Japanese Journal of Statistics and Data Science (2022) 5:279–301
https://doi.org/10.1007/s42081-022-00159-x

ORIGINAL PAPER

Statistics for COVID-19 Pandemic Data

Detection of space–time clusters using a topological hierarchy for geospatial data on COVID-19 in Japan

Yusuke Takemura[1] · Fumio Ishioka[2] · Koji Kurihara[3]

Received: 13 July 2021 / Revised: 15 March 2022 / Accepted: 1 April 2022 /
Published online: 12 May 2022

Abstract

In this paper, we detected space–time clusters using data on coronavirus disease 2019 (COVID-19) collected daily by each prefecture in Japan. COVID-19 has spread globally since the first confirmed case in China, in December 2019. Several people have to date been infected in Japan since the first confirmed case in January 2020. The outbreak of COVID-19 has had a significant impact on many people's lives. Studies are being conducted to detect regions, called clusters, which pose a significantly higher risk of infection than their surrounding areas, based on a spatial scan statistics of COVID-19 infections. Among these studies, space–time cluster detection has to date been actively performed to gain knowledge regarding infection status. Based on the spatial scan statistic, the cylindrical scan method is a widely used space–time cluster detection method. This method enables concurrent detection of the location and time of a cluster occurrence. However, this method cannot capture spatial changes in a cluster over time. When applying the existing method to a cluster whose shape changes over time, the number of calculations required becomes extremely large, and the analysis may become difficult. In this study, we focused on the hierarchical structure of the data obtained by conducting an echelon analysis and applied the space–time cluster detection method based on this structure to enable the capture of changes in a cluster's shape. Furthermore, we visualized the location and period of a cluster's occurrence and considered the cause of the cluster.

✉ Yusuke Takemura
yutakemu@mail.doshisha.ac.jp

Fumio Ishioka
fishioka@okayama-u.ac.jp

Koji Kurihara
kurihark@kyoto-wu.ac.jp

[1] Organization for Research Initiatives and Development, Doshisha University, Kyoto, Japan

[2] Faculty of Environmental and Life Science, Okayama University, Okayama, Japan

[3] Institute of Data Science, Kyoto Women's University, Kyoto, Japan

Keywords COVID-19 · Spatial scan statistic · Space–time cluster · Echelon analysis

1 Introduction

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, including research into detecting regions that have a significantly higher risk of infection than the surrounding areas. Detection of spatial clusters is very important for understanding the current status of infections and the factors involved in the spread of infection. To date, as methods for evaluating the presence or absence of a cluster, evaluation from the perspective of spatial autocorrelation (Moran 1948; Cliff and Ord 1973; Anselin 1995) and identification of the cluster position (Kulldorff 1997; Tango and Takahashi 2005; Ishioka et al. 2019) have been proposed. In particular, the spatial scan statistic (Kulldorff 1997) has been widely used for the detection of 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) used it to detect clusters of COVID-19 infections in New York City.

Clusters are often detected from the cumulative number of observations made in a specific period within the study area. It is important to simultaneously detect the location and duration of clusters from the number of observations that span multiple periods, such as the daily number of people infected with COVID-19. Kulldorff et al. (1998) proposed a method for detecting a space–time cluster based on the spatial scan statistic. The SaTScan$^{\text{TM}}$ software (the latest version is 10.0; Kulldorff 2021) can perform this method. Research to detect space–time clusters of COVID-19 infections using Kulldorff's method is currently underway (Hohl et al. 2020; Kim and Castro 2020; Martines et al. 2021).

The detection of a space–time cluster can be used to capture information regarding the status and the spread of infection up to a specific date. However, Kulldorff's method can only detect a cluster comprising the same regional area 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, analysis using any of the existing methods becomes difficult because of an increase in the number of calculations required.

In this paper, we detect space–time clusters using the scanning method proposed by Takemura et al. (2021) using the data on COVID-19-infected people data collected daily by each prefecture in Japan. Additionally, we consider the factors that caused the detected clusters and the changes in a cluster's shape.

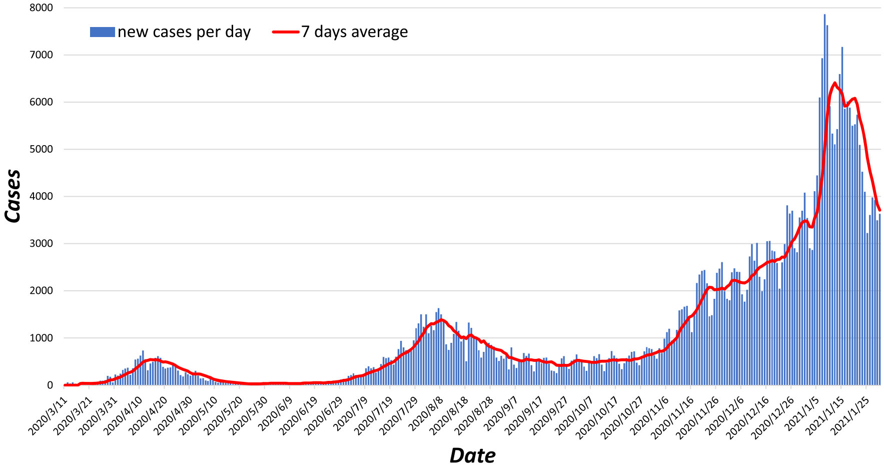


Fig. 1 Number of daily COVID-19 cases from March 11, 2020, to January 30, 2021

Section 2 introduces the data used to analyze and the spatial scan statistic. We then describe two types of methods for detecting space–time clusters. Section 3 shows Japan’s space–time cluster detection results, as obtained by the methods described here. In Sect. 4, we discuss these results. Section 5 provides conclusions to this paper.

2 Data and methods

2.1 Data on COVID-19-infected people 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 1 features a graph showing the number of newly infected people in Japan and the moving average for this number over the preceding 7 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 7863 on January 08, 2021.

2.2 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 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. A subset of regions adjacent to each other in the study area is called *a window* and represented by $\mathbf{Z}$. Let $O(\mathbf{Z})$ and $\xi(\mathbf{Z})$ be the random variable for the number of cases and the expected number of cases, respectively, within window $\mathbf{Z}$. The presence or absence of a cluster can then be given as the following hypothesis testing:

$$\begin{aligned} H_0 &: E(O(\mathbf{Z})) = \xi(\mathbf{Z}), \quad {}^{\forall}\mathbf{Z} \in \mathcal{Z} \\ H_1 &: E(O(\mathbf{Z})) > \xi(\mathbf{Z}), \quad {}^{\exists}\mathbf{Z} \in \mathcal{Z}, \end{aligned}$$

where $\mathcal{Z}$ is the universal set of $\mathbf{Z}$. At this time, performing the test for each $\mathbf{Z}$ gives rise to the problem of conducting multiple testing. Consequently, the likelihood ratio test statistic is given as follows:

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

where $o(\mathbf{Z})$ denotes the observed number of cases in window $\mathbf{Z}$, and $\mathbf{Z}^c$ is the complement of $\mathbf{Z}$. Typically, $\log \lambda_K(\mathbf{Z})$ is used to simplify the calculation. The window $\mathbf{Z}$ that maximizes the value of $\log \lambda_K(\mathbf{Z})$ is defined as the most likely cluster (MLC). The significance of the MLC is evaluated using the Monte Carlo method.

Let o_i be the number of cases observed in region i; it is desirable that a region included in the cluster be a high-risk region that satisfies $o_i > \xi_i$ when the cluster detection is performed using infectious disease data (e.g., COVID-19 data). However, since the spatial scan statistic is calculated based on $\mathbf{Z}$, which is a set that includes region 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

$$\lambda_T(\mathbf{Z}) = \begin{cases} \left(\dfrac{o(\mathbf{Z})}{\xi(\mathbf{Z})}\right)^{o(\mathbf{Z})} \left(\dfrac{o(\mathbf{Z}^c)}{\xi(\mathbf{Z}^c)}\right)^{o(\mathbf{Z}^c)}, & (o(\mathbf{Z}) > \xi(\mathbf{Z}), p_i < \alpha, {}^{\forall}i \in \mathbf{Z}) \\ 1, & (\text{otherwise}), \end{cases} \tag{2}$$

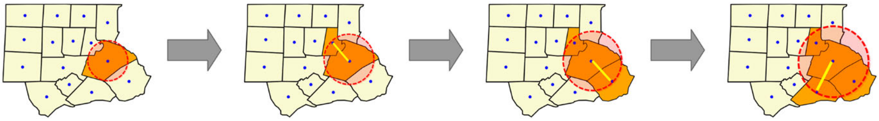


Fig. 2 Scanning process of the circular scan method

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

$$p_i = \Pr\{O_i \geq o_i + 1 \mid O_i \sim \mathrm{Pois}(\xi_i)\} + \frac{1}{2}\Pr\{O_i = o_i \mid O_i \sim \mathrm{Pois}(\xi_i)\} \quad (3)$$

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.

Tango's statistic considers each region's risk rate, thereby including only the regions that satisfy $o_i > \xi_i$ into the MLC.

2.3 Space–time cluster detection using the cylindrical scan method

The circular scan method (Kulldorff 1997) is widely used to scan **Z**. In this method, as shown in Fig. 2, a circular window expands from the representative point of the region to a user-defined limit. The regions within this range are sequentially included in **Z**. Based on this approach, the cylindrical scan method (Kulldorff et al. 1998) was used for detecting space–time clusters using a cylindrical window with a circular geographic base where the height corresponds to time. 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 same regional group are detectable. Accordingly (see Fig. 3), detection becomes difficult when the cluster's shape changes over time (Patil and Taillie 2004).

2.4 Space–time cluster detection based on the Echelon scan method

In the case of infectious diseases such as COVID-19, the disease may spread to the area 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. The Echelon scan method (Ishioka et al. 2007, 2019) searches for a cluster using the hierarchical structure of the spatial data obtained by conducting an Echelon analysis (Myers et al. 1997; Kurihara 2004; Kurihara et al. 2020). Echelon analysis is a method that systematically and objectively visualizes the topological

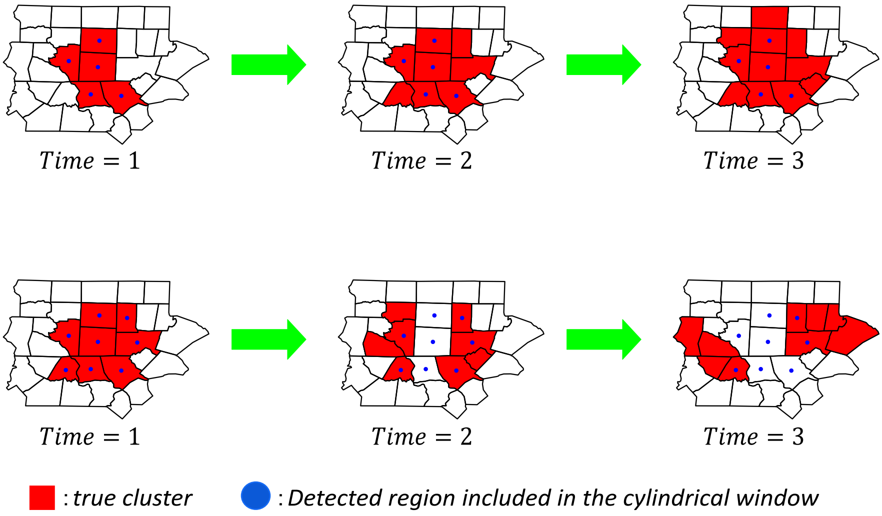


Fig. 3 Example of an expanding cluster (upper) and a dividing cluster (lower). The red indicates the regions included in the true space–time cluster. In real data, the true cluster may change over time, for example, when the number of regions included in a true cluster increases and its scale expands or when the cluster is divided into multiple clusters that move. However, the regions with blue dots, those detected by the cylindrical scan method, do not change over time. Therefore, because it cannot capture changes in these clusters, the true cluster is only partially detected, or regions not included in the true cluster are mistakenly detected by this method

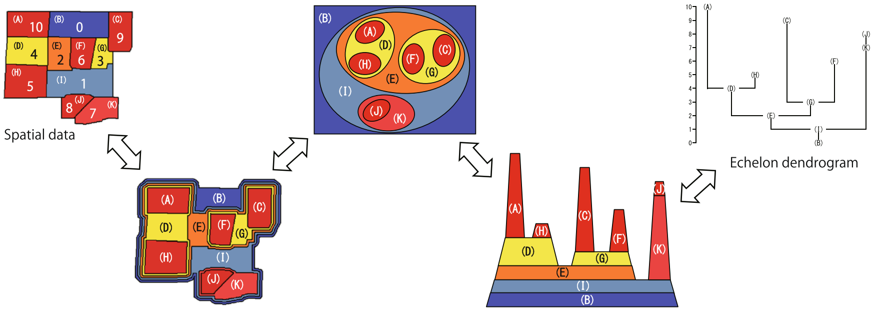


Fig. 4 Flow in the Echelon dendrogram created using the Echelon analysis method

structure of spatial data by dividing the spatial position based on the height of the surface for the univariate value of each region. Figure 4 shows the flow of the Echelon analysis; the structure of the spatial data obtained by the Echelon analysis is represented by a graph called the Echelon dendrogram. With the Echelon scan method, scanning is preferentially performed from the regions that constitute the upper hierarchies of the Echelon dendrogram, called the peak. In this way, it is possible to detect clusters with arbitrary shapes.

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. 5. Here, the attribute value of each region

	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

Fig. 5 Sample of spatiotemporal data

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 5d shows the location ID for each region. These data can be considered the spatial data of 75 regions (25 regions × 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

$$NB(l(i,t)) = \begin{cases} \{l(k,t) \mid \text{region } i \text{ and } k \text{ are neighbors}\} \cup l(i, t+1), & t = 1 \\ \{l(k,t) \mid \text{region } i \text{ and } k \text{ are neighbors}\} \cup l(i, t+1) \cup l(i, t-1), & t = 2 \\ \{l(k,t) \mid \text{region } i \text{ and } k \text{ are neighbors}\} \cup l(i, t-1), & t = 3, \end{cases} \tag{4}$$

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 6 shows the Echelon dendrogram for the data when the spatial adjacency

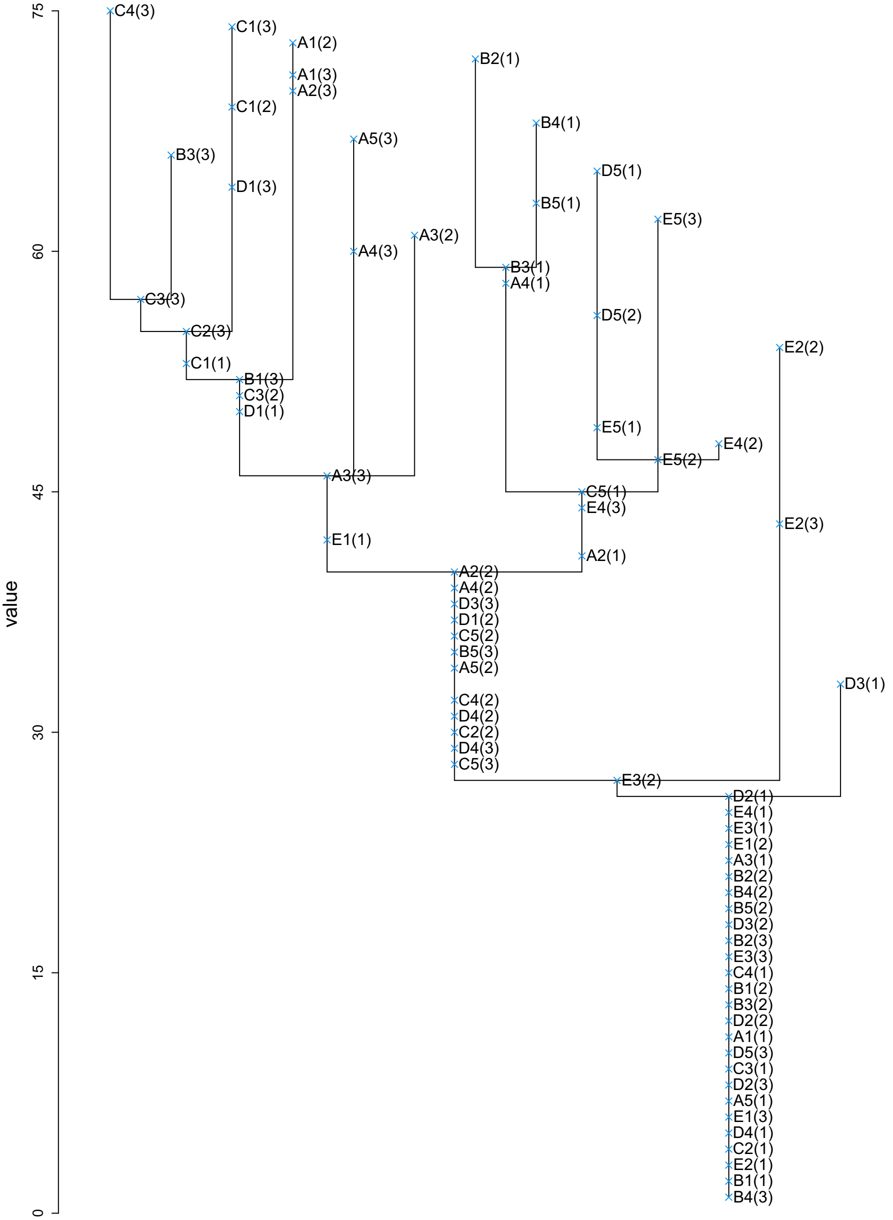


Fig. 6 Echelon dendrogram for the sample data

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 are 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. Additionally, the method scans down to the lower hierarchies of the dendrogram, which include the regions that satisfy $o_i < \xi_i$. The regions that should be detectable as a cluster are generally included in the upper hierarchies. For this problem, Takemura et al. (2021) proposed an improved technique (hereafter called the adjusted Echelon scan method [AESM]) for the Echelon scan method using p_i-value and Tango's α. 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:

$$p_{i,t} = \Pr\{O_{i,t} \geq o_{i,t} + 1 \mid O_{i,t} \sim \text{Pois}(\xi_{i,t})\} + \frac{1}{2}\Pr\{O_{i,t} = o_{i,t} \mid O_{i,t} \sim \text{Pois}(\xi_{i,t})\}, \tag{5}$$

where $O_{i,t}$ and $o_{i,t}$ are the random variable of cases and the observed number of cases, respectively, in region i at time point t, and $\xi_{i,t}$ is the expected number of cases in region i at time point t. In the AESM, the upper hierarchies of the spatiotemporal data are extracted using $p_{i,t}$ and Tango's α, and the Echelon scan method is applied to the extracted data. Specifically, the steps followed in this process are:

Step 1. Extract the data of region i at time point t that satisfies $p_{i,t} < \alpha$ 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 hierarchy of the dendrogram is taken into $\mathbf{Z}$ in order, and $\mathbf{Z}$, which maximizes $\log \lambda_K(\mathbf{Z})$, is the MLC.

Figure 7 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.

3 COVID-19 data analysis

3.1 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. 2 to detect space–time clusters

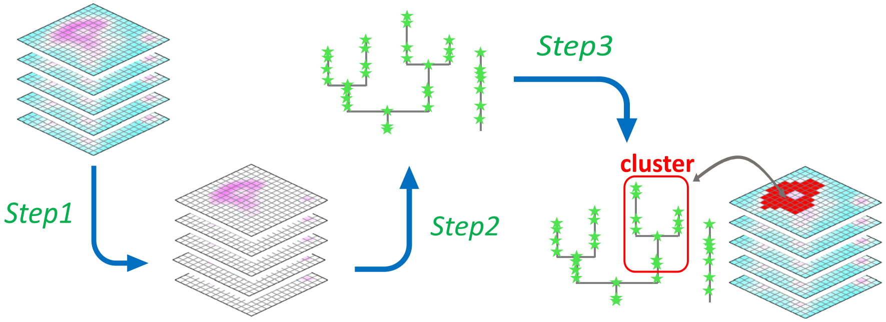


Fig. 7 Flow of space–time cluster detection using the AESM. The upper left figure represents spatiotemporal data, where red-colored regions are high-risk and satisfy $p_{i,t} < \alpha$ and blue-colored regions do not satisfy $p_{i,t} < \alpha$. In step 1, only the red-colored regions are extracted from 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

based on population. We collected the data of residents in each prefecture as the population data. We first used the SaTScan$^{\text{TM}}$ software to apply the cylindrical scan method. We then created a new function in R for applying the AESM based on the function implemented in the existing echelon package (Ishioka 2020), which is R package.

The setting of each method is described as follows. In the cylindrical scan method, we restricted the spatial window size 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 10–20% of the total population, which made it easy to interpret the results. In addition, we restricted the temporal window size 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 that 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 the maximum window size 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

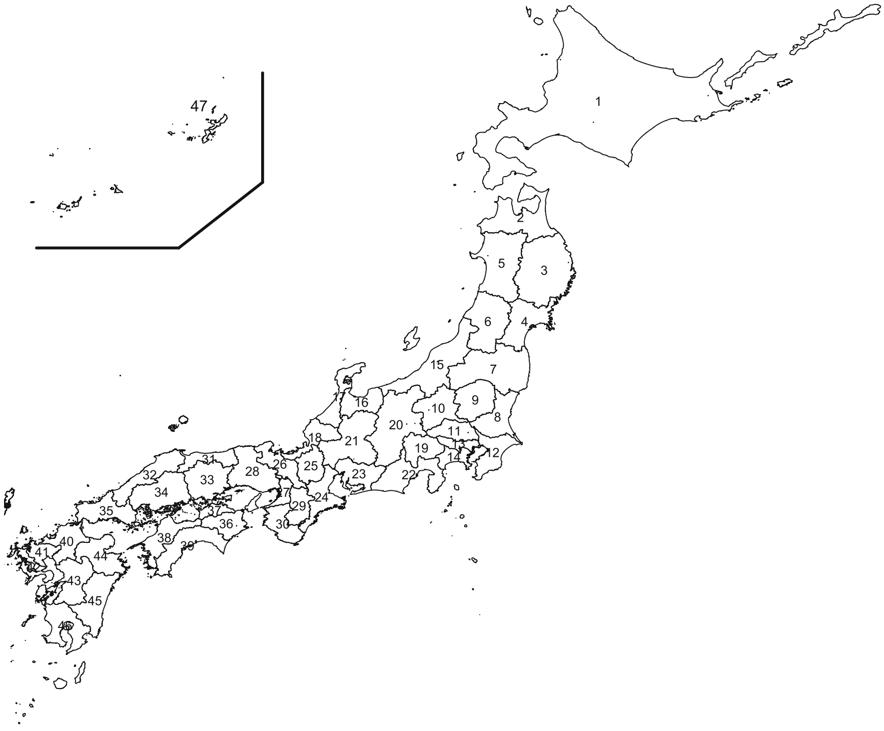


Fig. 8 Geographical location of each prefecture in Japan; Okinawa (No.47), shown in the upper left of the figure, is actually located in the southwestern part of Japan

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

$$\theta_{i,t} = \frac{o_{i,t}}{\xi_{i,t}}. \tag{6}$$

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

$$\xi_{i,t} = w_{i,t} \times \frac{\sum_{i=1}^{47} o_{i,t}}{\sum_{i=1}^{47} w_{i,t}}, \tag{7}$$

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 these 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. These data were included, because that Okinawa does not have geographically adjacent prefectures. Figure 8 shows the geographical location of each prefecture in Japan, and Table 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:

$$
\begin{aligned}
&NB(l(i,t))\\
&= \begin{cases}
\{l(k,t) \mid \text{region } i \text{ and } k \text{ are neighbor}\} \cup l(i,t+1), & t = 1\\
\{l(k,t) \mid \text{region } i \text{ and } k \text{ are neighbor}\} \cup l(i,t+1) \cup l(i,t-1), & 1 < t < 326\\
\{l(k,t) \mid \text{region } i \text{ and } k \text{ are neighbor}\} \cup l(i,t-1), & t = 326,
\end{cases}
\end{aligned}
\tag{8}
$$

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

The analytical results using the cylindrical scan method with the above settings are shown in Table 2 and Fig. 9a, and the results from the AESM are shown in Table 3 and Fig. 9b. Figure 9 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 2 and 3 lists the relative risk (RR), which is calculated as follows:

$$
RR = \frac{o(\mathbf{Z})/\xi(\mathbf{Z})}{o(\mathbf{Z}^c)/\xi(\mathbf{Z}^c)}. \tag{9}
$$

Figure 10 visualizes each prefecture; the numbered areas in the figure are the prefectures that were included as a cluster, even if only for 1 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. "Secondary clusters" refers to clusters other than the MLC that were judged to have a significantly high value of $\log \lambda_K(\mathbf{Z})$. Table 2 and Fig. 9a show that clusters (excluding cluster 5) were detected for an extended period, and the MLC was a cluster that lasted approximately 5 months. In cluster 4, which was detected in Okinawa, $RR = 3.79$ (see Table 2), indicating that it was a high-risk cluster, but, as seen in Fig. 9a, 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. 9b, 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.

Table 1 Neighboring information for each prefecture

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

Table 1 continued

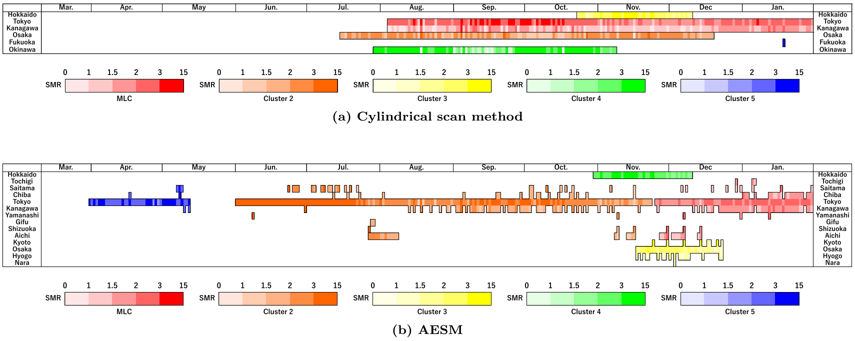
No.	Location	Neighbors
40	Fukuoka	35, 41, 42, 43, 44
41	Saga	40, 42
42	Nagasaki	40, 41, 43
43	Kumamoto	40, 42, 44, 45, 46
44	Oita	35, 38, 40, 43, 45
45	Miyazaki	43, 44, 46
46	Kagoshima	43, 45, 47
47	Okinawa	46

Table 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

3.2 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



(a) Cylindrical scan method

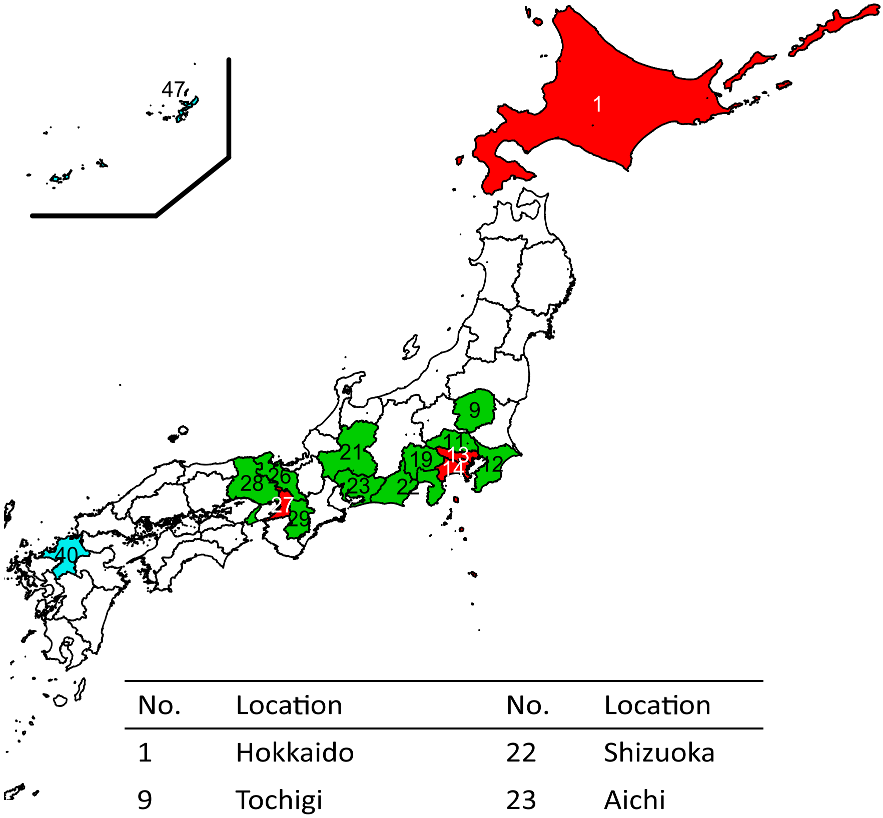
(b) AESM

Fig. 9 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 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

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. 2. 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



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

Fig. 10 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

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 3564 on day 12, then by calculating 3564/12 = 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. 3.1, 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 4 and Fig. 11. The numbered prefectures shown in Fig. 12 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. 11 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 $RR = 4.34$.

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. 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. 9a, 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 that the cluster expansion and contraction would be reflected in the areas surrounding Tokyo. Figure 1 shows the number of infected people rapidly increasing in late December 2020. Figure 9b 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

Table 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	12/22–1/29	13,308.03	83,852	47,578.69	1.97
	Tochigi					
	Gunma					
	Saitama					
	Chiba					
	Tokyo					
	Kanagawa					
	Yamanashi					
	Shizuoka					
Cluster 2	Ibaraki	11/10–12/21	3581.24	28,432	16,710.80	1.75
	Tochigi					
	Gunma					
	Saitama					
	Chiba					
	Tokyo					
	Kanagawa					
	Yamanashi					
	Shizuoka					
	Aichi					
	Mie					
	Kyoto					
	Osaka					
	Hyogo					
	Nara					
	Wakayama					
	Tokushima					
Cluster 3	Saitama	4/15–5/7	2069.27	2970	687.80	4.34
	Chiba					
	Tokyo					
	Kanagawa					
Cluster 4	Shizuoka	12/2–12/10	1145.93	1902	498.52	3.83
	Aichi					
Cluster 5	Ibaraki	3/23–4/11	940.47	2213	746.89	2.97
	Tochigi					
	Saitama					
	Chiba					
	Tokyo					
	Kanagawa					
	Shizuoka					

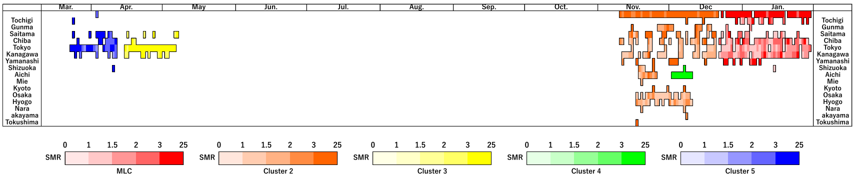
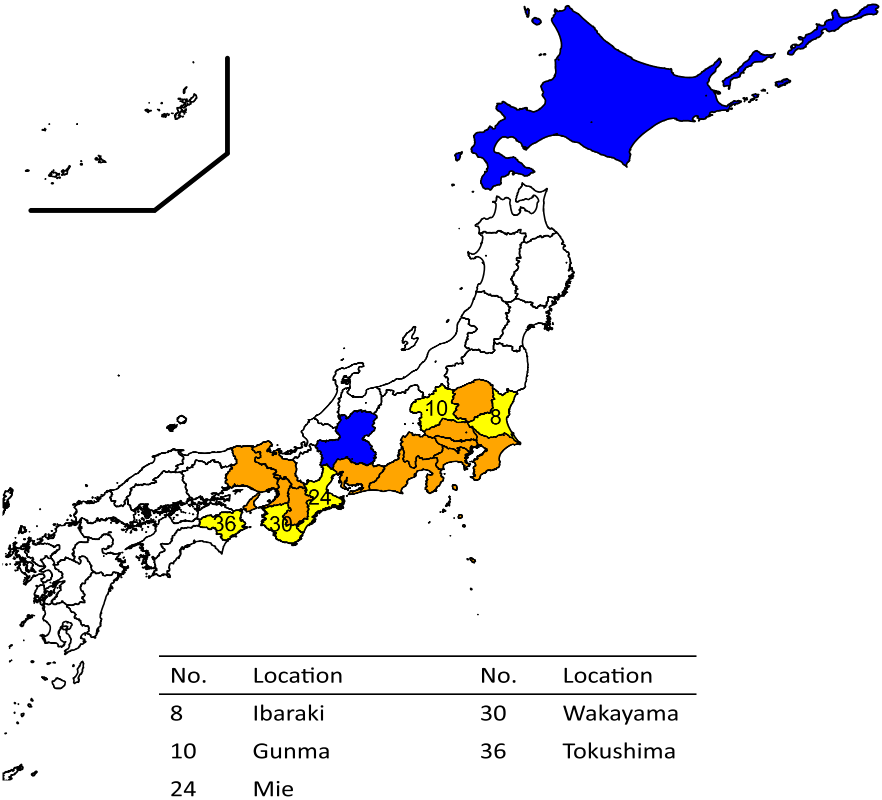


Fig. 11 Space–time clusters based on the number of PCR tests



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

Fig. 12 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

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 9b shows that cluster 2 included Tokyo in late June, and Fig. 13 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.

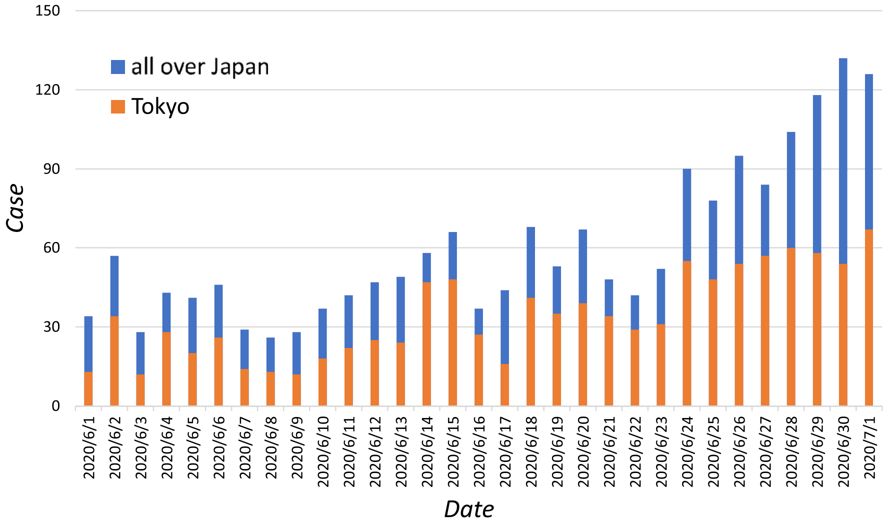


Fig. 13 Number of daily cases throughout Japan and in Tokyo from June 1 to July 1, 2020

Next, we considered the space–time clusters based on the number of PCR tests. Figure 11 shows that the clusters detected based on these tests lasted for approximately 1 month. It is also shown that cluster 2 expanded to an extremely wide area, including the regions surrounding Osaka and Tokyo. Figure 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 14 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, 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. 15 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. In this study, we used the data described in Sect. 2.1 and calculated the positive test rate. However, the values may differ from those published by each prefecture due to the base date of the data and the processing for the number of PCR tests described in Sect. 3.2. For example, in the case of Tokyo, we calculated the positive test rate using positive cases based on the day of notification from the public health center. On the other hand, Tokyo Metropolitan Government's monitoring site (https://stopcovid19.metro.tokyo.lg.jp/en/cards/positive-rate/) publishes the values using positive cases based on the day when the test result was confirmed. Therefore, it should be noted that there is a difference between the positive test rate in this paper

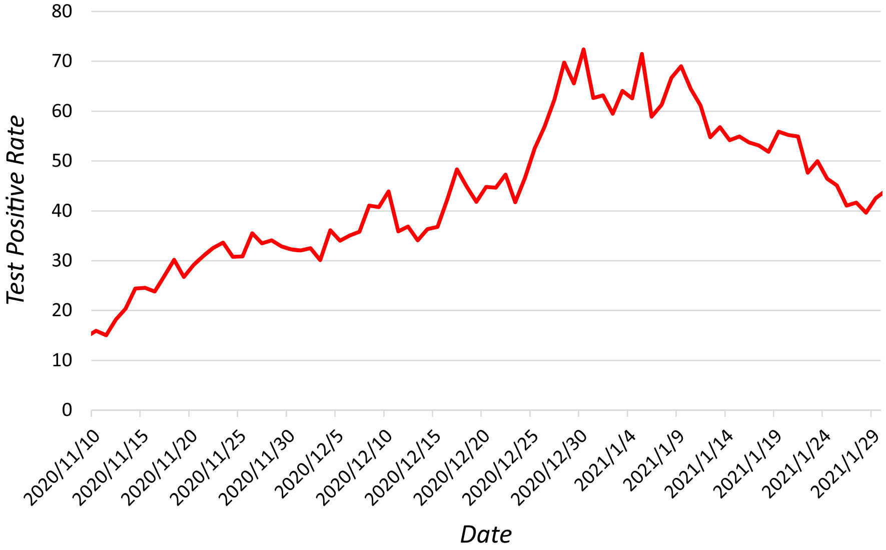


Fig. 14 Changes in the rate of positive PCR tests in Ibaraki within the period it was included in MLC and cluster 2

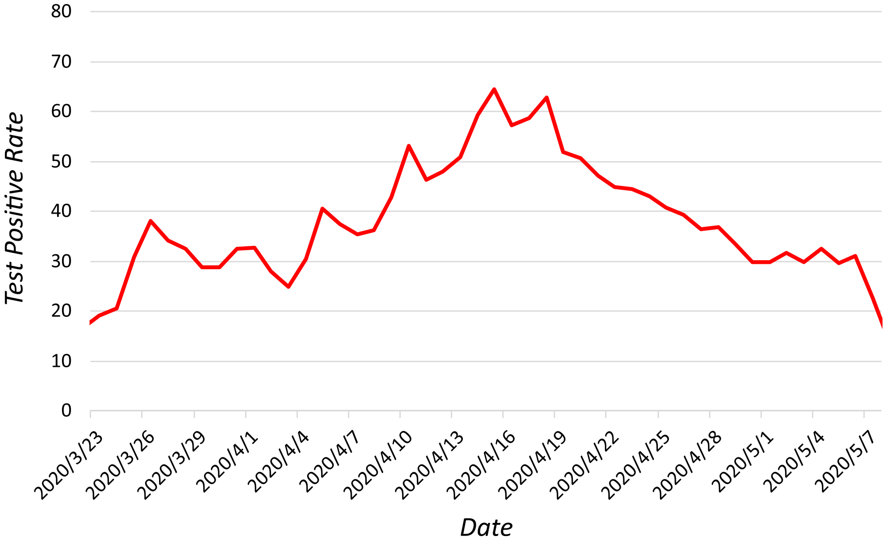


Fig. 15 Changes in the rate of positive PCR tests in Tokyo within the period it was included in cluster 3 and 5

and on this site. Like Tokyo and Ibaraki, 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.

5 Conclusion

In this study, we applied the cylindrical scan method and the AESM to detect space–time clusters in COVID-19 infection data in Japan. The results of the 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.

We detected space–time clusters based on the retrospective method (Kulldorff et al. 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. 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.

Acknowledgements This work was partly supported by JSPS KAKENHI Grant Numbers JP21K11786 and JP20K11710.

References

Andrade, A. L., Silva, S. A., Martelli, C. M., Oliveria, R. M., MoraisNeto, O. L., SiqueiraJunior, J. B., Melo, L. K., & 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., & Ord, J. K. (1973). *Spatial Autocorrelation*. London: Pion.

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

Hohl, A., Delmelle, E. M., Desjardins, M. R., & Lan, Y. (2020). Daily surveillance of COVID-19 using the prospective space-time scan statistic in the United States. *Spatial and Spatio-Temporal Epidemiology, 34*, 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. Accessed 10 Jan 2020.

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

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

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

Kim, S., & 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 statistic. *Communications in Statistics, Theory and Methods, 26*, 1481–1496.

Kulldorff, M. (2001). Prospective 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., & 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., & Harvard Medical School, Boston and Information Management Services Inc. (2021). SaTScan^{TM}v10.0: Software for the spatial and space-time scan statistics. http://www.satscan.org/. Accessed 11 Oct 2021.

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

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

Manabe, T., Yamaoka, K., Tango, T., Binh, G. N., Co, X. D., Tuan, D. N., Izumi, S., Takasaki, J., Chau, Q. N., & 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*, 64.

Martines, M. R., Ferreira, R. V., Toppa, R. H., Assuncao, L., Desjardins, M. R., & Delmelle, E. M. (2021). Detecting space–time clusters of COVID-19 in Brazil: Mortality, inequality, socioeconomic vulnerability, and the relative risk of the disease in Brazilian 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., & 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., & Maloney, S. (2008). Multidrug-resistant tuberculosis outbreak among US-bound Hmong refugees, Thailand, 2005. *Emerging Infectious Diseases, 14*, 1715–1721.

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

Takemura, Y., Ishioka, F., & Kurihara, K. (2021). Detection of spatial clusters with high-risk regions by using restricted hierarchical structure. *Bulletin of the Computational Statistics of Japan, 34*(1), 23–43.

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

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