Spatio-Temporal Cluster Detection of Point Events by Hierarchical Search of Adjacent Area Unit Combinations

Ryo Inoue[1], Shiho Kasuya and Takuya Watanabe

[1]Tohoku University, Sendai, Japan
email corresponding author: rinoue@plan.civil.tohoku.ac.jp

Abstract

Several methods for detecting spatial or spatio-temporal clusters of point events, based on the "space scan statistics" presented by Kulldorff and Nagarwalla (1995), have been proposed. In most of them, it is implicitly assumed that the propagation process starts from a certain point, and an attempt is made to extract a circular-like spatial region as a cluster. However, when analyzing socio-economic phenomenon in small areas, the assumption of an isotropic propagation process is not always appropriate, since the shape of a transport network and the non-homogeneity of attributes in space affects the speed of propagation significantly. In this paper, a method for detecting flexible-shaped clusters by searching combinations of adjacent area units hierarchically is proposed. The applicability of the proposed method to the detection of spatio-temporal clusters is tested on transaction location and time data of real estate in Tokyo.

1. Introduction

Nowadays, it is becoming easier to utilize a large amount of information with detailed spatial location data for regional analysis because of improved access to statistical information, promotion of open governmental policies, and popularization of geospatial technologies such as positioning

and mapping. Moreover, some spatial information is time-stamped; therefore, the environment for spatio-temporal analysis is ready.

One of the analyses that use spatio-temporal data is cluster detection. This study focuses on the cluster detection of point events for which position and time information is available. One of the cluster detection methods that deal with spatio-temporal point event data is "space time scan statistics," developed in the field of spatial epidemiology (Kulldorff et al., 1998; Kulldorff, 2001; Kulldorff et al., 2005). This method can detect a cluster region and period, and it is used to analyze the regional disparity and temporal transition of the distribution of point events. It is an extension of "space scan statistics," which detects spatial clusters of point events (Kulldorff and Nagarwalla, 1995; Kulldorff, 1997). Both space and space-time scan statistics are quite famous and are in use, being implemented in and distributed as free software called "SaTScan" (Kulldorff, 2010).

Although several extensions of scan statistics have been proposed thus far, the major differences among them lie in their settings of the spatial shape of clusters. There are two types of settings for the cluster shapes. One assumes that the point distribution is approximately based on the isotropic propagation process starting from a certain point, and outputs circle-like clusters (Kulldorff and Nagarwalla, 1995; Kulldorff et al., 2006; Tango and Takahashi, 2005). The other type does not assume an isotropic propagation process of point events, and outputs flexibly shaped clusters (Duczmal and Assunção, 2004; Yao et al., 2011). It is quite natural to consider that the shape of a transportation network may affect the propagation process of events caused by human activities. In particular, when a small-area analysis is performed, the assumption that the propagation process is isotropic is not suitable, since the attributes of spatial location have some randomness.

Duczmal and Assunção (2004) and Yao et al. (2011) proposed a method of the latter type. It uses the count data of point events according to certain spatial units, sets the combinations of adjacent spatial units as cluster candidates, and searches the cluster using the simulated annealing method. When the number of spatial units is small, it allows for a significant cluster to be searched in a short calculation time; however, when the number of spatial units becomes large, it becomes difficult to execute the cluster searching process quickly. The expansion of the method to treat spatio-temporal data is also difficult; the setting of temperatures for the simulated annealing method becomes complex.

In this study, we propose a new, fast detection method for spatio-temporal clusters that have flexible shapes. The cluster evaluation is based on previous space and space-time statistical approaches (Kulldorff and Nagarwalla, 1995; Kulldoff et al. 1998), and the settings of the window shape are based

on Duczmal and Assunção (2004) and Yao et al. (2011). By constructing a combination of adjacent spatial units hierarchically, the method is able to detect clusters with a relatively short calculation time. The proposed method is evaluated by an application to the real estate transaction place and time data observed in Tokyo.

2. Cluster detection by scan statistics

In this chapter, we review the cluster detection procedures of space and space-time scan statistics.
Since Kulldorff and Nagarwalla (1995) first proposed the "space scan statistics," several other methods have been proposed based on it. However, the differences among the previous methods are mostly in the settings of cluster shapes, as described above. The procedures of cluster detection in Kulldorff and Nagarwalla (1995) and its expansion to spatio-temporal space in Kulldorff et al. (1998) are commonly used. The procedures for cluster detection are introduced in Section 2.1, and the settings of cluster shapes in previous studies are introduced in Section 2.2.
Hereafter, in this paper, it is assumed that the positions of point events are aggregated according to certain spatial units, and the detection of clusters based on these spatial units is considered. When the exact positions of point events are observed, it might be possible to detect clusters in continuous space. However, the population that affects the occurrence of point events may not be distributed uniformly in space. For example, when searching the spatial cluster of patients with a certain disease, the cluster significance should not be judged by the number of patients over the size of the cluster; it should be judged by the ratio of the number of patients over the population in the cluster. Usually, people are not uniformly distributed in the study area; therefore, the population distribution should be considered. Additionally, the population data are usually aggregated based on certain spatial units, such as a municipality, zip code zone, or census tract, and it is difficult to express the population distribution in the study area using mathematical expressions over the continuous space. Then, scan statistics usually handles data in discrete space and not in continuous space. When handling data in discrete space that is divided into spatial units, the previous studies usually set the shape of the window, find the spatial units whose centroids are located inside the window, and then consider the total area of the spatial units found as a candidate cluster.

2.1. Procedures of cluster detection by scan statistics

The cluster detection procedure of scan statistics, proposed by Kulldorff and Nagarwalla (1995) and followed in many cluster detection methods, is given below.

1. The stochastic process of point event distribution is assumed. The Poisson point process is commonly assumed as the point process. When analyzing the spatio-temporal point event data, the point process in spatio-temporal space is assumed.
2. The shapes of windows that are cluster candidates are set. The details of the settings of window shapes are explained in Section 2.2. The whole study area is scanned by moving and resizing the windows.
3. The windows are evaluated by comparing two hypotheses, an alternative and a null hypothesis. The alternative hypothesis is that there are more point-event occurrences inside than outside the window; the null hypothesis is that there is no difference in the number of event occurrences inside and outside the window. The likelihood ratio, which is the likelihood of the alternative hypothesis in relation to that of the null hypothesis under the assumption of point process, is calculated for every window.

   Here, we briefly show the evaluation of windows based on the likelihood ratio. Let the whole study area where point events are distributed be denoted by G; a window, which is a subset of G, is denoted by Z; the outside of the window is denoted by Z^C; the number of point events in Z is denoted by $n(Z)$; the population in Z is denote by $\nu(Z)$. Assume that the point events are distributed based on the Poisson point process. The alternative hypothesis H_1 is "the density of point events is greater inside than outside the window Z" and the null hypothesis H_0 is "the intensity of point events inside and outside the window Z is the same." The likelihood ratio of the window Z, $\lambda(Z)$, is given in Equation (1).

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

   The window with the maximum likelihood ratio is selected as the most likely cluster (MLC).
4. The significance of a detected MLC is examined through a Monte Carlo simulation. The examination process first simulates the random point distribution in the study area according to the assumption of

point process; then, the cluster detection on the simulated point distribution outputs the maximum likelihood ratio when the points are randomly distributed in the study area. The significance of a detected MLC is evaluated by comparing the likelihood ratio of the MLC with the simulated maximum likelihood ratios.

2.2. Settings of cluster shapes

The first part of this section describes the methods that output simple-shaped spatial clusters, and the second part describes the method that outputs flexible-shaped clusters.

Simple-shaped cluster detection

In Kulldorff and Nagarwalla (1995) and Kulldorff (1997), a circular window is proposed, and in Kulldorff et al. (1998), Kulldorff (2001), and Kulldorff et al. (2005), a circular spatial window is extended to a cylinder spatio-temporal window by adding a temporal period to the cluster setting. To enhance the flexibility of the spatial window shape, Kulldorff et al. (2006) proposed an elliptic window, the parameters of which are the lengths of the major and minor axes and the azimuth of the major axis.
Another approach for detecting flexibly shaped clusters was presented by Tango and Takahashi (2005), who proposed setting the combination of adjacent spatial units inside a certain size of circular region. In their application, the maximum number of spatial units inside the window is limited to approximately thirty since the number of combinations of neighboring spatial units increases exponentially as the number of units to be considered increases. This approach is extended to a spatio-temporal analysis method in Takahashi et al. (2008).
Mori and Smith (2009) proposed setting windows as convex regions in the space in which the metric is the travel time on the transport network, although the significance of cluster candidates is evaluated by a method that is similar to but different from the spatial scan statistics.
These settings of window shapes are based on the assumption of cluster formation that the prevalence of point events starts from a certain location and spreads uniformly in any direction. However, especially when a small-area analysis is performed, the assumption that the diffusion process is isotropic and continuous is not suitable since some randomness in the attributes of spatial location can be observed in a small area.

Flexible-shaped cluster detection

Duczmal and Assunção (2004) proposed detecting flexible-shaped clusters by considering the combinations of adjacent spatial units as windows, and searching the MLC using simulated annealing. This method can search clusters whose shapes are flexible and deal with the flexible propagation process of point events. The method was applied to the spatial cluster detection of homicide scenes in Belo Horizonte, Brazil. It detected a long and narrow cluster region along an expressway, whose detection would not have been possible if circular windows had been used. In this application, since the total number of spatial units and point events are 240 and 273, respectively, which are not very large numbers for spatial analysis, the method outputs the cluster region and its significance test in a short calculation time. However, when the number of spatial units is large, the number of combinations of neighboring regions increases exponentially. If the cluster is spread over a broad area, it is impossible to find the highly significant cluster based on the simulated annealing method. In addition, the previous study considered only spatial clusters; when searching spatio-temporal clusters, it would therefore be difficult to define the temperature, a setting of a simulated annealing method.
Based on Duczmal and Assunção (2004), Yao et al. (2011) proposed two algorithms which limit the search area of combinations of neighboring regions. These methods might reduce the computational complexity, however it was not mentioned.

3. Proposal for spatio-temporal cluster detection with flexible spatial shapes

We propose a new spatio-temporal cluster detection method with flexible spatial shapes in this section.
The setting of the window shape for spatial extent is the same as that in Duczmal and Assunção (2004); the combinations of adjacent spatial units are considered spatial windows. Here we clarify the setting of spatial windows and the definition of MLC in this study. Let G denote the graph structure of spatial units and their adjacency in the study area. G consists of V, the set of vertices which represent spatial units, and E, the set of edges which connects the vertices representing neighboring spatial units. The setting of spatial window W in this study is given as W must be a subgraph of G ($W \subset G$) and connected. Let Ω denote the set of spatial windows Ws; then the problem to detect a spatial cluster is written as

$$\hat{W} = aug \max_{W \in \Omega} \lambda(W) \tag{2}$$

This study also consider the time domain. The setting of the window shape for the temporal extent is the same as that in Kulldorff et al. (2005); the time period is the same for all spatial areas inside the window. Let T denote the study period, and P denote the period of time in T. Then the problem to detect a spatio-temporal cluster is described as

$$\left(\hat{W}, \hat{P}\right) = aug \max_{W \in \Omega, P \in T} \lambda(W, P) \tag{3}$$

The number of combinations of adjacent spatial units increases exponentially as the number of spatial units increases. When the number of connected spatial units is N, the number of combinations of adjacent spatial units is between $N(N+1)/2$ and $\sum_{i=1}^{N} {}_{N}C_{i}$; it is not practicable to search all combinations of neighboring spatial units.

This paper proposes a new algorithm for flexible-shaped cluster detection. First, it divides the study area into small aggregation groups consisting of adjacent spatial units; then, it searches the spatial window with the maximum likelihood ratio value by combining adjacent spatial units in each aggregation group and designates them as cluster candidates. Next, it divides the study area into larger aggregation groups that consist of the cluster candidates detected at the former stage and combines the adjacent cluster candidates in each aggregation group. By repeating the cluster candidate search within the groups of adjacent candidates and expanding the aggregation groups by combining the cluster candidates detected at the former stages, this method is able to detect clusters in the whole study area without checking all combinations of adjacent spatial units. The procedures of the proposed method are given below.

1. The method removes the spatial units that do not have point events inside, since the inclusion of such spatial units always leads to lower likelihood values, as is clear from Equation (1). Other spatial units are designated as cluster candidates.
2. The method selects a candidate randomly, searches its adjacent candidates, and forms a spatial aggregation group. If the number of adjacent candidates exceeds the limit value *max_neighbors*, adjacent candidates are randomly chosen according to the limit values. *max_neighbors* limits the number of combinations of adjacent candidates to reduce the calculation time; its value is set to nine in the applications in this study; thus, the maximum number of candidates in a spatial aggregation group is at most ten. The formation of spatial ag-

gregation groups is repeated until all cluster candidates are included in some spatial aggregation groups.

3. In each spatial aggregation group, the method creates all the combinations of adjacent candidates. When calculating the likelihood ratio, it searches all the combinations of time points inside the combination of candidate as start and end times of the cluster periods, and records the highest likelihood ratio values and its cluster period.
4. The method selects the combination of candidates that has the largest likelihood ratio in each spatial aggregation unit, and records it as the cluster candidate for the next aggregation stage.
   The combinations of candidates that share borders with other candidates in other spatial aggregation groups have a chance to be a part of clusters; then, the combinations of candidates that share borders with other candidates and do not overlap with the selected candidates inside the same spatial aggregation group are also kept for the larger aggregation.
5. The method repeats Procedures 2-4 until a single candidate remains.

This method is able to reduce the calculation time by limiting the number of spatial units in the search of combinations of adjacent spatial units. However, the order in which adjacent candidates are combined affects the cluster detection results. Random selection of candidates when forming the spatial aggregation groups may lead to a failure to detect the clusters with a high likelihood value. The performance of the proposed method is evaluated in the next chapter. The source code of the proposed method is available at http://www.plan.civil.tohoku.ac.jp/inoue/research/code/.

4. Application

The applicability of the proposed method is tested on the real estate transaction-price information (Fudosan Torihiki kakaku Joho (in Japanese)), published by the Ministry of Land, Infrastructure, Transport, and Tourism of Japan (2012), in the twenty-three special wards in Tokyo. The information consists of the records of locations and dates of real estate transactions and other attributes of the transferred estates, such as size, accessibility, and the relevant regulations of the Urban Planning Act. In this study, we used the locations and dates of vacant lot transactions within the Tokyo metropolitan area from January 1999 to December 2009, and searched clustered areas and periods.

The spatial units used in this study were the city districts, cho and aza (in Japanese); the data source was the Digital Map 2500 (Spatial Data Frame-

work) published by the Geospatial Information Authority of Japan. Two spatial units were determined to be adjacent if they shared at least one node on their borders.

The spatial extents of three cases are shown in Fig. 1. Case 1 uses the data of the Setagaya ward; Case 2, the data of the Josai area (Setagaya, Suginami, Nakano, and Nerima wards); and Case 3, the data of the 23 special wards of Tokyo. Summaries of the three cases are shown in Table 1.

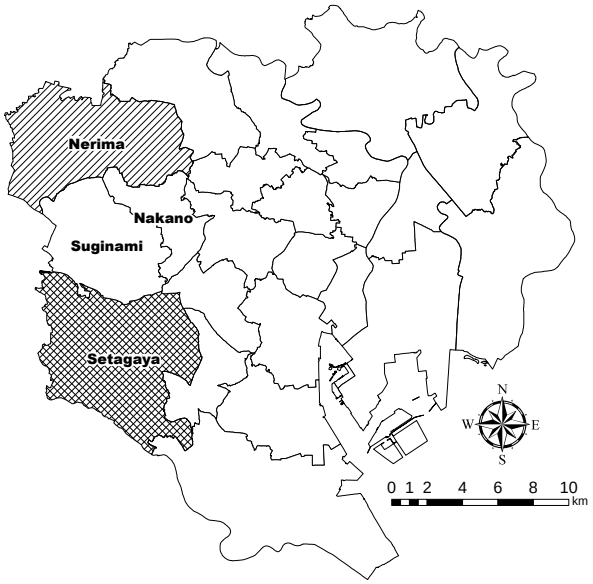


Fig. 1. Study areas–Case 1: Setagaya ward; Case 2: Josai region: Setagaya, Nakano, Suginami, and Nerima wards; Case 3: Twenty-three special wards of Tokyo

Table 1. Summary of data

	Case 1	Case 2	Case 3
Number of spatial units	279	705	3,130
Number of pairs of adjacent spatial units	833	2,171	9,834
Number of real estate transactions	2,672	8,466	26,380
Study area (km^2)	58.4	156.1	621.4

The density of transactions is calculated over the spatial area (km^2) and temporal period (years). It would be better if the density is evaluated by the number of transactions over the number of land parcels if the information on land parcels is available. The regions used in Cases 1 and 2 are residential suburbs of Tokyo. The density calculation over the size of the area in this study may not be a problem because the size of the parcels

does not vary much. However, the analysis of Case 3 may present some problems since the central business districts, where land parcels are quite large as compared those in residential areas, are included in the study area.
The cluster detection results of the proposed method are compared with those of SaTScan, the space -time scan statistics that uses circular or elliptic spatial windows and temporal periods: cylindrical or elliptical cylindrical windows in spatio-temporal space. The version of SaTScan used was 9.1.1x64, and the settings were as follows: The type of analysis was retrospective space-time; the probability model was discrete Poisson; and the time aggregation length was 7 days.
The proposed method is implemented using C++ codes with OpenMP libraries and Intel C++ compiler, whereas the SaTScan is implemented using JAVA. The calculation times of both methods are shown for reference; however, the comparison of calculation time would not be very significant since the environments of these implementations are different. The calculation times shown in this paper do not include the time taken by the significance tests using the Monte Carlo simulation. The calculations were executed on a PC with Xeon X5670 (2.93GHz/6 cores) with 8GB of memory.

4.1. Case 1 – Setagaya ward

The cluster detection results of Case 1, i.e., the analysis of data in Setagaya ward, are shown in this section. The input data, spatial units, and position and time of transactions are shown in Fig. 2. First, we examine the effect of the order of the aggregation group formation on the cluster detection results. Table 2 shows the 20 cluster detection results achieved using the proposed method. Fig. 3 indicates the spatial extents of clusters that have the largest and smallest likelihood ratio of MLCs in these calculations detected using the proposed method. Although the periods of the detected clusters almost coincide, their spatial extents do not. The log-likelihood ratio values of MLCs increase by approximately 20%, from 317 to 382. It is confirmed that the spatial aggregation group formation in the proposed method affects the cluster detection results.
Next, we compare the results of the proposed method with those of SaTScan. Table 3 shows the attributes of the detected clusters. The circular and elliptic spatial windows output the same results in this case. The spatial extent of MLC using SaTScan is shown in Fig. 4. It is clear that the likelihood ratio of MLC using the proposed method is higher than that using SaTScan, even when the lowest likelihood ratio results are used for the comparison; the flexibility in the spatial cluster shapes allows the detection

of clusters with a high likelihood ratio value. The time aggregation length, which is the temporal resolution of a cluster, is set at 7 days in the SaTScan analysis; the periods of the clusters detected by the two methods can be judged as the same results.

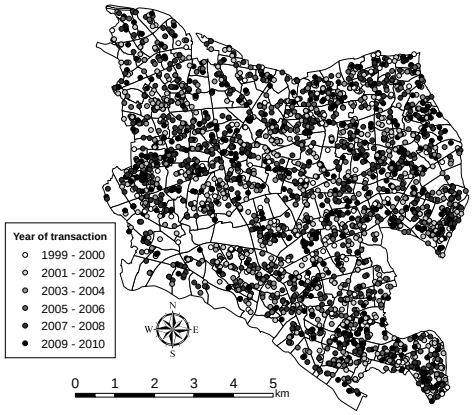


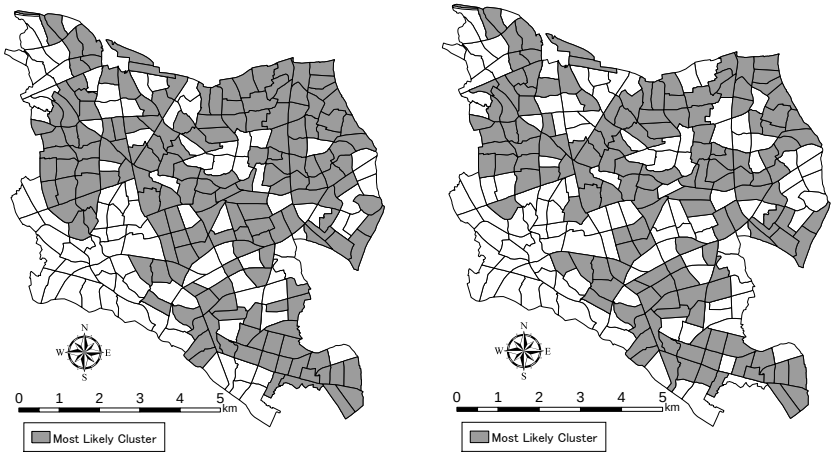
Fig. 2. Input data (Case 1

Table 2. Cluster detection result using the proposed method (Case 1)

Calculation	Log likeli-hood ratio	Area (km^2)	Period		Number of transactions
			Start	End	
1st	362.1	36.1	2005/7/19	2009/7/30	1,246
2nd	338.1	31.5	2005/7/19	2009/7/30	1,125
3rd	332.5	41.7	2005/7/19	2009/7/30	1,333
4th	351.2	39.5	2005/7/19	2009/7/30	1,307
5th	354.4	35.4	2005/7/19	2009/7/30	1,225
6th	**382.2**	35.2	2005/7/19	2009/7/30	1,245
7th	**317.2**	36.2	2005/7/19	2009/7/30	1,205
8th	351.2	35.4	2005/7/19	2009/7/30	1,222
9th	345.0	36.4	2005/7/19	2009/7/30	1,237
10th	361.7	36.6	2005/7/19	2009/7/30	1,256
11th	361.1	29.5	2005/7/19	2009/7/30	1,102
12th	364.1	37.4	2005/7/19	2009/7/30	1,276
13th	349.5	34.9	2005/7/19	2009/6/30	1,193
14th	371.5	33.7	2005/7/19	2009/7/30	1,203
15th	345.1	34.4	2005/7/19	2009/7/30	1,195
16th	354.7	35.9	2005/7/19	2009/7/30	1,236
17th	368.5	39.1	2005/7/19	2009/7/30	1,315
18th	342.4	42.4	2005/7/19	2009/7/30	1,357
19th	368.0	33.0	2005/7/19	2009/7/30	1,185
20th	377.1	32.2	2005/7/19	2009/7/30	1,175

Table 3. Comparison of the proposed method with SaTScan (Case 1)

Attributes of MLC	Proposed method	SaTScan	
		Circular	Elliptic
Logarithm of likelihood ratio	382.2	210.4	210.4
Number of transactions	1,245	883	883
Area (km^2)	35.2	29.0	29.0
Period	2005/7/19	2005/7/16	2005/7/16
	2009/7/30	2009/7/31	2009/7/31
Calculation time (s)	13	167	8,186



(a) Largest likelihood ratio 6^{th} calculation

(b) Smallest likelihood ratio 7^{th} calculation

Fig. 3. Most likely clusters obtained using the proposed method (Case 1)

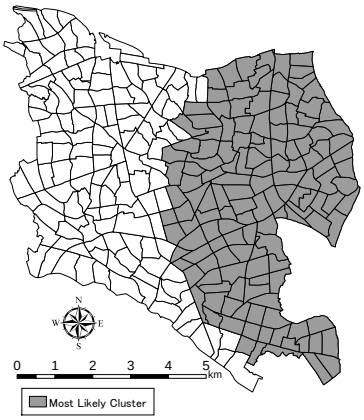


Fig. 4. Most likely cluster obtained using SaTScan (Case 1)

4.2. Case 2 – Josai region

The cluster detection results of Case 2, the analysis of data in the Josai region, are described in this section. The input data, spatial units and position, and time of transactions are shown in Fig. 5.

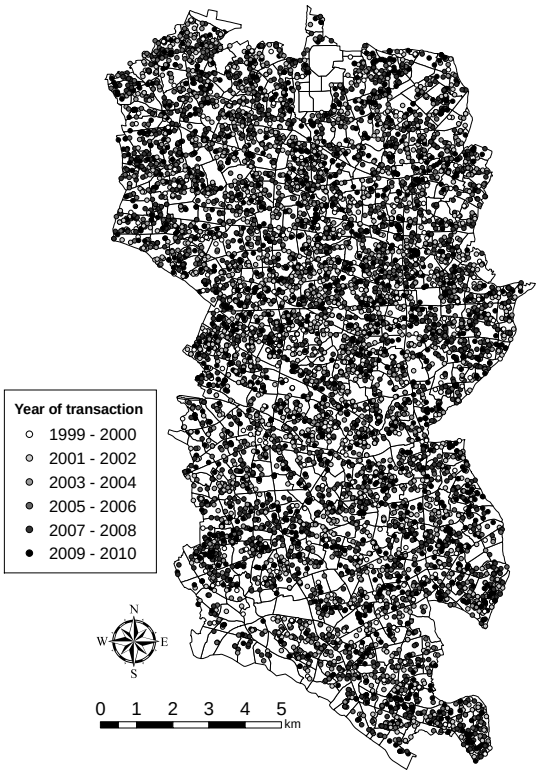


Fig. 5. Input data (Case 2)

Table 4 shows the attributes of the cluster detection results of twenty executions of the proposed method, and Table 5 shows the attributes of the clusters detected by circular and elliptic spatial windows using SaTScan. The spatial extents of MLCs obtained using the proposed method are shown in Fig. 6, and those of MLCs obtained by circular and elliptic spatial windows using SaTScan are shown in Fig. 7. Similar to the results of Case 1, the results of cluster detection using the proposed method are different according to calculations; the logarithm of the likelihood ratio values of MLCs increases by more than ten percent, from 717 to 797. Howev-

er, these values obtained by the proposed method exceed those obtained by the circular and elliptic spatial windows of SaTScan.
It is noteworthy that the spatial extents of MLCs obtained using SaTScan in Cases 1 and 2 differ, although the study area of Case 1 is included in Case 2. The MLC of Case 1 is not detected as a cluster region; the strict constraint on the spatial shape of clusters prevents its being selected as a cluster. On the other hand, the proposed method detects the area of MLC in Case 1 in the analysis of Case 2.

Table 4. Cluster detection result using the proposed method (Case 2)

Calculation	Log likeli-hood ratio	Area (km^2)	Period		Number of transactions
			Start	End	
1st	790.8	108.5	2005/7/19	2008/12/25	3,487
2nd	**717.4**	106.2	2005/7/19	2008/12/25	3,357
3rd	774.3	107.6	2005/7/19	2008/12/25	3,450
4th	733.0	103.0	2005/7/19	2008/12/25	3,306
5th	749.4	110.3	2005/7/19	2008/12/25	3,479
6th	771.7	109.5	2005/7/19	2008/12/25	3,487
7th	760.2	108.0	2005/7/19	2008/12/25	3,442
8th	746.2	103.5	2005/7/19	2009/3/31	3,496
9th	794.8	108.2	2005/7/19	2008/12/25	3,484
10th	728.8	105.2	2005/7/19	2008/12/25	3,348
11th	784.9	107.9	2005/7/19	2008/12/25	3,468
12th	791.0	105.8	2005/7/19	2009/3/31	3,597
13th	774.2	98.7	2005/7/19	2008/12/25	3,259
14th	743.5	108.4	2005/7/19	2008/12/25	3,433
15th	750.3	102.0	2005/7/19	2008/12/25	3,305
16th	780.6	103.2	2005/7/19	2008/12/25	3,363
17th	779.8	96.5	2005/7/19	2008/8/28	3,018
18th	790.3	104.6	2005/7/19	2008/12/25	3,403
19th	778.6	98.4	2005/7/19	2008/12/25	3,258
20th	**796.5**	105.2	2005/7/19	2008/12/25	3,424

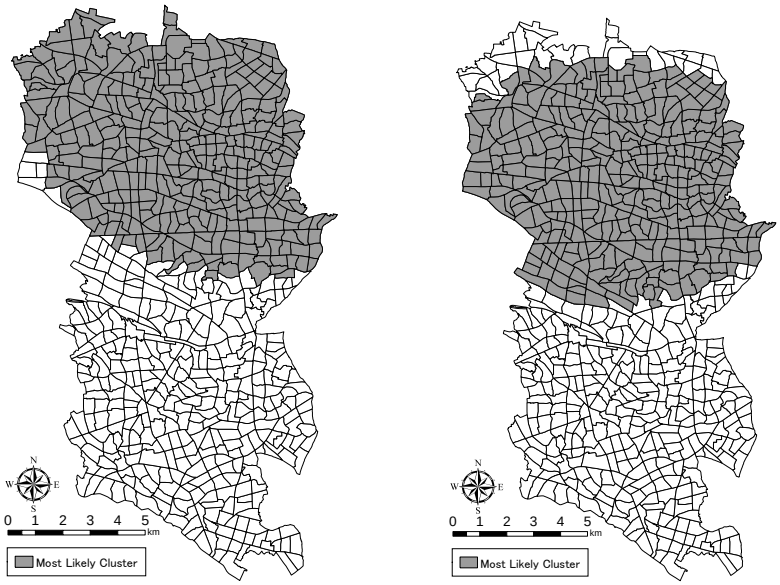
Table 5. Comparison of the proposed method with SaTScan (Case 2)

Attributes of MLC	Proposed method	SaTScan	
		Circular	Elliptic
Logarithm of likelihood ratio	796.5	422.6	439.8
Number of transactions	3,424	2,235	2,253
Area (km^2)	105.2	78.0	77.8
Period	2005/7/19 2008/12/25	2005/7/23 2008/12/26	2005/7/23 2008/12/26
Calculation time (s)	221	1,075	52,920

(a) Largest likelihood ratio 20^{th} calculation

(b) Smallest likelihood ratio 2^{nd} calculation

Fig. 6. Most likely clusters obtained using the proposed method (Case 2)



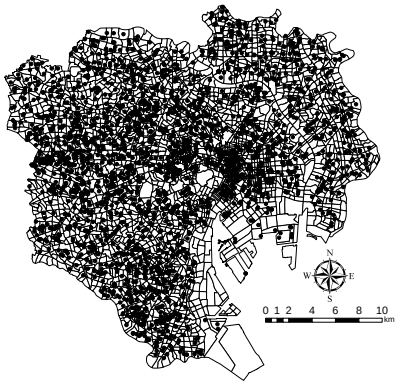
(a) Circular spatial windows

(b) Elliptic spatial windows

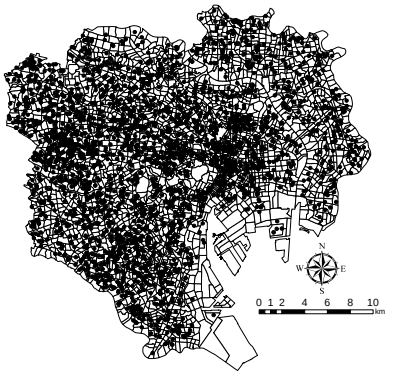
Fig. 7. Most likely clusters obtained using SaTScan (Case 2)

4.3. Case 3 – the twenty-three special wards of Tokyo

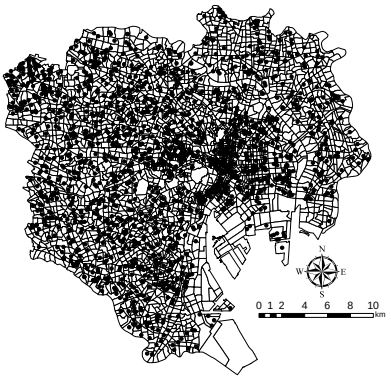
The cluster detection results of Case 3, the analysis of data in the 23 special wards of Tokyo, are described in this section. The input data, spatial units and position, and time of transactions are shown in Fig. 8.
Table 6 shows the attributes of the cluster detection results obtained by 20 executions of the proposed method, and Table 7 shows the attributes of clusters detected by the circular and elliptic spatial windows of SaTScan. The spatial extents of MLCs obtained using the proposed method are shown in Fig. 9, and those of MLCs obtained using the circular and elliptic spatial windows of SaTScan are shown in Fig. 10.
Similar to Cases 1 and 2, the proposed method outputs the clusters with large likelihood ratio values, although their likelihood ratios are in a wide range. However, the cluster periods detected by the proposed method do not coincide with those of the SaTScan in Case 3, although they almost coincide in Cases 1 and 2. Since the proposed method has a large degree of freedom for the spatial cluster shape, it allows the spatial disparity of density to be take into account more than the temporal disparity. As a result, the proposed method searches the spatial area with small land parcels, where the number of real estate transactions per unit area is relatively large for every time period; it outputs a temporal cluster period that almost covers the whole study period. On the other hand, in the methods using circular or elliptic spatial windows, the spatial shape is more strongly restricted; they output the temporal periods of clusters under the restriction on the spatial cluster shapes. These results show the necessity to control the population, especially when a method with a higher degree of freedom for spatial cluster shape is used for the analysis.
Through our analysis of the three cases, it is confirmed that the proposed method is applicable to real data. It is able to detect spatio-temporal clusters of point events with larger likelihood ratio values in a short calculation time as compared with SaTScan methods. However, since the random formation process of spatial aggregation groups affects the cluster detection results, the necessity to consider the aggregation process further is indicated.



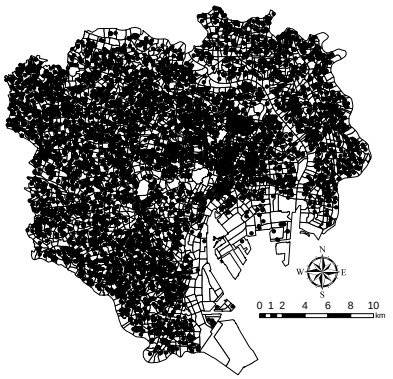
(a) Transactions in 1999 and 2000



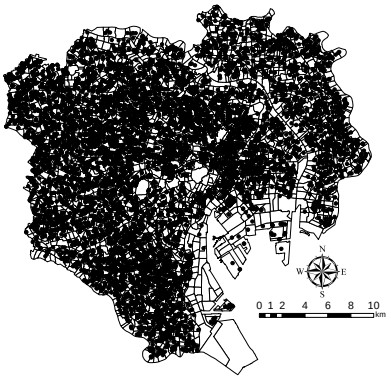
(b) Transactions in 2001 and 2002



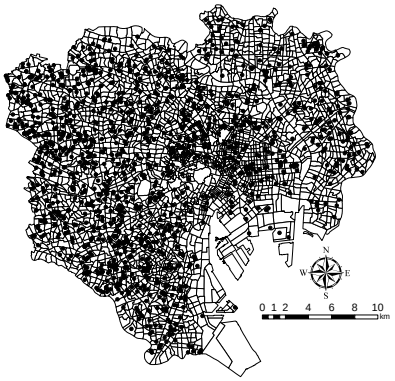
(c) Transactions in 2003 and 2004



(d) Transactions in 2005 and 2006



(e) Transactions in 2007 and 2008



(f) Transactions in 2009 and 2010

Fig. 8. Input data (Case 3)

Table 6. Cluster detection results using the proposed method (Case 3)

Calculation	Log likeli-hood ratio	Area (km^2)	Period		Number of transactions
			Start	End	
1st	4117.5	334.5	1999/4/14	2009/7/30	20538
2nd	**3805.5**	307.8	1999/4/18	2009/6/30	19189
3rd	4207.2	301.3	1999/4/18	2009/7/30	19383
4th	**4373.9**	316.3	1999/3/23	2009/7/30	20159
5th	3977.0	330.0	1999/4/18	2009/7/30	20243
6th	4065.9	323.0	1999/4/18	2009/7/30	20066
7th	4011.6	328.4	1999/4/18	2009/7/30	20216
8th	4038.2	325.1	1999/4/14	2009/7/30	20136
9th	3984.2	300.6	2005/7/18	2009/3/31	10403
10th	4173.2	326.2	1999/4/18	2009/7/30	20269
11th	4204.7	316.1	1999/4/14	2009/7/30	19948
12th	3858.5	320.6	1999/4/18	2009/7/30	19804
13th	4241.5	345.3	1999/3/28	2009/7/30	21063
14th	4174.8	337.0	1999/3/28	2009/6/30	20623
15th	4237.1	328.7	1999/3/23	2009/7/30	20497
16th	4294.1	321.5	1999/4/18	2009/7/30	20199
17th	4238.4	321.1	1999/3/28	2009/7/30	20206
18th	4361.9	322.9	1999/4/18	2009/7/30	20304
19th	4351.6	317.3	1999/4/18	2009/7/30	20096
20th	4042.4	314.0	1999/4/14	2009/7/30	19734

Table 7. Comparison of the proposed method with SaTScan (Case 3)

Attributes of MLC	Proposed method	SaTScan	
		Circular	Elliptic
Logarithm of likelihood ratio	4373.9	1733.2	1871.3
Number of transactions	20,159	7,923	8,086
Area (km^2)	316.3	310.5	310.6
Period	1999/3/23 2009/7/30	2005/7/22 2008/12/25	2005/7/22 2008/12/25
Calculation time (s)	6,316	20,525	920,654

(a) Largest likelihood ratio value 4th calculation

(b) Smallest likelihood ratio value 2nd calculation

Fig. 9. Most likely clusters obtained using the proposed method (Case 3)

(a) Circular spatial windows

(b) Elliptic spatial windows

Fig. 10. Most likely clusters obtained using SaTScan (Case 3)

5. Concluding remarks

In this paper, a new cluster detection method is proposed that is able to output spatio-temporal clusters with flexible spatial shapes. The applicability of the proposed method was verified through the analysis of three cases using real estate transaction data in Tokyo.

The comparison with the methods that output cylindrical and elliptic cylindrical spatio-temporal clusters demonstrated that the proposed method outputs clusters with larger likelihood ratio values because of the higher flexibility of the spatial cluster shapes. The calculation time of the proposed method became longer as the number of point events and spatial units for aggregation increased; however, its level was acceptable even for the larger dataset, which covered the whole area of the special wards of Tokyo.

However, in the proposed method several unsolved problems remain. The first is the dependency of detected cluster shapes on the order of the formation of spatial aggregation groups. The proposed method randomly chooses the aggregation groups in the algorithm; the obtained shapes of detected clusters and their likelihood ratio are different in every execution, as shown in the application. To find the "truly" most likely cluster and its maximum likelihood ratio is difficult. However, the difference in the cluster shapes obtained in each execution is relatively small, as shown in Figs. 3, 6, and 9.

The second problem is that the proposed method tends to create large clusters because of the hypothesis used in the selection of MLC. The MLC is selected according to the comparison of likelihood ratio shown in Equation (1), the assumption behind which is that there is only one cluster in the study area. Even if there are multiple clusters whose densities are different, the evaluation based on the likelihood ratio of Equation (1) may connect these clusters as one big cluster. A solution to this problem is proposed by Mori and Smith (2009), who introduced a Bayesian Information Criterion approach for model selection in cluster candidate evaluation and compared the models that had different numbers of clusters. The integration of a model selection approach in the proposed method is one of the future tasks toward enhancing the cluster detection ability of the proposed method.

The third problem is the difficulty with the interpretation of resultant cluster shapes. As the detected clusters almost cover the whole study area, they do not manifest the spatial pattern of the point events. This might be caused by the characteristics of the data which is influenced more strongly by time than by space; however, it indicates the necessity to restrict window shapes. Yiannakoulias et al. (2007) proposed the method to restrict window shapes based on a measure of connectivity. Since other types of spatial properties have been discussed (e.g. Shirabe (2005)), it is possible to introduce the spatial shape restriction of windows in the cluster detection analysis. At the same time, it is also indispensable to discuss what the adequate spatial cluster shape is.

The extension of the proposed method to network analysis is not difficult; it can be achieved only by replacing spatial units' adjacency with the con-

nectivity of links. As the detection of spatial and spatio-temporal clusters on a network is the present topic, as discussed in Steenberghen (2010), Shiode (2011), and Shiode and Shiode (2012), the extension of the proposed method to network analysis tools is one of the future tasks.

Acknowledgment

This work was supported by JSPS KAKENHI grant numbers 21241039 and 24241053. The real estate transaction-price information is provided by the Tokyo Association of Real Estate Appraisers.

References

Duczmal, L., and Assunção, R., 2004. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. Computational Statistic and Date Analysis, 45: 269–286.

Kulldorff, M., 1997. A spatial scan statistic. Communication Statistic Theory and Method, 26(6): 1481–1496.

Kulldorff, M., 2001. Prospective time periodic geographical disease surveillance using a scan statistic. Journal of the Royal Statistical Society. Series A (Statistics in Society), 164(1): 61–72.

Kulldorff, M., 2010. SaTScan™ User's guide version 9.0. http://www.satscan.org/ accessed: 3rd Dec. 2012.

Kulldorff, M., Athas, W. F., Feuer, E. J., Miller, B. A., and Key, C. R., 1998. Evaluating cluster alarms: a space-time scan statistics and brain cancer in Los Alamos, New Mexico. American Journal of Public Health, 88(9): 1377–1380.

Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R., and Mostashari, F., 2005. A space-time permutation scan statistic for disease outbreak detection. PLoS Medicine, 2(3): 216–224.

Kulldorff, M., Huang, L., Pickle, L., and Duczmal, L., 2006. An elliptic spatial scan statistic. Statistics in Medicine, 25: 3929–3943.

Kulldorff, M. and Nagarwalla, N., 1995. Spatial disease clusters - detection and influence, Statistics in Medicine, 14: 799–810.

Ministry of Land, Infrastructure, Transport, and Tourism, 2012. Land General Information System. http://www.land.mlit.go.jp/webland/ accessed 3rd Dec. 2012.

Mori, T. and Smith, T. E., 2009. A probabilistic modeling approach to the detection of industrial agglomerations. Discussion paper No.682, Institute of Economic Research, Kyoto University.

Shiode, S., 2011. Street-level spatial scan statistic and STAC for analysing street crime concentrations. Transactions in GIS, 15(3): 365–383.

Shiode, S. and Shiode, N., 2012. Network-based space-time search window technique for hotspot detection of street-level crime incidents. International Journal of Geographical Information Science. (In print, uploaded online on 07 Nov 2012.)

Shirabe, T., 2005. Classification of spatial properties for spatial allocation modeling. GeoInfomatica, 9:3, 269-287.

Steenberghen, T., Aerts, K., and Thomas, I., 2010. Spatial clustering of events on a network. Journal of Transport Geography, 18: 411–418.

Tango, T. and Takahashi, K., 2005. A flexibly shaped scan statistic for detecting clusters. International Journal of Health Geographics, 4: 11.

Takahashi, K., Kulldorff, M., Tango, T., and Yih, K., 2008. A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. International Journal of Health Geographics, 7: 14.

Yao, Z., Tang, J., and Zhan, F.B., 2011. Detection of arbitrarily-shaped clusters using a neighbor-expanding approach: A case study on murine typhus in South Texas. International Journal of Health Geographics. 10: 23.

Yiannakoulias, N., Rosychuk, R.J., and Hodgson, J., 2007. Adaptations for finding irregularly shaped disease clusters. International Journal of Health Geographics. 6: 28.