GeoJournal
https://doi.org/10.1007/s10708-023-10834-5

Detection of irregular-shaped clusters on a network by controlling the shape compactness with a penalty function

Ryo Inoue · Shino Shiode · Narushige Shiode

Accepted: 23 January 2023

Abstract Recent development of cluster detection methods focuses on the improvement of efficiency or accuracy, with the latter yielding a wide range of variants in the shape of the search window, from a simple circle and elliptic shape to more irregular shapes. Detection of irregular-shaped clusters has seen various new approaches as it is considered to capture the shape and extent of clusters more accurately. One of these newly developed approaches achieves the irregularity of the clusters by placing a penalty on the shape complexity of a candidate cluster. This study extends this approach and applies it to a network-space to detect irregular-shaped clusters along a street network segments in a small urban area. The study uses a genetic algorithm to search candidate clusters and identify the most likely cluster using the framework of spatial scan-statistics. Application of the method to a small synthetic data and a real data set revealed that providing options of different cluster patterns with different compactness parameters helps find more accurate as well as geometrically and contextually more meaningful clusters, as opposed to those detected without a shape controlling parameter.

Keywords Cluster detection · Genetic algorithm · Network-based analysis · Scan statistic · Shape compactness

Introduction

In an urban environment, configuration of a street network affects and even determines the characteristics of each street block and segment, often in conjunction with zoning, land use plans, building control and other types of regulations. Attributes attached to these neighborhoods and streets often determine the type and the spatial arrangement of point features, events and activities observed in these areas. For instance, areas with narrower and intricate streets tend to result in residential usage, while areas with greater accessibility and a major throughway tend to be used for commercial facilities. Key urban and public facilities such as schools, churches and hospitals tend to be located in accessible places near a major junction. Consequently, each neighborhood and street come to form their own identity and characteristics, which allows these areas to provide a geometrical and

R. Inoue
Department of Human-Social Information Sciences, Graduate School of Information Sciences, Tohoku University, Sendai 980-8579, Japan
e-mail: rinoue@tohoku.ac.jp

S. Shiode
Department of Geography, Birkbeck, University of London, London WC1E 7HX, UK
e-mail: s.shiode@bbk.ac.uk

N. Shiode (✉)
Department of Geography, Geology and the Environment, Kingston University, Kingston Upon Thames KT1 2EE, UK
e-mail: n.shiode@kingston.ac.uk

Published online: 15 February 2023

contextual meaning to a cluster of features and events detected at that scale and extent. In a small-scale urban setting, these clusters are formed either on a single street or on a connected collection of street segments, which may take a simple or an irregular shape. Such variation in the shape of clusters is difficult to capture with existing cluster detection methods, especially when those clusters are confined by the street network.

This study proposes a cluster detection method that can identify irregular-shaped clusters of features that are confined by and formed along an urban street network. It extends the notion of spatial scan statistics that is widely used in spatial epidemiology and applies it to network space, whilst also adding a shape-controlling procedure as a parameter during the scan statistics calculation so that more than one cluster sets with different shape irregularity are derived. The aim of this approach is to provide options that may increase the chance of detecting clusters that are closer to the shape and extent of true clusters.

Literature review

The purpose of detecting clusters among point features or cases of events is to find the shape and the intensity of their concentration and, thereby, to investigate its aetiology to characterise their determinants. The shape of a cluster is usually unknown *a priori*, there is no right answer as to what shape, size and extent we should expect a cluster to take. Existing studies on cluster detection tend to identify the extent of the detected clusters by (1) whether and how the data (point features) are aggregated, and/or (2) the analytical procedure used for identifying clusters.

Detecting clusters in aggregate and disaggregate data

Point features are often aggregated to areal units defined by administrative boundaries or regular square grids for variety of reasons, e.g. to standardise by the background population in each area to derive the relative intensity of the features; or to maintain privacy by concealing the exact locations of features. Aggregating point features by administrative boundaries and other predetermined shapes means that the detected cluster also becomes connected areas. There are many cluster detection methods for finding clusters of aggregated point features.

One such method developed in the domains of quantitative geography and spatial analysis focuses on the notion of *spatial autocorrelation*. Methods that belong to this group use a spatial correlation coefficient such as local Moran's *I*, local Geary's *C* and local Getis-Ord's G_i^* and G_i (Anselin, 1995; Getis & Ord, 1992; Ord & Getis, 1995) to extract local clusters. These statistics help us find a group of adjacent areas with similar attribute values. Aldstadt and Getis (2006) also developed A Multidimensional Optimum Ecotope-Based Algorithm (AMOEBA), which extends the concept of spatial autocorrelation and finds clusters as flexibly combined neighbouring areas.

Another group of methods use the concept of a search window. A search window usually takes a circular shape and is used for sweeping exhaustively across the study area to find an area with a high concentration of events or features. This approach is also used for detecting a sequence of connected areas that determines the shape and the extent of a cluster. A number of relevant statistical techniques have been developed to facilitate these methods (e.g. Turnbull et al., 1990, Besag & Newell, 1991, Diggle & Chetwynd, 1991, Tango 1995, Kulldorff & Nagarwalla, 1995, Rushton & Lolonis, 1996). In particular, Spatial Scan Statistics (Kulldorff, 1997; Kulldorff & Nagarwalla, 1995) has become arguably the most prominent search-window method, partly because it addressed most of the limitations of the previous search window type methods, and also because the method has been offered as software called SaTScan that found users in epidemiology and many other fields (Kulldorff, 2022). In principle, most of these methods can be also used for analysing the distribution of *disaggregate* point features. In these contexts, the cluster detection methods are not explicitly used for delineating the boundary of a cluster area; rather, they are intended for identifying a point set that constitutes a cluster. However, in the process of identifying a significant concentration of points as a cluster, the shape of that cluster will be explicitly defined. Furthermore, as explained below, the shape of the detected cluster will be also affected by the shape and the procedure of the search window used for identifying that cluster.

The shape of a search window

In pursuit of a method for that captures the shape and extent of clusters accurately, several methods have been developed mainly in the computer science field. The main focus of these studies is to configure a search window with the right shape and size so that it can efficiently capture the most highly concentrated set of point features. Search-window-type methods have most rigorously explored this point, especially after Kulldorff's scan statistics became widely known. Many of the search-window-type methods, including the original scan statistic, use a *circular* search window which means that the shape of the detected clusters will be also bound by that circular shape. Although it is simple and computationally efficient, the circular spatial scan statistic works well with compact-shaped clusters only, and it may struggle to correctly identify non-circular clusters (Kim & Jung, 2017). Tango and Takahashi (2005), Tango and Takahashi (2012), Tango (2021) suggest that the original spatial scan statistic using circular windows tends to extract an area that is larger than the true cluster by absorbing neighbouring areas with non-elevated risk. This phenomenon may occur more easily if the true cluster is non-circular in shape. Patil (2004) refers to the case of a cholera outbreak along a winding river floodplain and explains how poorly the circular window fits—the small circles miss out on much of the outbreak while the large circles include many unwanted areas. Kim and Jung (2017) also note how a circular search window is ill suited for detecting clusters of events along roads. To overcome this shortfall, various extensions were added to the original circular scan statistic to find clusters with a non-circular (and non-compact) shape. These include Kulldorff et al. (2003), Tango and Takahashi (2005), Neill et al. (2005), Asunção *et al.* (2006), Kulldorff et al. (2006), Moura et al. (2007), Takahashi et al. (2008), Costa et al. (2012), Neill (2012), Neill et al., 2013 with a comprehensive review of these methods given by Duczmal et al. (2009), Duczmal and Cançado (2017).

Firstly, as a natural extension to a circular-shaped search window, Kulldorff et al. (2003) proposed an *elliptic* shaped search window for capturing clusters that are extend in a certain direction. Although their power comparison indicates that the overall performance of elliptic scan statistic is only slightly better than that of the circular variant, they suggest that the advantage of the elliptic scan statistic lies in its ability to give a better estimate of the true cluster area, especially if the true cluster area was elongated. Interestingly, they also point out that the elliptic, like the circular, is a parametric shape which comes with some constraint and, to detect very irregular shaped clusters, a non-parametric spatial scan statistic would yield a better outcome.

Unlike the circular and the elliptic scan statistic, the non-parametric type of scan statistics is designed to find irregularly shaped clusters. For instance, Patil and Taillie (2004) proposed new method called an *upper-level set scan statistic*. It adopts a data-driven search algorithm for detecting a cluster of connected areas by utilising a tree structure for identifying the adjacency among the areas nearby. They used the notion of upper-level set to reduce the size of windows to be scanned. Similarly, Duczmal and Assunção (2004) proposed a new technique of *adaptive Simulated Annealing* (SA). It is also based on the likelihood ratio test formulated in the same way as the circular spatial scan statistic was. However, in this method, the set of areas comprising the irregularly shaped cluster could result in a large collection of areas, and it would not be feasible to derive the likelihood for all areas. It is for this reason that an SA method has been adopted, where they test only the most promising windows to derive the local maxima with a certain likelihood function over a subset of areas comprising the irregularly shaped cluster.

Both Patil and Taillie (2004) and Duczmal and Assunção (2004)'s approaches were designed to avoid computationally infeasible searches, but they also had weaknesses in that Patil and Taillie (2004) failed to select the upper-level set (Tango and Takahashi 2005). Also, while their method offers faster alternatives, there is a risk of overlooking many interesting clusters in this procedure, due to the small cardinality of the upper-level set tree (Duczmal, 2009). Duczmal and Assunção's method (Duczmal & Assunção, 2004), on the other hand, detects a cluster of irregular shape but is the results tend to be much larger than the extent of a true cluster (Tango and Takahashi, 2005). The overshooting of the clusters mainly arises from the application of the algorithm where it tends to find elongated clusters that link the highest likelihood ratio cells only in the map.

Tango and Takahashi (2005) proposed another way to find clusters of flexible shapes, facilitated by their

software called FleXScan (Takahashi et al., 2010). They state that their method is suitable for situations where relatively small clusters are expected, as they have an upper limit of 30 areas that can fit in a single cluster. Later, Tango (2008), Tango and Takahashi (2012), Tango (2021) eliminated this limitation of 30 nearest neighbours and achieved faster computational time than the original flexible spatial scan statistic took by proposing a flexible spatial scan statistic with a *restricted likelihood ratio*.

Bounding the cluster shape with penalties

The algorithms presented above managed to identify irregular-shaped clusters. However, without any control on their search, the detected irregular shaped clusters would likely exceed the extent of true clusters. To alleviate this, Tango and Takahashi (2005) constrained the Maximum Search Window Size (MSWS) to a small value to keep the detected cluster from becoming excessively large. Another way would be to introduce some form of a *penalty function* to constrain the cluster from taking too irregular a shape.

The idea of applying a penalty function on spatial clusters for controlling the irregularity of its shape was first used for ellipses (Kulldorff et al., 2006), but many studies presented alternatives since then. For instance, Duczmal et al. (2006) proposed a variant of simulated annealing scan statistic that considers the geometric *non-compactness* of the cluster shape. The penalty on the shape irregularity is introduced in the form of a modified maximum likelihood function, which is a general form of that used in the case of ellipses (Kulldorff et al., 2006). Compactness correction is a parameter that is incorporated into the likelihood ratio function in the scan statistics and serves as a penalty to control the extent of irregularity of the cluster shape. Similarly, Yiannakoulias et al. (2007) proposed *adjacency-constrained spatial scans*, which applies double penalties on the standard scan statistics; the first penalty pertains to controlling highly irregular shapes with non-connective relationship, while the other constrains searches to prevent small clusters from getting merged into a single large cluster. More recently, Cançado et al. (2010), Tango and Takahashi (2012), Somanchi et al. (2015), and Tango (2021) respectively introduced notions linked to a penalty for the purpose of controlling the irregularity of a cluster shape. For instance, Somanchi et al. (2015) proposed Star Scan, where changes in the radius of a cluster is penalised to ensure smoothness of the circumference of a cluster.

Besides the use of a penalty to control the shape of a cluster, efforts have been placed on developing new algorithms to rapidly and efficiently find the most likely cluster. One of the increasingly widely used approaches is Genetic algorithm (GA). GA is a heuristic method that adopts the notion of evolution and natural selection to simulate certain scenarios for computing the optimal solutions. The method adopts the equivalent of biological operators such as mutation, crossover and selection to run the iterative process of natural selection to derive the solution, including the optimisation of decision trees for better performances. It is known to date back to Turing's learning machine (1950) that took on the principles of evolution, followed by computer simulation of evolution proposed by Barricelli (1957), Fraser (1957) and others for controlling measurable traits using a heuristic selection process (Fogel, 2006).

Sahajpal et al. (2004) and Conley et al. (2005) are early examples that adopted GA in the context of cluster detection across point data. Duczmal et al. (2007a) also proposed the application of GA called genetic-algorithm scan for extracting clusters using a flexible search window. Their algorithm repeats the crossover and the selection processes under a set of predefined parameters until the solution converges to the most likely cluster. Graph-related operations are minimised by adopting a fast offspring generation and efficient evaluation of Kuldorff's spatial scan statistic. It is a faster and more robust alternative to their previous method, simulated annealing scan (Duczmal & Assunção, 2004); yet performance tests suggest that both methods return a comparable power of detection for clusters that are moderately irregular, but GA-based method is much better when it comes to highly irregular clusters.

In summary, a number of studies have been proposed for using search windows that are irregularly shaped and they have demonstrated that these irregular search windows provide better performance on detecting irregularly shaped clusters. Tango (2021) notes that the original scan statistics tend to detect an area much larger than the true cluster as the most likely cluster, and it does so by merging the neighbouring regions with non-elevated risks. While the

performance of a cluster detection method depends on the underlying shape and the spatial arrangement of true clusters, the flexible shaped cluster detection methods discussed so far seems to return better performance than the circular or the elliptic types. However, they also have shortfalls. For instance, the flexible-type scan statistics can detect a very irregular and complex shape, but these clusters often do not provide a geographically meaningful solution that delineates the location of the true clusters correctly. To avoid this, applying some control over the shape of a cluster by imposing a penalty seems to be effective. Duczmal et al. (2006) developed a significant improvement in shape control by using a geometric "non-compactness" as a penalty function against highly irregular-shaped clusters.

However, there is no conclusive evidence on which method delivers the closest solution to the true irregular-shaped clusters, as there is lack of research that has comprehensively compared these methods. Given the lack of consensus on which searching algorithms and which shape controlling method are the best, this study extends Duczmal et al. (2007a) that employed a penalty function to control the shape and GA for deriving the solution fast and develops a variant that can be applied to the network dimension in an urban space where certain types of events and features are confined by the street network. We will be using GA because, algorithmically, it is a well-established algorithm and is proven to provide fast calculation for flexible shaped cluster detection. Also, while some of the methods proposed in the literature do not show any programmable procedure of their method, GA is readily available for extension.

While some recent studies have developed network-based cluster detection methods, they are not designed to detect flexible-shaped clusters. For instance, Shiode and Shiode (2020) proposed a concept of a networks-shaped search window called NetScan. It detects clusters along a street network but remains focused on the detection of relatively compact clusters which may not be suited for scenarios where we need to extract larger irregular clusters that are stretched across a neighbourhood. The size of their target clusters is small, not because they set a MSWS but because it is a disaggregate method and the true cluster size that they aimed to detect in their application are small in nature. This is understandably so, as Shiode and Shiode (2020) proposed their method in response to the recent trend towards micro-scale crime opportunities, i.e. identifying crime hotspots at specific locations, as many crime events concentrate at very specific small places. This shows that the suitability of the methods may vary and depends on the actual data and the context. There are also some other studies that involve street networks in the cluster detection. For instance, Duczmal and Buckeridge (2006) extended the spatial scan statistic to account for the mobility of individuals between their home address and workplace. Also, Duczmal et al. (2007b) presented a concept of creating a graph structure where traffic between cities is represented by the population, similar to the framework of a spatial interaction model. However, both studies presented their concepts only and did not analyse data from the physical street network. For this reason, this study proposes a network-segment based approach that can detect irregular-shaped clusters along a street network by (1) utilising a penalty controlling procedure as means to increase a chance to identify cluster with better performance; and (2) utilising GA to find a solution within a reasonable amount of time.

Methodology

Scan statistics (and network-based scan statistics)

A standard spatial scan statistic (Kulldorff, 1997; Kulldorff & Nagarwalla, 1995) creates a search window around the centroid of each spatial region and the radius of this window changes continuously to take any value between zero and a predefined upper limit MSWS. Using the likelihood ratio test, the scan statistic detects areas where the underlying event occurrence rates are significantly higher within the window than it is outside. In other words, for each scanning window, a likelihood ratio test statistic is calculated for comparing the event rate within and outside the window. The window with the maximum likelihood ratio will be detected as the most likely cluster; i.e. a cluster that is most likely to be generated under the alternative hypothesis of clustering. The test statistic is evaluated through Monte Carlo simulations. Randomisation testing is used for computing the p-value of each detected region, correctly adjusting for multiple hypothesis testing, and identifying potential clusters and assessing whether they are significant.

Let Z denote the extent of the candidate cluster comprising a combination of adjacent (connected) regions, n_Z denote the number of point features that exist within Z, and Z^C denote the area outside Z (i.e. the remainder of the study area). Next, we define $\lambda(Z)$ as the expected number of points within Z, $\lambda\left(Z^C\right)$ as the expected number of points in Z^C and assume that their point pattern follows that of the Poisson distribution. The null hypothesis H_0 is that $n_Z = \lambda(Z)$ and the alternative hypothesis H_1 is $n_Z > \lambda(Z)$. The Likelihood Ratio (LR) is calculated for the network cluster in much the same way as *standard scan statistic*:

$$\mathrm{LR}(Z) = \begin{cases} \left(\frac{n_Z}{\lambda(Z)}\right)^{n_Z}\left(\frac{n_{Z^C}}{\lambda\left(Z^C\right)}\right)^{n_{Z^C}}, & \text{if} n_Z > \lambda(Z), \text{ and} \\ 1 & \text{, otherwise} \end{cases} \quad (1)$$

Compactness index and compactness correction

To control the shape complexity of a detected cluster through scan statistic, Duczmal et al. (2006) introduced a *compactness index*. Let $A(Z)$ denote the area of Z, and $H(Z)$ denote the perimeter of a convex hull of Z. The compactness index $K(Z)$ is defined by the ratio of $A(Z)$ and the area of a circle with perimeter $H(Z)$; i.e.

$$K(Z) = A(Z)/\pi\left(\frac{H(Z)}{2\pi}\right)^2 \quad (2)$$

$K(Z)$ takes values between 0 and 1–it approaches 0 if Z takes a complex shape with long perimeter and reaches 1 if Z has a compact, circular shape. Duczmal et al. (2006, 2007a) employ $LR(Z)^{K(Z)^a}$ as a penalty function by modifying the original likelihood ratio $\mathrm{LR}(Z)$ derived in the scan statistic (Eq. 1), where a is a *compactness correction* parameter which is a user-specified exponent that controls the degree of *penalty* placed on the geometric shape of the candidate clusters. Like the Compactness Index, the Compactness Correction can take values between 0 and 1—*a* of a greater value increases the effect of the penalty, prioritising more compact clusters, whereas a lower a value allows more flexibility for the shape of a cluster. Specifically, no compactness constraints are imposed on the shape of a cluster when a is 0, and the penalty on the shape becomes stricter as a increases (e.g. $a=0.5$ means medium compactness correction is made, and $a=1$ means full compactness correction is in effect).

The notions of compactness index and compactness correction can be also adopted for measuring the shape complexity of network-segment clusters. This study proposes a network-based compactness index (hereafter *NT-compactness index*) to quantify the compactness of the candidate network-segment clusters. Unlike a regular, circular-shaped cluster in the Euclidean space, the shape of a network-segment cluster is constrained by the geometrical configuration of the sub-network area; i.e. it is not possible to define the single most compact form of a network that applies across the study area, as what constitutes the most compact network differs from place to place. Therefore, this study defines the most compact network as the link set in which '*all endpoint vertices are at the same shortest-path distance along the network from the central node*'. This definition retains the equidistance property of a circle in the Euclidean space in which '*all end points (points on the circumference) are at the same distance from the centre of the circle*.'

Let $C(Z)$ denote the most compact sub-network corresponding to a candidate sub-network Z. In other words, $C(Z)$ has the same central node as Z does as its generator point from which the minimum-spanning tree is extended in all direction to cover the same total length as Z. Here, the central node Z_O of network Z can be defined as *the node that has the shortest network distance to the farthest node(s) within Z*. As this study uses network segments as the smallest unit of measurement, we can revise and simplify this statement as follows: the central node Z_O of a network Z is defined as *the node that has the shortest network distance to the end node of the farthest network segment(s) in Z*. Then, $C(Z)$ is generated by extending the minimum-spanning tree from Z_O in all possible directions and selecting the connected links in the shorter-distance order from the central node until the total length exceeds that of Z. Let $N(Z)$ denote the node set in a sub-network Z and $D(n, Z_O)$ denote the network distance to the farthest node n from the central node Z_O of Z. Then, NT-compactness index $K_{\mathrm{NT}}(Z)$ can be derived as

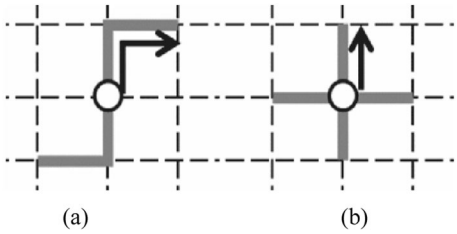

Fig. 1 An illustrative example of **a** an irregular sub-network Z, and **b** the corresponding most compact network $C(Z)$

$$K_{NT}(Z) = \frac{\max\limits_{n \in N(C(Z))} D(n, Z_0)}{\max\limits_{m \in N(Z)} D(m, Z_0)} \quad (3)$$

By replacing $K(Z)$ with $K_{NT}(Z)$, we aim to find Z that maximises the $LR(Z)^{K_{NT}(Z)^d}$ under the predetermined compactness correction parameter.

Figure 1 shows an illustrative example of a candidate sub-network Z (Fig. 1a) and its corresponding most compact sub-network $C(Z)$ (Fig. 1b). The white circle is the central node Z_0, and the grey bold line segments show the extent of Z and $C(Z)$, respectively. The black arrows show examples of the shortest-path route from Z_0 to the end note of the farthest network segment.

Figure 2 shows another example of the comparison between a complex sub-network ($K_{NT}(Z) = 0.3$) and the corresponding most compact network ($K_{NT}(Z) = 1$) defined on a real street network. While both sub-networks share the same starting point Z_0 and the same total length, the two subnetworks have contrasting appearance in that Fig. 2a shows a sub-network with an irregular and elongated sub-network, whereas Fig. 2b shows the arrangement of a compact network.

Extracting an irregular shaped network-segment cluster using genetic algorithm (GA)

Using NT-compactness correction as a means to apply penalties to regulate the shape of sub-networks, this study employs the Genetic-Algorithm (GA) Scan for detecting clusters that consist of spatially contiguous network segments. It expands on the GA Scan proposed by Duczmal et al. (2007a) for searching flexible-shaped clusters on a Euclidean plane. In the case of network-segment cluster detection using GA,

Fig. 2 Illustrative examples of sub-networks with **a** an irregular, complex network ($K_{NT}(Z) = 0.3$), and **b** the most compact network ($K_{NT}(C(Z)) = 1.0$)

each *individual* in the GA scan represents a candidate cluster, where the set of network segments that comprises a candidate cluster is recorded as its chromosome information. Here are the steps of GA for searching candidate network-segment clusters.

Step 1: Initialisation

To improve the odds of identifying clusters faster, the initial population is generated against all network segments that contain at least one of the point features. Each individual (candidate cluster) is expanded by randomly selecting adjacent segments and adding them to construct its sub-network area. In order to generate sub-networks with a higher concentration of point features, segments that contain no features are only added stochastically. This process will be repeated until the number of segments in the sub-network reaches a predetermined threshold value, or the likelihood of the respective sub-network falls below a threshold.

Step 2: Crossover

Select two individuals randomly from the initial population and examine if they share any network segment(s). If they do, implement crossover using those segment(s) to breed the second generation. Specifically, for each parent, create an ordered list of network segments in ascending order of their connectivity (those with the same degree of connectivity will be ordered randomly). Next, implement an

order-based crossover by combining the parent segments in the order of the list to create the second-generation population. If the remnants outside the crossover points constitutes m and n number of segments respectively, generate m x n sets of individuals and retain all solutions that satisfy a predefined shape complexity constraint.

Step 3: Selection (replacement)

The crossover step generates numerous solutions from the same combination of parents which, in turn, reduce the diversity of the gene pool among the individuals. For this reason, this study selects a portion of the parent generation that have low likelihood ratios and replaces them with the child generation that have higher likelihood ratios.

The process of crossover and selection (replacement) will be repeated until such time that the same individual has retained the maximum likelihood ratio for a sufficient number of iterations, which will be identified as the Most Likely Cluster (MLC). The significance of this MLC will be tested by comparing it against the MLC distribution derived from the random point distribution obtained by Monte Carlo simulation.

Sensitivity analysis with a synthetic data set

To test the effectiveness of the proposed method, the impact of the NT-compactness correction parameter on the shape of the detected clusters (i.e. $K_{NT}(Z)$ of the most likely cluster) and the performance of the proposed method were examined using a simple synthetic data. Performance of the method was assessed using *sensitivity* and *Positive Predictive Value* (PPV), where the sensitivity is defined as the proportion of network segments correctly detected among all network segments in the true cluster, and PPV is defined as the proportion of network segments correctly detected among all detected segments. These tests were conducted against two small synthetic cluster data sets (with distinct shapes and arrangements). Figure 3 shows a regular grid-type street network that consists of 220 same-length links, on which the following cluster(s) are generated: (a) a single cluster (in red) which mimics a non-compact cluster; e.g. a cluster formed along arterial streets in an urban setting, and (b) two separate small clusters (moderately compact and non-compact clusters) that are located

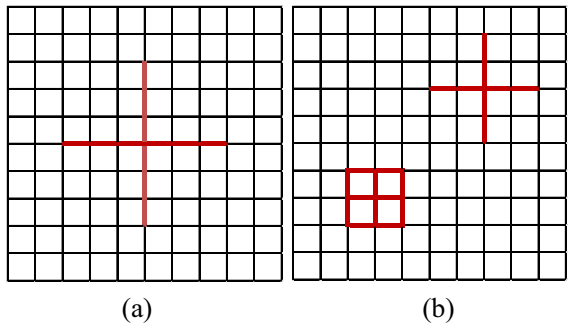

Fig. 3 Examples of synthetic data: **a** a single non-compact cluster, **b** two clusters located near each other

near each other. Points were then placed along the street network within and outside the cluster(s) using Poisson distribution with the following parameters: (1) the expected number of points inside the cluster(s) (red links) was set to 100 points per link, and (2) the expected number of points outside the clusters (black links) was set to 10 points per link.

Results from the analysis are shown in Fig. 4. Firstly, the true non-compact cluster is detected when no compactness correction was made ($a = 0$). As the compactness correction value increases, links outside the true cluster are gradually added to the detected cluster which reduces the PPV to 0.5 in the end when the compactness correction value reaches 1 (Fig. 5a and Table 1).

Secondly, the two clusters located near each other are detected as a single large cluster, if there is no penalty on the shape. Although the sensitivity is 1 (i.e. all true cluster links are included in the detected cluster) (Table 2), some excessive links (those located between these two links——8 in total) are also included, and this is reflected in the low PPV value. As the compactness correction value increases, the constraint on the shape compactness is tightened and PPV gradually decreases (i.e. the number of links that have been incorrectly detected increases gradually (Table 2). However, when the compactness correction value reaches 0.7, only one of the two true clusters gets detected (the cluster in red) as the MLC, and this continues until the compactness correction value reaches 1. This is why the PPV suddenly drops to 0.6 when the compactness correction reaches 0.7. To avoid multiple testing problems, spatial scan statistics restricts us from detecting clusters simultaneously at multiple

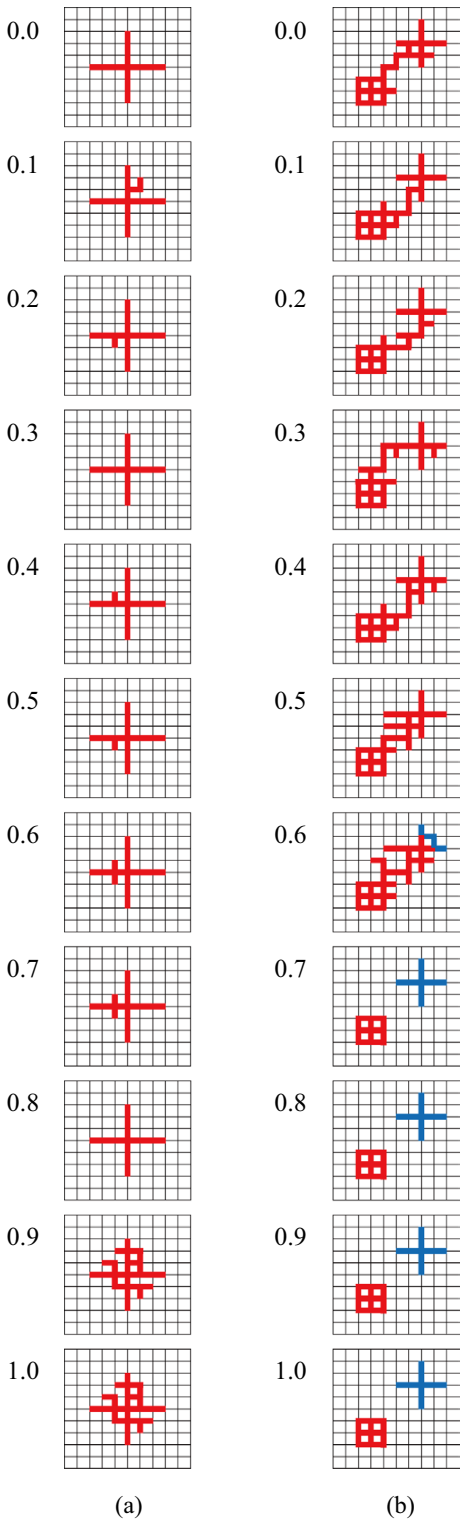

Fig. 4 Detected clusters across different compactness correction values for **a** a single non-compact shape cluster, and **b** two clusters located near each other

locations (Zhang et al., 2010). For this reason, we detect the second MLC, the third, and so on in different rounds of execution by eliminating the extent of previously detected cluster(s). The blue cluster in Fig. 4b therefore is detected as the MLC in the second round of cluster detection under the condition that the first cluster exist at the detected place. This suggests that when the penalty is low, the method allows more flexibility on the shape of a cluster and can detect irregular-shaped clusters, but the method may fail to distinguish nearby clusters separately and may detect them together as a single large cluster. The results from two sample distributions exhibit an interesting comparison. In the case of simple small clusters used here, a strong penalty hampered the detection of a non-compact cluster, but it also helped distinguish the two separate clusters located near each other.

Application

Data

To test the effectiveness of the proposed method with real-world data, this study uses distribution of taverns (casual restaurant-style pubs that tend to cluster in highly populated areas) along a street network of a regional city, the central area of Sendai City, Japan. The street area extends to roughly 7 km by 6 km (Fig. 6). The area serves as an ideal test case, as it covers a variety of street configurations and street densities, ranging from a grid-based tightly knit urban structure of downtown Sendai, to the radial configuration in the suburb, as well as the sparse and irregular street configuration on the outskirt. Taverns are casual diners serving alcohols and are prevalent across different types of urban landscape in Japan, including downtown, suburbs and outskirts. The street network was obtained through the Digital Topographic Map 25,000 dataset, covering a total of 647 km across 7953 street segments in the area. Locations of taverns were extracted from the TelePoint® Pack database (September 2010 edition) which listed the xy-coordinates of 492 taverns in the area. Each tavern is assigned to the nearest street segment and the number of taverns is recorded for each street segments. Of the 7953

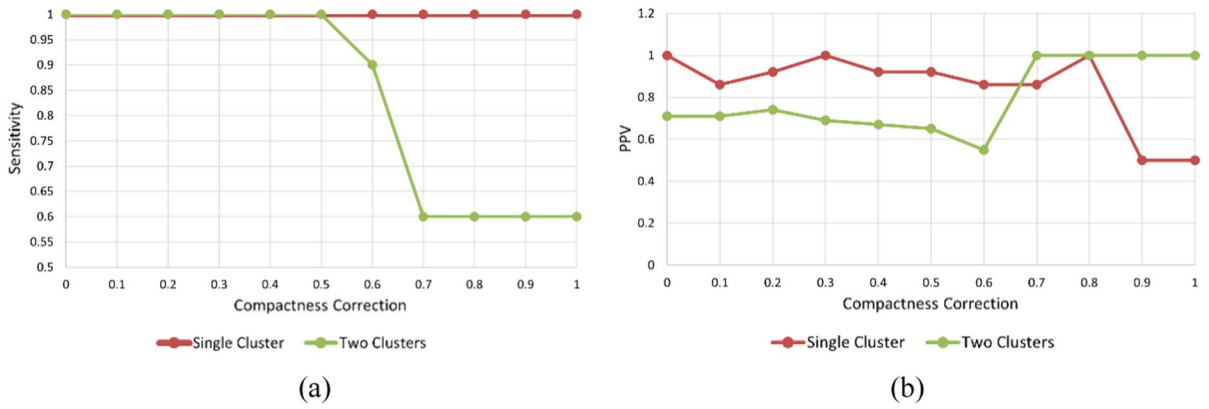

Fig. 5 Assessment of the performance in the forms of **a** sensitivity and **b** PPV

Table 1 Result statistics for the single non-compact cluster (cluster (a))

Compactness correction	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Compactness index ($K_{NT}(Z)$)	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	1	1
#Dectected links	12	14	13	12	13	13	14	14	12	24	24
#True cluster links detected	12	12	12	12	12	12	12	12	12	12	12
#Links wrongly dectected	0	2	1	0	1	1	2	2	0	12	12
Sensitivity	1	1	1	1	1	1	1	1	1	1	1
PPV	1	0.86	0.92	1	0.92	0.92	0.86	0.86	1	0.5	0.5

Table 2 Result statistics for the two clusters located near each other (clusters (b))

Compactness correction	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Compactness index ($K_{NT}(Z)$)	0.5	0.5	0.5	0.5	0.5	0.58	0.64	1	1	1	1
#Dectected links	28	28	27	29	30	31	33	12	12	12	12
#True cluster links detected	20	20	20	20	20	20	18	12	12	12	12
#Links wrongly dectected	8	8	7	9	10	11	15	0	0	0	0
Sensitivity	1	1	1	1	1	1	0.9	0.6	0.6	0.6	0.6
PPV	0.71	0.71	0.74	0.69	0.67	0.65	0.55	1	1	1	1

street segments in the area, 217 segments contained one or more taverns.

Detection of non-weighted network-segment clusters

During the initialisation stage, parameters for the GA model were set as follows: the tolerance level for connecting a segment with no tavern locations was set at 1/3, the maximum number of street segments that could be connected as a single cluster was set at 217 (i.e. the number of segments containing at least one tavern), and the lower bound of the log-likelihood ratio was set at 5. In the selection process, 10% of the parent generation were replaced by individuals from the child generation; and the iterative process of crossover and selection was set to be terminated after the same individual has remained as the MLC for 10 generations. These parameters were determined in an exploratory manner, as restraining the maximum number of connected street segments too low could lead to compact, local optimal solutions, whilst allowing too high a log-likelihood ratio may result in generating a homogeneous set of initial population that cover similar areas in such a way that each individual of the initial population would comprise sufficient

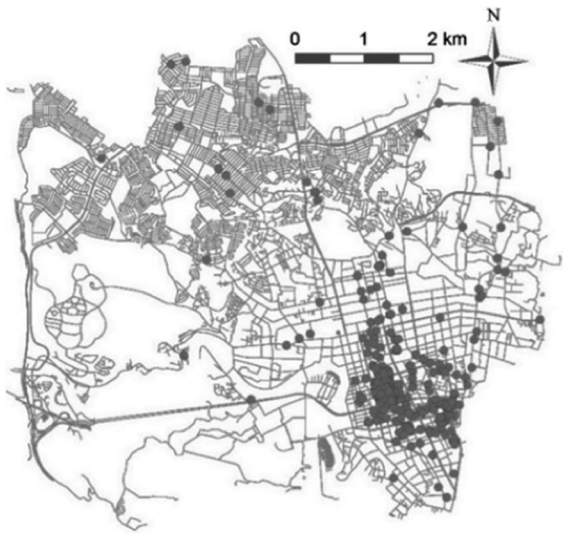

Fig. 6 Distribution of 492 taverns (black points) in the study area

number of segments but are different in the area they cover and maintain diversity.

The first step of the empirical analysis was carried out with the maximum extent of an individual set at the total length of all segments. Under the assumption that the distribution of taverns along the street network would follow a Poisson point distribution, we use GA to search for MLC by exploring areas that have higher density of taverns within the sub-network than outside. Using Monte Carlo simulation, the distribution of stores was generated on each link (under the null hypothesis that there is no difference in the point density within and outside a sub-network of interest), repeated the process of detecting clusters 1,000 times, and determined the rejection region of MLC. In the case of detecting clusters with no penalty on shape, the rejection region by one-sided test with a significance level of 1% was greater than the log-likelihood ratio of 24.5.

Next, different degrees of compactness correction were set and clusters of taverns were detected. The red lines in Fig. 7 denote the detected sub-network clusters under different compactness correction values. All results yielded high log-likelihood ratio, confirming their significance. It turns out that 50–70% of all taverns are found on a small number of street segments that account for approximately 1% of the total length of the streets in the study area, supporting the validity of the outcomes. It should be also noted that the relative risk, which represents the ratio of point densities within and outside the sub-network, remained high under all shape constraints.

Figure 7a shows the outcome under the condition of no penalties. The result is the least compact sub-network (with the $K_{NT}(Z) = 0.35$); i.e. a long and complex form of sub-network that covers the main boulevards stretching west from the Sendai Terminal Train Station on the right and extending all the way to the city's main entertainment district area in the upper left (highlighted in pale blue). The result marks the highest log-likelihood ratio but lowest tavern density among all scenarios with different compactness correction values (Table 3). This is partly because the likelihood ratio tends to become large for a larger spatial extent, prioritising a greater difference in point densities within and outside the search window, despite that the point density decreases as the spatial extent increase. In other words, when there is no penalty on shape, the result tends to take a highly irregular, wide area which may be a collection of multiple clusters.

As the penalty becomes stronger, the network segment cluster reduces in size, and its shape becomes more compact (Fig. 7b, c). This shows that the penalty prevents the network segments from annexing the surrounding regions freely, resulting in the detection of a compact area with a moderate likelihood ratio but with a high point density. At the same time, the shape of the network segment cluster retains some degree of freedom which may lead to the inclusion of adjacent segments without points. Finally, as we impose the strongest penalty ($a = 1$) (Fig. 7d), it yields the most compact cluster, but it also includes segments with no points, thus reducing the point density and the relative risk.

It seems that the excessively strong penalty has created some noise (in Case (d)). The other extreme (Case (a)) does not also seem to be the most suitable cluster due to the lowest point density among all results. Case (c) may be the most appropriate cluster (because of the highest point density and a moderately high relative risk), but Case (a) also provides important information in that it identified a geographically meaningful cluster—the boulevard with a relatively high density of taverns.

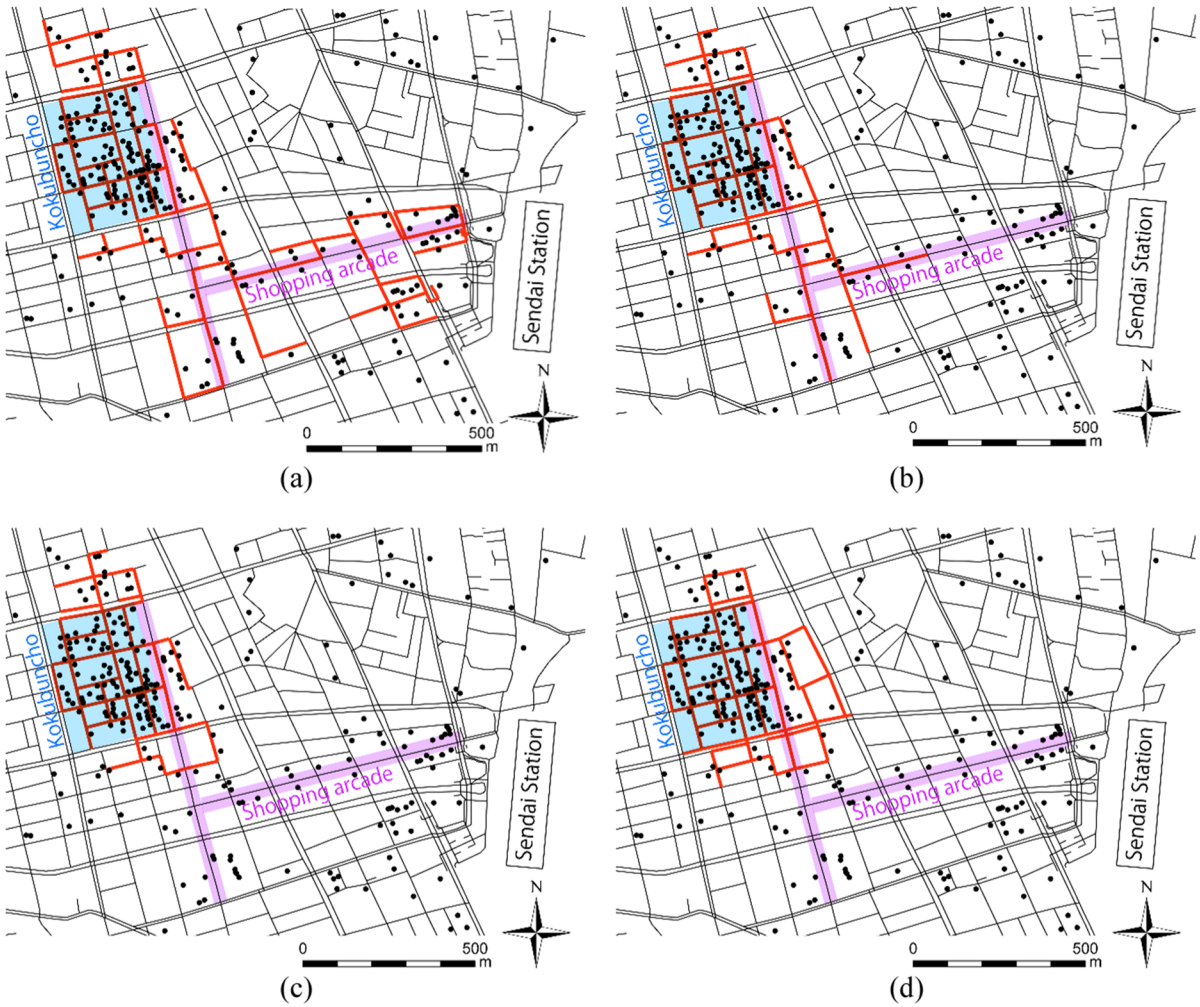

Fig. 7 Detection of clusters of taverns along the street segment network. The black circles denote tavern locations and the red line segments show the extent of the detected NT-segment cluster. The main boulevard and the entertainment district are shown in purple lines and pale blue polygon, respectively. **a** MLC NT-segment cluster detected at $a=0.0$ (no NT-compactness constraint); **b** MLC NT-segment cluster at $a=0.6$; **c** MLC NT-segment cluster at $a=0.8$; and **d** MLC NT-segment cluster at $a=1.0$

Table 3 Result statistics of NT-segments detection against taverns in Sendai City

	Compactness correction (a)	Compactness index ($K_{NT}(Z)$)	Log-likelihood Ratio	p-value	# Taverns (% of total)	Street length (km) (% of total)	Relative risk	Tavern density (/km)
(a)	0.0	0.35	1201	>0.001	350(71%)	8.99(1.4%)	174	38.9
(b)	0.6	0.60	924	>0.001	270(55%)	6.03(0.93%)	129	44.8
(c)	0.8	0.81	867	>0.001	241(49%)	4.32(0.67%)	142	55.8
(d)	1.0	0.90	802	>0.001	242(49%)	5.80(0.90%)	107	41.7
Study area	–	–	–	–	492	647	–	0.764

Detection of weighted network-segment clusters

As this study uses street segments as the aggregation unit (equivalent to an areal unit in the Euclidean space analysis), we could apply a weight factor against each street segment (instead of a simple street length) in the same analysis. The weight factor could be an environmental, social and any other attribute. As an illustrative example, this study uses the total floor space of buildings in each area as a weight factor to calculate the expected number of stores against the floor space available. The floor space of each building is derived from the building footprint and the number of floors for each building is obtained from Zenrin Co's commercial "ZmapTown II". These values are assigned to the nearest street segment as a weight factor. The assumption is that the distribution of taverns along the street network would follow a Poisson distribution with the parameter of the total floor space for each street segment (instead of the street length). This works in the same way as the weighting of the *population at risk* used in the scan statistics analysis in spatial epidemiology.

Figure 8a shows the outcome of the analysis with respect to the floor space of taverns as the weight factor with no penalty for compactness ($a = 0.0$). As the central district around the station has a larger number of multi-story buildings and a relatively high volume of floor space assigned to the street segments in that area, the expected number of taverns in the area has increased in the analysis, which in turn reduced the log-likelihood ratio to 773 and relative risk to 51.0, respectively; compared to the outcomes of a non-weighted scenario. The spatial arrangement of the resulting cluster still resembles that of Fig. 7a (Fig. 7a: $K_{NT}(Z) = 0.35$, #Taverns = 350; Fig. 8a: $K_{NT}(Z)$=0.36, #Taverns = 331).

Figure 8b shows the outcome of the same weighted scenario but with $a = 0.8$. The cluster highlighted in red shows the MLC, which is much more compact and smaller, created by truncating the areas that are less significant within the non-constrained MLC. It should be noted, however, that areas that fell outside the MLC may still hold significant concentration of taverns. Therefore, the method was applied repeatedly by eliminating the extent of previously detected cluster(s) as we have done with a small synthetics data. Figure 8b shows the three most likely clusters: namely, red, blue and green clusters—in the order of highest likelihood—all of which are statistically significant. Interestingly, the total extent covered collectively by the three clusters is similar to that of the single MLC with no penalty (Fig. 8a).

The fact that the three clusters are detected in that order implies that the strongest concentration evidently exists within the entertainment district; whilst

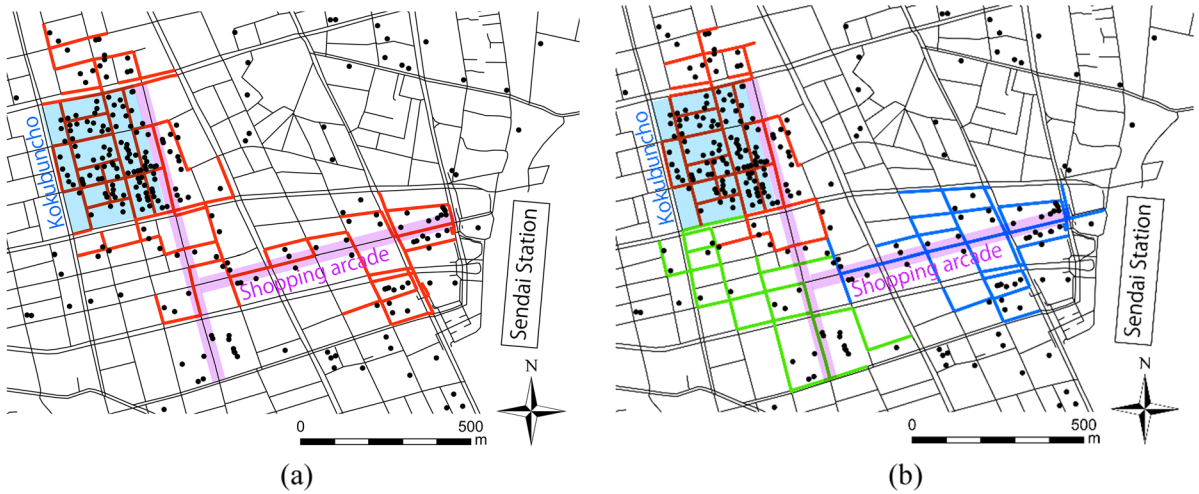

Fig. 8 Detection of clusters of taverns along the street segment network around Sendai Station, weighted by the floor space. The black circles denote tavern locations, and the red line segments show the extent of the detected weighted NT-segment cluster. The main boulevard and the entertainment district are shown in purple lines and pale blue polygon, respectively. **a** MLC NT-segment cluster detected at $a = 0.0$; **b** MLC NT-segment cluster at $a = 0.8$

the secondary and tertiary clusters are respectively located around the station and the in-fill between the two areas. Interestingly, these areas seem to form a clear territory of their own in the sense that the clusters did not result in an intricate combination that goes in and out of the streets in the other areas but, rather, seem to form a relatively compact cluster of its own. Each district naturally forms clusters of different likelihood ratios, i.e. the expected density and the likelihood within each cluster will likely reflect the extent and the nature of their concentration.

Discussion

In this study, we proposed a new spatial cluster detection method to find irregular-shaped clusters along a street network. The study took an approach to derive different cluster sets by changing the compactness correction parameter values which controls the degree of geometric shape penalisation. The method was firstly applied to a simple, synthetic data set and then to a real-world data. If we call a detected cluster that achieves the highest detection accuracy as the optimum solution, then the optimum solution for each of the two small synthetic data sets was at the opposite extreme ends; one was achieved with no compactness correction ($a = 0$) and the other with full compactness correction ($a = 1$). This is partly because the shape of the synthetic test clusters was too simple, but it also means that the result would depend on the shape of the underlying true cluster and their spatial arrangement. The accuracy of the cluster detection was measured with respect to sensitivity and PPV, but this was only possible because we knew the exact shape and location of the synthetic clusters. In other words, we cannot measure the sensitivity or PPV with the real-world data, as there is not enough information about the shape and the arrangement of the true clusters.

The sensitivity analyses have indeed demonstrated a benefit of running cluster detection across different compactness correction values. In the case of the taverns in Sendai City, the optimal solution did not seem to be at the either extreme end of the compactness. What forms the optimum could be a debatable, but the relative risk and the point density of the detected clusters could be referred to as two diagnostic measurements for selecting the optimum, as they reflect accuracy of cluster detection. Under the unweighted scenario (Fig. 7), the optimum cluster with the highest accuracy seemed to be achieved at $a = 0.8$ (Fig. 7c). This may change if other criteria are imposed in the form of weight factors, and the subsequent analysis demonstrated how this can be facilitated.

This study delivered a proof-of-concept analysis to demonstrate that it can deliver solutions of network-segment-based clusters with a flexibility in their shape. It would benefit from applications to other events and features observed along a street network, including events such as crime and epidemiological outbreaks that could benefit from different criteria and scenarios to weigh and configure the shape of the clusters. While its effectiveness in different contexts still needs to be explored, the proposed method is expected to offer the analysts a better chance of finding more accurate clusters among a range of possible solutions along networks.

Detecting clusters whilst changing the extent of penalties on the compactness also shows that there is a range of values that enables us to increase the chance of discovering more than one statistically significant clusters which, with a different compactness penalty, may be detected as a single larger cluster. This would give new interpretation on the context of cluster formation and meaning to the detected clusters in that some clusters may be close enough to be related with one another, but they can be also recognised as a group of conjoined clusters. In the context of taverns, this typically arises because they are concentrated in an entertainment district, but they tend to create small sub-clusters within the district around a landmark or along a main boulevard to make them visible to their potential customers. In this sense, the capacity to offer solutions with different degrees of complexity could help capture the clusters under different criteria and, thereby, provide new knowledge and understanding of the features studied.

Funding This research was conducted in part with support from Japan Society for the Promotion of Science (JSPS), the Grants-in-Aid for Scientific Research (KAKENHI) programme [Grant Numbers 18H01552 and 21H01447].

Declarations

conflict of interest We are also happy to confirm that there is no conflict of interest.

Ethical approval This article is an original, unpublished research and it is not under consideration for publication with any other journals. As this research focuses on secondary data analysis, no human subjects or sensitive data have been used in the study.

References

Aldstadt J., & Getis, A. (2006). Using AMOEBA to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis, 38*(4), 327–343. https://doi.org/10.1111/j.1538-4632.2006.00689.x

Anselin, L. (1995). Local Indicators of Spatial Association-LISA. *Geographical Analysis, 27*(2), 93–115.

Besag, J., & Newell, J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society A, 154*(1), 143–155.

Barricell, N. A. (1957). Symbiogenetic evolution processes realized by artificial methods. *Methodos, 9*(35–36), 143–182.

Cançado, A. L., Duarte, A. R., Duczmal, L. H., Ferreira, S. J., Fonseca, C. M., & Gontijo, E. C. (2010). Penalized likelihood and multi-objective spatial scans for the detection and inference of irregular clusters. *International Journal of Health Geographics, 9*, 55.

Conley, J., Gahegan, M., & Macgill, J. (2005). A genetic approach to detecting clusters in point-data sets. *Geographical Analysis, 37*(3), 286–314.

Costa, M. A., Assunção, R., & Kulldorff, M. (2012). Constrained spanning tree algorithms for irregularly-shaped spatial clustering. *Computation Statistics and Data Analysis, 56*(6), 1771–1783.

Diggle, P. J., & Chetwynd, A. D. (1991). Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics, 47*(3), 1155–1163.

Duczmal, L., & Assunção, R. (2004). A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics and Data Analysis, 45*(2), 269–286.

Duczmal, L., & Buckeridge, D. L. (2006). A workflow spatial scan statistic. *Statistics in Medicine, 25*(5), 743–754.

Duczmal, L., & Cançado, A. (2017). Irregular shaped spatial clusters: Detection and inference. In S. Shekhar, H. Xiong, & X. Zhou (Eds.), *Encyclopedia of GIS* (pp. 1086–1092). Cham: Springer International Publishing, Switzerland.

Duczmal, L., Cançado, A., Takahashi, R., & Bessegato, L. (2007a). A genetic algorithm for irregularly shaped spatial scan statistics. *Computational Statistics and Data Analysis, 52*(1), 43–52.

Duczmal, L., Duarte, A., & Tavares, R. (2009). Extensions of the scan statistic for the detection and inference of spatial clusters. In J. Glaz, V. Pozdnyakov, & S. Wallenstein (Eds.), *Scan statistics statistics for industry and technology* (pp. 153–177). Boston: Birkhäuser.

Duczmal, L., Kulldorff, M., & Huang, M. (2006). Evaluation of spatial scan statistic for irregularly shaped clusters. *Journal of Computational and Graphical Statistics, 15*(2), 428–442.

Duczmal, L., Moreira, G. J. P., Ferreira, S. J., & Takahashi, R. H. C. (2007b). Dual graph spatial cluster detection for syndromic surveillance in networks. *Advances in Disease Surveillance, 4*, 88.

Fogel, D. B. (2006). Nils Barricelli - artificial life, coevolution, self-adaptation. *IEEE Computational Intelligence Magazine, 1*(1), 41–45.

Fraser, A. S. (1957). Simulation of genetic systems by automatic digital computers II. Effects of linkage on rates of advance under selection. *Australian Journal of Biological Sciences, 10*(4), 492. https://doi.org/10.1071/BI9570492

Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis, 24*(3), 189–206.

Kim, J., & Jung, I. (2017). Evaluation of the Gini coefficient in spatial scan statistics for detecting irregularly shaped clusters. *PLoS ONE, 12*(1), e0170736.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and Methods, 26*(6), 1481–1496.

Kulldorff, M. (2022) *SaTScan User Guide for Version 10.1* (http://www.satscan.org)

Kulldorff, M., Huang, L., & Pickle, L. (2003). An elliptic spatial scan statistic and its application to breast cancer mortality data in Northeastern United States. *Journal of Urban Health, 80*(Suppl 1), i130–i131.

Kulldorff, M., Huang, L., Pickle, L., & Duczmal, L. (2006). An elliptic spatial scan statistic. *Statistics in Medicine, 25*(22), 3929–3943.

Kulldorff, M., & Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine, 14*(8), 799–810.

Moura, F. R., Duczmal, L., Tavares, R., & Takahashi, R. H. C. (2007). Exploring multi-cluster structures with the multi-objective circular scan. *Advances in Disease Surveillance, 2*, 48.

Neill, D.B., Moore, A.W., Pereira, F. and Mitchell, T. (2005). Detecting significant multidimensional spatial clusters. In: *Advances in Neural Information Processing Systems 17 - Proceedings of the 2004 Conference, NIPS 2004* (Advances in Neural Information Processing Systems). Neural information processing systems foundation.

Neill, D. B. (2012). Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 74*(2), 337–360.

Neill, D. B., McFowland, E., 3rd., & Zheng, H. (2013). Fast subset scan for multivariate event detection. *Statistics in Medicine, 32*(13), 2185–2208.

Ord, J. K., & Getis, A. (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis, 27*(4), 286–306.

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

Rushton, G., & Lolonis, P. (1996). Exploratory spatial analysis of birth defect rates in an urban population. *Statistics in Medicine, 15*(7–9), 717–726.

Sahajpal, R., Ramaraju, G.V. & Bhatt, V. (2004). Applying niching genetic algorithms for multiple cluster discovery in spatial analysis In: *International Conference on Intelligent Sensing and Information Processing*.

Shiode, S., & Shiode, N. (2020). A network-based scan statistic for detecting the exact location and extent of hotspots along urban streets. *Computers, Environment and Urban Systems, 83*, 101500.

Somanchi, S., Choi, D. & Neill, B. D. (2015). StarScan: A novel scan statistic for irregularly-shaped spatial clusters. *Online Journal of Public Health Informatics* **7**(1).

Takahashi, K., Yokoyama, T. & Tango, T. (2010). FleXScan user guide: for version 3.1. Retrieved from https://sites.google.com/site/flexscansoftware/download_e

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

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

Tango, T. (2021). Spatial scan statistics can be dangerous. *Statistical Methods in Medical Research, 30*(1), 75–86.

Tango, T., & Takahashi, K. (2005). A flexible shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics, 3*, 17.

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

Turing, A. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460.

Turnbull, B., Iwano, E. J., Burnett, W. S., Howe, H. L., & Clark, L. C. (1990). Monitoring for clusters of disease: Application to leukemia incidence in Upstate New York. *American Journal of Epidemiology, 132*(1 Suppl), 136–143.

Yiannakoulias, N., Rosychuk, R. J., & Hodgson, J. (2007). Adaptations for finding irregularly shaped disease clusters. *International Journal of Health Geographics, 6*, 28.

Zhang, Z., Assunção, R. and Kulldorff, M. (2010) Spatial scan statistics adjusted for multiple clusters. *Journal of Probability and Statistics*, 11, Article ID 642379.