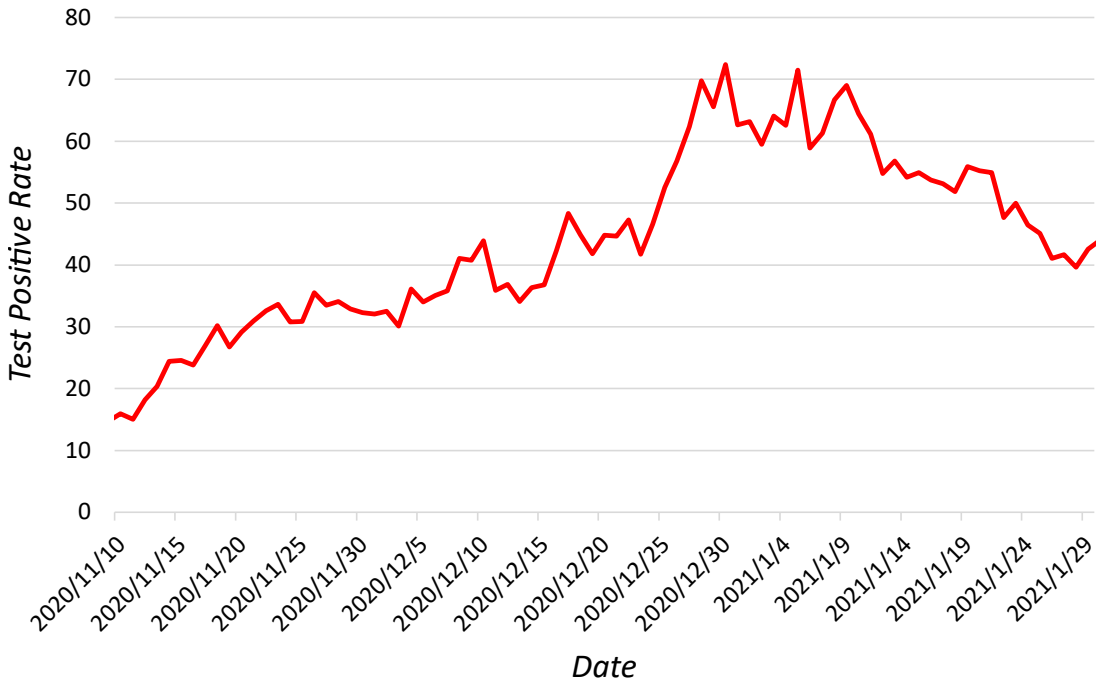


Figure 8.10: Changes in the rate of positive PCR tests in Ibaraki within the period it was included in MLC and Cluster 2.

designated a positive test rate of 5% or more as one of the criteria identifying prefectures where measures are required to avoid a rapid increase in the number of infected people and occurrence of major obstacles to the medical care provision system. Ibaraki shows a sufficiently high value compared to this criterion. Additionally, Fig. 8.11 shows the positive test rate in Tokyo during the period when Cluster 3 and 5 were detected. Tokyo also had a high positive test rate when infections first began to spread in Japan. High positive test rates can make it difficult to provide tests for potentially infected people who have not yet developed symptoms. We considered that these potentially infected individuals eventually contributed to the expansion of the cluster.

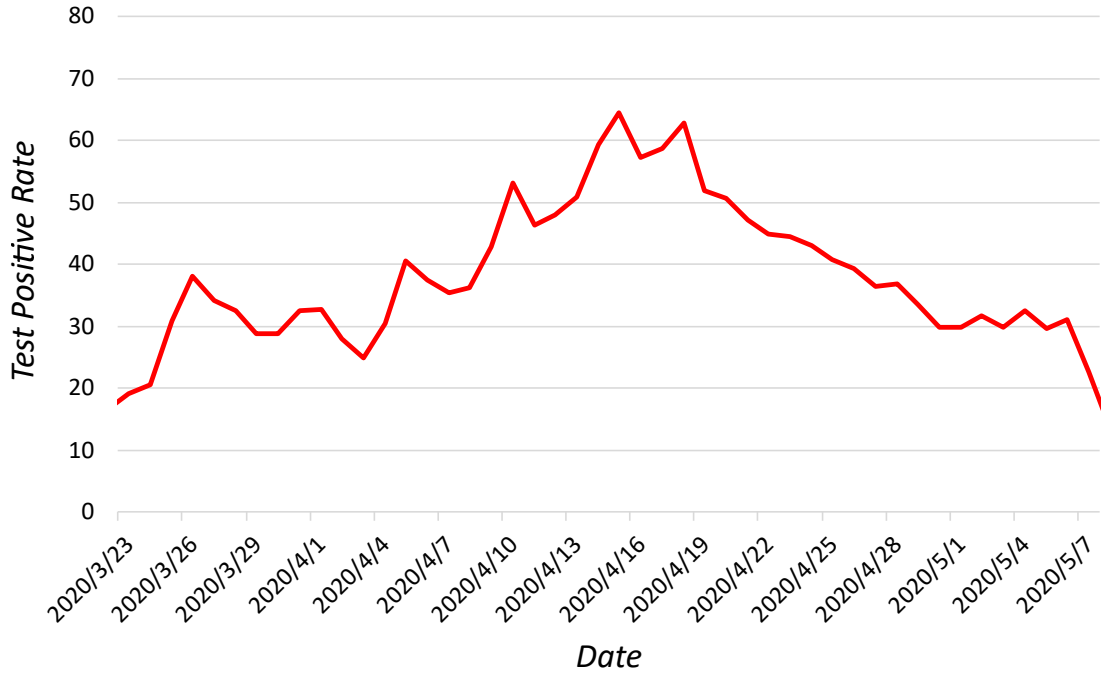


Figure 8.11: Changes in the rate of positive PCR tests in Tokyo within the period it was included in Cluster 3 and 5.

9 Conclusion

In the first half of this paper, we proposed the AESM, which is a method that combines Tango's alpha and the spatial hierarchical structure obtained by the echelon analysis, as a new method that can detect high-risk clusters. In addition, we compared the detection accuracy of the AESM with existing methods by simulation and evaluated them visually. The simulation result showed that the proposed method has ICDA equal to or higher than that of the existing method and reduced the analysis time for large-scale spatial data. Therefore, the AESM is effective as a cluster detection method for large-scale spatial data. From this result, we expect the AESM can detect clusters of large-scale spatial data, which has been difficult to apply by existing methods, even though it has been collected in various fields, and it is possible to obtain new knowledge.

In the second half of this paper, as real data analysis, we applied the cylindrical scan method and the AESM to detect space-time clusters in COVID-19 infection data in Japan. The results of analysis show population-based clusters in densely populated and well-traveled areas, such as Tokyo, suggesting that a large amount of human movement in these areas is one of the factors influencing the spread of infection. Furthermore, results of an analysis based on the number of PCR tests conducted showed detected clusters during the period when the positive test rate was high. The clusters expanded to a wide range when there were more infected persons. Therefore, we emphasize that it is important to secure a sufficient number of tests to be prepared for the increase in the number of infected people, which can be achieved by establishing cooperative relationships between the medical systems of each prefecture. However, the properties of each of the clusters may differ. Therefore, it is necessary to analyze each prefecture in more detail.

Finally, we discuss future works. First, we will consider how to set the value of α in the AESM. This is a value arbitrarily determined according to the characteristics of the cluster that the analyst wants to detect. However, from the simulation results in this study, it can be seen that the detected cluster may change depending on the value of α even if the data and scan method are the same. The criteria given by Tango (2008) are just guidelines and may not apply to all data. For example, when analyzing large-scale spatial data using the AESM, the size of the detected cluster may be large even if $\alpha = 0.05$. Therefore, the analyst must determine the value of α suitable as the cluster for each data and method to be used. This problem is like the setting of MSWS in existing methods in that it must be determined by the analyst. Han et al. (2016) proposed a method using the Gini coefficient as a method for setting MSWS. We consider that it

is necessary to carefully consider the value of α in AESM by utilizing various statistical indicators including this method.

Second, we detected space-time clusters based on the retrospective method (Kulldorff 1998), which also detects clusters that had already ended at the time of the analysis. In the case of people infected with COVID-19, where the data are updated daily, it is important to identify ongoing clusters. These are referred to as "alive cluster." Kulldorff (2001) proposed the prospective method for detecting such clusters. This method can be performed with the same software as the retrospective method and apply to the analysis of various surveillance problems (Takahashi and Tango 2008). It is extremely important to capture the shape change of alive clusters; however, this is currently difficult to do using the AESM. Therefore, a new detection method is required. We consider this to be a worthwhile direction for future work.

References

Andrade, A. L., Silva, S. A., Martelli, C. M., Oliveria, R. M., Morais Neto O. L., Siqueira Junior, J. B., Melo, L. K. and Di Fabio, J. L. (2004). Population-based surveillance of pediatric pneumonia: use of spatial analysis in an urban area of Central Brazil. *Cadernos De Saude Publica*, **20**, 411–421.

Anselin, L. (1995). Local indicators of spatial association-LISA. *Geographic Analysis*, **27**(2), 93–115.

Cliff, A. D., and Ord, J. K. (1973). Spatial Autocorrelation. London: Pion.

Cordes, J., and Castro, C. M. (2020). Spatial analysis of COVID-19 clusters and contextual factors in New York City. *Spatial and Spatio-temporal Epidemiology*, **34**, Article number: 100355.

Han, J., Zhu, L., Kulldorff, M., Hostovich, S., Stinchcomb, D. G., Tatalovich, Z., Lewis, D. R. and Feuer, E. J. (2016). Using Gini coefficient to determining optimal cluster reporting sizes for spatial scan statistics. *International Journal of Health Geographics*, **15**, Article number: 27

Hohl, A., Delmelle, E. M., Desjardins, M. R., and Lan, Y. (2020). Daily surveillance of COVID-19 using the prospective space-time scan statistic in the United States. *Spatial and Spato-temporal Epidemiology*, **34**, Article number: 100354.

Ishioka, F. (2020). echelon: The Echelon Analysis and the Detection of Spatial Clusters using Echelon Scan Method, R package version 0.1.0. `https://cran.r-project.org/web/packages/echelon/index.html` (Accesed 10 January 2020.)

Ishioka, F., Kawahara, J., Mizuta, M., Minato, S., and 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., and Kurihara, K. (2012). Hotspot Detction Using Scan Method Based on Echelon Analysis. *Proceedings of Institute of Statistical Mathematics*, **60**(1), 93–108.

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

Kammerer, J. S., Shang, N., Althomsons, S. P., Haddad, M. B., Grant, J., and Navin, T. R. (2013). Using statistical methods and genotyping to detect tuberculosis outbreaks. *International Journal of Health Geographics*, **12**, 15.

Kim, S., and Castro, M. C. (2020). Spatiotemporal pattern of COVID-19 and government response in South Korea (as of May 31, 2020). *International Journal of Infectious Diseases*, **98**, 328–333.

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

Kulldorff, M. (2001). Prospctive time-periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society*, **A164**, 61–72.

Kulldorff, M., Athas, W., Feuer, E., Miller, B., and Key, C. (1998) Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos. *American Journal of Public Health*, **88**, 1377–1380.

Kulldorff, M., and Harvard Medical School, Boston and Information Management Services Inc. (2022). SaTScan^{TM}v10.0.2: Software for the spatial and space-time scan statistics. `http://www.satscan.org/` (Accesed 11 January, 2022.)

Kurihara, K. (2004). Classification of geospatial lattice data and their graphical representation. In D. Banks et al. (Eds), *Classification, clustering, and data mining applications*, 251–258, Springer. Berlin.

Kurihara, K., Ishioka, F., and 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.

Manabe, T., Yamaoka, K., Tango, T., Binh, G. N., Co, X. D., Tuan, D. N., Izumi, S., Takasaki, J., Chau, Q. N., and Kudo, K. (2016). Chronological, geographical, and seasonal trends of human cases of avian influenza A (H5N1) in Vietnam, 2003-2014: a spatial analysis. *BMC Infectious Diseases*, **16**, Article number: 64.

Martines, M. R., Ferreira, R. V., Toppa, R. H., Assuncao, L., Desjardins, M. R., and Delmelle E. M. (2021). Detecting space-time clusters of COVID-19 in Brazil: mortality, inequality, socieconomic vulnerability, and the relative risk of the disease in Brazillian municipalities. *Journal of Geographical Systems*, **23**, 7–36.

Moran, P. A. P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B*, **10**(2), 243–251.

Myers, W. L., Patil, G. P., and Joly, K. (1997). Echelon approach to areas of concern in synoptic regional monitoring. *Environmental and Ecological Statistics*, **4**(2), 131–152.

Oeltmann, J. E., Varma, J. K., Ortega, L., Liu, Y., O'Rourke, T., Cano, M., Harrington, T., Toney, S., Jones, W., Karuchit, S., Diem, L., Rienthong, D., Tappero, J. W., Ijaz, K., and Maloney, S. (2008). Multidrug-Resistant Tuberculosis Outbreak among US-bound Hmong Refugees, Thailand, 2005. *Emerging Infectious Diseases*, **14**, 1715–1721.

Otani, T. and Takahashi, K. (2019). rflexscan: The Flexible Spatial Scan Statistic, R package version 0.2.0. `https://cran.r-project.org/package=rflexscan` (Accesed 13 November, 2019.)

Patil, G. P., and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Echological Statistics*, **11**(2), 183–197.

Takahashi, K., and Tango, T. (2008). Syndoromic Surveillance Using a Cluster Detection Test. *Journal of the Natinal Institute of Public Health*, **57**(2), 122–129.

Takahashi, K., Yokoyama, T., and Tango, T. (2013). FleXScan v3.1.2: Software for the Flexible Scan Statistic. `https://sites.google.com/site/flexscansoftware/` (Accesed 10 April, 2018.)

Tango, T. (2008). A spatial scan statistic with a restricted likelihood ratio. *Japanese Journal of Biometrics*, **29**(2), 75–95.

Tango, T., and Takahashi, K. (2005). A flexible scan statistic for detecting clusters. *International Journal of Health Geographics*, **4**, Article number: 11.

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

Acknowledgements

I would like to express my sincere gratitude to Associate Professor Fumio Ishioka of Graduate School of Environmental and Life Science, 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 Professor Koji Kurihara of Graduate School of Environmental and Life Science, Okayama University, for his valuable teaching, advice and encouragement on my study.

Furthermore, grateful appreciation and sincere thanks are extended to Professor Wataru Sakamoto of Graduate School of Environmental and Life Science, Okayama University, for his teaching, suggestions and encouragement.

Finally, I am thankful to all people who gave me encouragement, support, and advice warmly throughout my doctor course.