A method for comparing and classifying point distributions

Yukio Sadahiro* and Yan Liu**

July 2017

*Center for Spatial Information Science, The University of Tokyo
**School of Geography Planning and Environmental Management, The University of Queensland

Corresponding author:
Yukio Sadahiro
Center for Spatial Information Science, The University of Tokyo
5-1-5, Kashiwanoha, Kashiwa-shi, Chiba 277-8568, Japan
Phone: +81-471-36-4310
Fax: +81-3-5841-8521
sada@cis.u-tokyo.ac.jp
Abstract

Keywords: point distributions, comparison and classification, spatial scale

Spatial analysis often faces point distributions representing a wide variety of spatial objects. A first step of point data analysis is to evaluate the distribution of the data and classify them into groups with similar patterns. It helps us to understand the relationship and common properties of point distributions as well as the underlying structure that determines the distributions. This paper proposes a novel method of classifying and comparing point distributions, and analyzing their relationships at a variety of spatial scales. Compared with existing approaches, this method is more robust to positional errors that are inevitable in spatial data. The validity of the method is tested through its application to the analysis of trip behavior data of the public transport users in Brisbane, Australia. The results support the technical soundness of the method, and reveals travel patterns that cannot be easily obtained by visual analysis and other existing methods.
1. Introduction

Spatial analysis often faces point distributions representing a wide variety of spatial objects. Retail geography discusses the spatial patterns in the distribution of retail stores, service shops, and restaurants. Sociology and demography analyze the mixture and segregation of different ethnic groups in urban areas. Ecology treats the relationship between the distributions of different species of plants and animals.

A first step of analyzing multiple distributions is to evaluate the data and classify them into groups with similar patterns. It helps us to understand the relationship and common properties of point distributions and explore the underlying structure that determines the distributions. Several options are available for this purpose.

A most popular approach to comparison is the quadrat method (Diggle (1983); Upton and Fingleton (1985)). Overlaying a lattice on point distributions, we count the number of points in each cell and compare them by the $\chi^2$-square statistic. Another option is the nearest neighbor spatial association measure $R^*$ proposed by Lee (1979). The measure evaluates the spatial closeness of two sets of points, by which we can statistically test whether the sets are spatially close, separated, or independent. A revised version of $R^*$ is the conditional nearest-neighbor spatial-association measure proposed by Okabe and Miki 1984). Their method evaluates the similarity of one distribution with respect to another, which yields asymmetrical measures. The method is appropriate when an asymmetrical association is expected between point distributions such as cause and result relationship. Instead of a single measure, Ripley proposed a function that evaluates the similarity between two sets of points. The cross $K$-function represents the similarity as a function of distance variable (Ripley (1976); Ripley (1977); Ripley (2005); Cressie (2015)).

Classification of point distributions can be performed by evaluating the similarity between point distributions by a single measure. Among the methods mentioned above, the quadrat method, $R^*$, and its revised forms satisfy this condition. The distances between every pair of distributions form a distance matrix, which become the basis of classification using an existing method of cluster analysis such as single-linkage, complete-linkage, and Ward’s method.

The evaluation of similarity depends on the spatial scale at which one compares the point distributions. For example, suppose there are five patterns on a one-dimensional space (Figure 1). Distributions $\Gamma_1$, $\Gamma_2$ and $\Gamma_3$ appear similar at a first glance, but a detailed view reveals a small difference between the distributions. We can say that $\Gamma_1$, $\Gamma_2$ and $\Gamma_3$ are similar as a whole with a slight variation, and that $\Gamma_1$ is more similar to $\Gamma_2$ than $\Gamma_3$. Distribution $\Gamma_4$ consists of four points distributed at regular intervals and another four clustered at other locations. One may think that $\Gamma_4$ is not totally but partially similar to $\Gamma_1$. Distribution $\Gamma_5$ also contains points clustered at one location, though further apart from the others. It would be easier to reach the consensus that $\Gamma_4$ and $\Gamma_5$ are partially similar given that both have points that are clustered at one location. However, it would be more difficult to think that the overall patterns of $\Gamma_4$
and $\Gamma$s are similar. Evidently, the evaluation of similarity of point data is scale-dependent, thus requiring explicit considerations of the scale factor in the comparison and analysis of point distributions.

Unfortunately, the nearest neighbor spatial association measure, $R^*$, does not consider the spatial scale factor. Quadrat analysis and cross $K$-function, on the other hand, do take into consideration the scale effect in the analysis. In the quadrat method, the size of cells used for counting points represents the spatial scale. A large cell size conceals small differences between points and thus the distributions tend to look similar. A small cell size emphasizes the difference between cells so that the distributions appear different. The cross $K$-function represents the similarity as a function of spatial scale and thus evaluates the similarity at various scales. Quadrat analysis and cross $K$-function, however, are more prone to errors in positional accuracy as these methods measure the number of points contained in cells divided by crisp boundaries. A slight error in the positioning of points may drastically change the evaluation of similarity. Such instability in evaluation is critical since spatial data are generally subject to noise and measurement errors. In addition, the assessment of similarity using quadrat analysis depends not only on the spatial scale but also the starting location to tessellate the space. Therefore, the effects of scale and location are inseparable and thus it is not possible to evaluate the role of spatial scale explicitly when comparing and classifying point distributions using this method. The cross $K$-function is free from the location issue, but is nevertheless inappropriate for classifying point data as it does not provide a single measure.

This paper proposes a new method applicable to the classification and comparison of point

![Figure 1: Point distributions on a one-dimensional space.](image-url)
Section 2 describes the method with illustrations. Section 3 applies the method to the analysis of trip patterns in Brisbane, Australia. Section 4 summarizes the conclusions with discussion.

2. Methods

2.1 Representation of point distributions

There has been a long debate on the definition of scale in geography (extensive reviews include Lam and Quattrochi (1992), Quattrochi and Goodchild (1997), Atkinson and Tate (2000), Wu and Li (2009), and Zhang et al. (2014)). Lam and Quattrochi (1992) proposes three meanings of scales: geographic scale, operational scale, and cartographic scale. Geographic scale refers to the spatial extent in which spatial analysis is performed. Operational scale is the spatial extent at which a spatial phenomenon operates. Cartographic scale refers to the ratio between the length of objects on a map and that in the real world. Lam and Quattrochi (1992) treats resolution as a concept closely related to scale. Study in a small region, i.e., small-scale study generally requires spatial data of fine resolution, while large-scale study often uses spatial data of coarse resolution. The terms scale and resolution are often used interchangeably because of their close relationship. Scale and resolution, however, are fundamentally different concepts and thus they are not totally dependent with each other. Quattrochi and Goodchild (1997) clearly distinguishes resolution from other definitions of scales, and Atkinson and Tate (2000) calls it measurement scale. Wu and Li (2009) distinguishes modeling scale from operational scale, and add policy scale. Modeling scale is the scale at which a spatial phenomenon is modelled, which can be different from the actual scale of spatial phenomenon operation. Policy scale refers to the scale at which political decisions are made. Liu et al. (2015) distinguishes the scale of analysis from measurement scale. Measurement scale defines the maximum resolution at which analysis can be performed. Analysts often reduce the resolution of original data to find spatial patterns of coarser scales.

This paper uses the term "scale" to refer to the analytical scale, because analytical methods mentioned in Section 1 implicitly assume this usage. Cell size in quadrat method, circle radius in $K$-function, and window width in kernel smoothing represent analytical scale. The term "fine/coarse scale" indicates analysis performed based on spatial data of high/low resolution. For instance, quadrat method of small cells is an analysis at a fine scale while a wide window in kernel smoothing provides a coarse scale analysis.

Analytical scale is equivalent to the concept of scale discussed in scale-space theory developed in image processing and computer vision (Witkin (1984); Koenderink (1984); Lindeberg (1994)). This gives a theoretical basis of representing images at multiple scales. Scale-space theory claims that representations at coarse scales should constitute simplifications of representations at finer scales. Smoothing transformation is performed based on the convolution by the Gaussian kernel (Leung et al. 2000).
This paper evaluates, at present, a point distribution at every location and treats all points equally within a given distance $h_C$ from the location. This is equivalent to the cross $K$-function except that we evaluate the distribution at every location. Suppose $N$ types of point distributions denoted by $\Gamma = \{ \Gamma_1, \Gamma_2, \ldots, \Gamma_N \}$. Point distribution $\Gamma_i$ consists of $n_i$ points ($i=1, \ldots, N$). Let $P_j$ and $z_{ij}$ be the $j$th point and its location in set $\Gamma_i$, respectively. Standing at location $x$, we see points in the circle of radius $h_C$ centered at $x$ and treat them equally independent of the distance from $x$. This implies, in other words, that we interpret the point distribution at each location by the number of points within distance $h_C$. This view on point distribution $\Gamma_i$ is mathematically represented as a function of location $x$:

$$F_i(x, h_C; \delta_C) = \sum_j \delta_C(x, z_{ij}, h_C),$$

where $\delta_C(x, z_{ij}, h_C)$ is an indicator function:

$$\delta_C(x, z_{ij}, h_C) = \begin{cases} 
1 & \text{if } |x-z_{ij}| \leq h_C \\
0 & \text{otherwise} 
\end{cases}.$$

We call $F_i(x, h_C; \delta_C)$ the interpreted surface of points $\Gamma_i$, because it represents our analytical interpretation of point distribution, i.e., the way of treating the distribution in analysis. Figure 2 shows some examples of $F_i(x, h_C; \delta_C)$ calculated based on distributions $\Gamma_1$, $\Gamma_4$, $\Gamma_6$ and $\Gamma_9$ in Figure 1. As seen in the figure, Equation (1) transforms point distributions into stepwise functions that indicate the number of points within distance $h_C$ at each location. Parameter $h_C$ defines the spatial extent of view at each location; a large $h_C$ implies the consideration of a wide area while a small $h_C$ provides a narrow view. Since $h_C$ works as an indicator of analytical scale, we call $h_C$ a scale parameter, hereafter.

The role of $h_C$ emerges in the smoothness of obtained surface, i.e., a small $h_C$ converts a point distribution into a rough surface with many peaks (Figure 2a) while a large $h_C$ decreases the peaks of the surface (Figure 2b). The former is more similar to the original distribution because it evaluates the location of each point more accurately. A surface based on a large $h_C$ implies that we only consider the rough distribution of points based on their approximate location. Interpreted surface $F_i(x, h_C; \delta_C)$ approaches a uniform surface in any distribution of points when $h_C \to \infty$ and hence the obtained surfaces look quite similar. This is because the difference between point distributions interpreted at finer scales is concealed by the similarity between distributions observed at coarser scales.

Given a certain $x$, interpreted surface $F_i(x, h_C; \delta_C)$ represents the accumulation of spatial phenomena interpreted at scales finer than $h_C$. Consequently, $F_i(x, h_C; \delta_C)$ as a function of $h_C$ increases monotonically with $h_C$. The change in value of $F_i(x, h_C; \delta_C)$, on the other hand, indicates the point distribution interpreted exactly at $h_C$. $F_i(x, h_C; \delta_C)$ is a stepwise function that increases only when a point is on the ring of radius $h_C$ centered at $x$ (a similar discussion can be found in Kiskowski et al. (2009)).
Figure 2 Interpreted surfaces calculated based on the point distributions in Figure 1. (a) Interpreted surfaces with a small $h_C$. (b) Interpreted surfaces with a large $h_C$. Points in each distribution are labelled in ascending order from left to right.

Given that $F_d(x, h_C; \delta c)$ is a stepwise function, its value can increase or decrease rapidly at the boundary of the circle of radius $h_C$ centered at each point. This property is not desirable in analyzing spatial data since the result of analysis based on such functions can change drastically by even a slight error, which is often included in positional information of spatial data. To assure the robustness against such data errors, we introduce a different representation of point distribution. Standing at location $x$, we evaluate the spatial phenomena with a distance-decaying weight defined over an infinite space, i.e., we
interpret the spatial phenomena in the near neighborhood of \( x \) with more weight than those in distant places. We count the number of points with a distance-decaying weight defined by a function \( \delta(|x - z_j|, h) \). This gives another definition of the interpreted surface:

\[
F_i(x, h; \delta) = \sum_j \delta\left(|x - z_j|, h\right).
\]

(3)

Similar to \( h_C \), \( h \) in Equation (3) determines the scale of the analysis; a small \( h \) represents a fine scale that gives a more detailed view of the point distribution while a large \( h \) represents a coarse scale.

The interpreted surface \( F_i(x, h; \delta) \) is equivalent to kernel smoothing (Silverman 1986). Kernel smoothing puts a small bump called a *kernel* at each observation and sums the kernels up to obtain a surface function. Each kernel and its summation corresponds to \( \delta(|x - z_j|, h) \) and \( F_i(x, h; \delta) \), respectively. This paper adopts the Gaussian kernel as \( \delta(|x - z_j|, h) \) that is most frequently used in kernel smoothing:

\[
\delta\left(|x - z_j|, h\right) = e^{-\frac{|x - z_j|^2}{2h^2}}.
\]

(4)

The interpreted surface becomes

\[
F_i(x, h; \delta) = \sum_j e^{-\frac{|x - z_j|^2}{2h^2}}.
\]

(5)

Figure 3 shows the interpreted surfaces \( F_i(x, h; \delta) \) from the point distributions in Figure 1. Unlike surfaces in Figure 2, those in Figure 3 are continuous, and their shapes do not change drastically due to possible positional errors in the points data. The results of analysis based on surfaces \( F_i(x, h; \delta) \) are more stable than those on \( F_i(x, h_C; \delta) \), which assures the robustness of our method. A small \( h \) generates rough surfaces in Figure 3a while a large \( h \) yields smooth surfaces (Figure 3b). The roughness in surfaces \( F_1 - F_3 \) observed in Figure 3a disappears in Figure 3b, and hence all the surfaces look similarly uniform. This is because the coarse scale conceals the variation between the distributions in Figure 3b, which supports our earlier observation that the distributions are similarly uniform with a slight variation.
Figure 3 Interpreted surfaces calculated based on the point distributions in Figure 1. (a) Interpreted surfaces with a small scale parameter $h$. (b) Interpreted surfaces with a large scale parameter $h$.  

Similar to the interpreted surface $F_i(x, h; \delta C)$, $F_i(x, h; \delta)$ represents the accumulation of spatial phenomena interpreted at scales finer than $h$, and the change in value indicates the point distribution interpreted exactly at $h$. The latter is mathematically represented as

$$F'_i(x, h; \delta) = \frac{d}{dh} F_i(x, h; \delta)$$

$$= \sum_{j} \frac{d}{dh} e^{-\frac{|x-z_j|^2}{2h^2}}$$

$$= \sum_{j} \left| \frac{x-z_j}{h^2} \right|^2 e^{-\frac{|x-z_j|^2}{2h^2}}$$

(6)

$F'_i(x, h; \delta)$ is the derivative of $F_i(x, h; \delta)$ with respect to $h$, which we call the interpreted density of points $\Gamma_i$ at scale $h$. It represents the point distribution interpreted exactly at scale $h$, i.e., the distribution interpreted on the ring of radius $h$ centered at $x$. Thus, the interpreted surface $F_i(x, h; \delta C)$ can be obtained as:
\[ F'_i(x, h; \delta) = \int_{h_0=0}^{h} F'_i(x, h_0; \delta) \, dh_0. \]

Figure 4 shows the interpreted densities \( F'_i(x, h; \delta) \) calculated based on distributions \( \Gamma_1, \Gamma_4, \Gamma_6 \) and \( \Gamma_9 \) in Figure 1. \( F'_i(x, h; \delta) \) increases around points, the degree of which depends on the distance from the origin and that from points. \( F'_i(x, h; \delta) \) increases drastically near the origin when the both distances are small as seen in \( F'_1(x, h; \delta) \) and \( F'_4(x, h; \delta) \) in Figure 4. \( F'_i(x, h; \delta) \) becomes long-tailed in \( \Gamma_4, \Gamma_6 \) and \( \Gamma_9 \) that contain point clusters located far from the origin.

2.2 Comparison of point distributions

Using the interpreted surface, we evaluate the similarity between point distributions. The similarity measure between \( \Gamma_i \) and \( \Gamma_k \) is defined as

Figure 4 Interpreted density \( F'_i(0, h; \delta) \) as a function of \( h \) calculated based on the point distributions in Figure 1, where \( x=0 \) is the starting point of the horizontal axis on which points are distributed.
\[ S_{ik}(h) = 1 - \frac{1}{2} \int \left[ f_i(x, h; \delta) - f_k(x, h; \delta) \right] dx , \]

(8)

where \( f_i(x, h; \delta) \) is a standardized form of \( F_i(x, h; \delta) \):

\[
f_i(x, h; \delta) = \frac{1}{n_i} \sum_j \delta \left( \left| x - x_j \right|, h \right) \]

\[
= \frac{1}{\sqrt{2\pi n_i h}} \sum_j e^{-\frac{|x - x_j|^2}{2h^2}} ,
\]

(9)

\( (k=1, \ldots, N, \text{ but } k \neq i) \). We use \( h \) in logarithmic form \( S_{ik}(h) \) in the following since \( S_{ik}(h) \) increases very rapidly when \( h \) is small. The measure \( S_{ik}(h) \) reaches its minimum and maximum values when \( h=0 \) and \( h \to \infty \), respectively. The maximum value of \( S_{ik}(h) \) is 1 while its minimum value is

\[
S_{ik}^{MIN} = 1 - \frac{1}{2} \left( \frac{n_i - m_k}{n_i} + \frac{n_k - m_{ik}}{n_k} \right)
\]

\[
= \frac{n_i + n_k}{2n_i n_k} m_{ik}
\]

(10)

where \( m_{ik} \) is the number of points in \( \Gamma_i \) that shares the same location with a point in \( \Gamma_k \). If the locations of all the points in \( \Gamma_i \) and \( \Gamma_k \) are different, \( m_{ik}=0 \); therefore, \( S_{ik}(h) \) ranges from zero to one.

The measure \( S_{ik}(h) \) permits us to evaluate the similarity between \( \Gamma_i \) and \( \Gamma_k \) with an explicit consideration of spatial scale. The measure \( S_{ik}(h) \) evaluates the accumulation of similarity between \( F_i(x, h; \delta) \) and \( F_k(x, h; \delta) \) at scales finer than \( h \), and the change of \( S_{ik}(h) \) indicates the similarity exactly at scale \( h \). The latter is mathematically represented as

\[
S'_{ik} = \frac{d}{dh} S_{ik}(h)
\]

\[
= -\frac{1}{2} \frac{d}{dh} \int \left[ f_i(x, h) - f_k(x, h) \right] dx ,
\]

(11)

which we call differential similarity measure.

Figure 5a and Figure 5b illustrate the changing patterns of \( S_{ik}(h) \) and \( S'_{ik}(h) \) between distribution \( \Gamma_1 \) and all other distributions (\( \Gamma_2 - \Gamma_9 \)), respectively. Figure 5c also shows \( S_{ik}(h) \) between \( h=0.01 \) and \( h=10 \), where \( S_{ik}(h) \) has very low peaks. The peaks are critical, however, since \( S_{ik}(h) \) at large \( h \) occupies a considerable portion of \( S_{ik}(h) \) (recall \( h \) is shown in logarithmic scale). The scale of observation gradually
changes from high to low with an increase in $h$ and thus the difference between distributions vanishes. Consequently, $S_d(h)$ is an increasing function of $h$, and $S'_d(h)$ is non-negative for any $h$ with multiple peaks of different height.

Figure 1 shows that distributions $\Gamma_2$ and $\Gamma_3$ are similar to $\Gamma_1$ as a whole, while $\Gamma_4$ to $\Gamma_9$ are different since they contain point clusters. Measures $S'_{12}(h)$ and $S'_{13}(h)$ are characterized by their significant peaks at small $h$, which are much higher than their lower peaks observed in Figure 5c. Measures from $S'_{14}(h)$ to $S'_{19}(h)$ have multiple peaks, those of which at large $h$ are not negligible as shown in Figure 5c. This implies that the similarity as a whole emerges as a significant peak at a fine scale. It is consistent with our earlier discussion, i.e., a slight difference vanishes at a fine scale when point distributions are similar as a whole. When point distributions are not so similar, a coarse scale is necessary to regard the distributions similar with each other.

Distributions $\Gamma_4$, $\Gamma_6$ and $\Gamma_7$ are partially similar to $\Gamma_1$ because they contain several points that are closely located to those in $\Gamma_1$. This results in peaks of $S_d(h)$ at a small $h$ as seen in Figure 5b. These distributions also have peaks at large $h$ (Figure 5c), which implies that a coarse scale is necessary to recognize the similarity between these distributions and $\Gamma_1$. From this we can say that multiple peaks containing peaks at a fine scale indicate the partial similarity between point distributions.

The highest peaks of $\Gamma_4$, $\Gamma_6$ and $\Gamma_7$ are observed at $h=0.005$, which is the half of the distance between some points in $\Gamma_4$, $\Gamma_6$, and $\Gamma_7$ and their nearest points in $\Gamma_1$. For instance, the distance between four points on the left in $\Gamma_4$ and their nearest points in $\Gamma_1$ is 0.01 as seen in Figure 1. Similarly, distributions $\Gamma_3$, $\Gamma_5$ and $\Gamma_8$ have peaks at $h=0.025$, which is also the half the distance between some points in $\Gamma_3$, $\Gamma_5$ and $\Gamma_8$ and their nearest points in $\Gamma_1$. The value of $h$ of the highest peaks contains the information on the distance between neighboring points in different distributions. Appendices A1 and A2 discuss this relationship in detail with a theoretical support.

Distribution $\Gamma_9$ is totally different from $\Gamma_1$. Measure $S'_{19}(h)$ stays zero where $h \leq 0.05$, and reaches its maximum at $h=0.115$. Lack of peaks at a fine scale and a significant peak at a coarse scale indicate that point distributions are different on both fine and coarse scales.

We may summarize the above observation as follows. Given two distributions, we say that they are similar as a whole when $S_d(h)$ has a significant peak at a small $h$. If $S'_d(h)$ has a low peak at a small $h$, the distributions are partially similar. If $S'_d(h)$ does not have any peak at a small $h$, the distributions are totally different.
Figure 5 The relationship between scale parameter $h$ and (a), similarity measure $S_{ik}(h)$ (b)(c), and differential similarity measure $S'_{ik}(h)$. Examples of $\eta_{ik}(0.9)$ are also shown in (a).

The similarity measures can also be defined using quadrat analysis or cross $K$-function. These measures, however, are stepwise functions whose value changes discontinuously at many $h$ values. We cannot define the differentials of $S_{ik}(h)$ effectively to compare point distributions under different spatial scales. In addition, stepwise functions are less robust against positional errors than smooth functions such as $S_{ik}(h)$ and $S'_{ik}(h)$. A slight error may drastically affect the evaluation of similarity at scales around which stepwise functions discontinuously change (see a concrete example in Sadahiro and Liu (2017)).

Given a certain parameter $h$, we can classify point distributions by using the similarity measures. For instance, distance matrix whose elements are defined by $1-S_{ik}(h)$ provides a basis of cluster analysis methods (Everitt et al. (2011); Hennig et al. (2015)). Classification of point distributions at $m$ scales runs in

$$O \left( km \left( \frac{N - k}{N} \right)^2 \sum_{i} n_i \right) \approx O \left( kmN^2 \bar{n} \right)$$

(12)
time, where a fast $k$-medoids algorithm is applied at $m$ scales (Tiwari and Singh 2012). This computational complexity, unfortunately, is acceptable only at a single scale since it increases in proportion to $N^2$. Multiscale classification is practically impossible, which is highly recommended to find an appropriate classification useful for detecting interesting spatial patterns and relationships.

An alternative to multiscale classification is to classify point distributions by a single measure that summarizes the similarity between distributions at various scales. The similarity between point distributions increases monotonically by accumulating the similarity at increasing scales. The measure $S_{ik}(h)$ increases very rapidly when $\Gamma_i$ and $\Gamma_k$ are similar at a fine scale, while it increases slowly when $\Gamma_i$ and $\Gamma_k$ are dissimilar as a whole. Therefore, we can evaluate the overall similarity of distributions by evaluating the rate at which $S_{ik}(h)$ becomes similar with the increase in $h$. We define a measure $\eta_{ik}(\alpha)$ in an implicit form as:

$$\alpha = \int_{h=0}^{\eta_{ik}(\alpha)} S'_{ik}(h) \, dh = S_{ik}\left(\eta_{ik}(\alpha)\right)$$

(13)
The right side of the equation represents the accumulation of similarity at scales from $h=0$ to $\eta_{ik}(\alpha)$. $\eta_{ik}(\alpha)$ is equal to $h$ when the accumulation reaches $\alpha$. When $\Gamma_i$ and $\Gamma_k$ are highly similar at a fine scale, $S_{ik}(h)$ increases rapidly and the accumulation reaches $\alpha$ at a small $h$. On the other hand, if $\Gamma_i$ and $\Gamma_k$ are not so similar, $S_{ik}(h)$ increases slowly and hence $\eta_{ik}(\alpha)$ becomes large. Measure $\eta_{ik}(\alpha)$ serves as a distance...
measure in cluster analysis.

Parameter $\alpha$ can take any value. However, the comparison of more than two distributions requires a consistent value. The overall similarity is evaluated by considering the point distributions at various scales using a large $\alpha$, as a small $\alpha$ implies that the comparison of distributions is only at fine scales. On the other hand, $\alpha=1$ is meaningless as $\eta_{ik}(1)=\infty$ in any case. We thus recommend using a large $\alpha$ value such as $\alpha=0.9$ or $\alpha=0.95$ as is often adopted as the level of significance in statistical tests.

Figure 5a shows examples of $\eta_{ik}(\alpha)$ where $\alpha=0.9$. The values of $\eta_{ik}(\alpha)$ reveal that distribution $\Gamma_1$ is most similar to $\Gamma_2$ but least similar to $\Gamma_9$. The values also indicate that distributions $\Gamma_2$ and $\Gamma_3$ are even more similar to $\Gamma_1$ than to $\Gamma_4$-$\Gamma_9$, which is consistent with our intuition mentioned earlier.

$\eta_{ik}(\alpha)$ works as a distance measure between $\Gamma_i$ and $\Gamma_k$, and forms a distance matrix that represents the similarity between the point distributions. The matrix gives a basis of classifying the distributions using existing cluster analysis methods (Everitt et al. (2011); Hennig et al. (2015)). Cluster analysis include both hierarchical methods such as the single-linkage, complete linkage and Ward's method, and non-hierarchical methods such as K-means and K-medoids method. The non-hierarchical methods are popular especially in data mining because these methods run faster than the hierarchical clustering methods. While K-means method is most popular in non-hierarchical cluster analysis, it requires the attribute data of elements to calculate the distance between groups repeatedly in the clustering process. In contrast, the K-medoids method is based only on a distance matrix of elements, that is, by choosing an initial set of medoids from the elements, the K-medoids method assigns each element to its nearest medoid. This method recomposes the set of medoids step by step until the summation of the distance within each medoid is minimized. The K-medoids method is more appropriate to use when dealing with large number of point types, hence, this approach is adopted for our purpose.

The method proposed in this section aims for exploratory spatial analysis. An emphasis is on the detection rather than the validation of interesting relationship between point distributions. The method itself does not include statistical test since a strict view of statistics often conceals subtle but interesting patterns in spatial phenomena. Once having detected spatial patterns, however, we usually test their significance through a statistical procedure since it would yield more objective and persuasive results. Statistical test will be considered in subsequent studies.

3. Empirical study

3.1 Study area and data

The method proposed in Section 2 is applied to the analysis of trip patterns of the public transport users travelling on go card in South East Queensland (SEQ), Australia (Figure 6). go card is a transport smart card whose owners can travel on bus, train, ferry and tram services by tapping on and off when they board and alight a service. We collected all trip transaction data made by Pensioner Concession Card (PCC) holders for one week, from 9th to 15th in March, 2015 and randomly sampled 3026 PCC
holders (which is just under 10% of all users of this card type) for analysis of their travel patterns. The PCC is one type of concession cards issued by Australian government, where the card holders are entitled to travel at concession fares. This sampling approach is necessary to reduce the data processing time when testing the proposed method, given that there were 32,970 PCC users in the week with over 200,000 transaction records. Each trip record consists of boarding and alighting times, bus route number, and the identification number of bus/train/ferry stops where each trip commences and ends. The trip transaction data were mapped to the spatial data of road service network using the alighting locations for visualization and spatial analysis.

Figure 6 South East Queensland consisting Brisbane and its surrounding local government areas.

The first boarding and last alighting bus stops in a day by each PCC holder are identical in most
cases. This permits us to infer the home locations of the card holders. Some card holders make no trip in one day while others make multiple trips. In the latter case each card holder’s transaction records compose a sequence of trips in a day. This results in temporal gaps between trips, i.e., the gap between the alighting time of a trip and the boarding time of its subsequent trip. Card holders may stay at a location (i.e., end of a journey) or walk for a short distance to another stop to continue a journey. This paper distinguishes the gap between trips as transits and stays based on the duration of the gap time. Transits are the gaps shorter than 30 minutes while stays are gaps longer than transits.

Some card holders make stays many times in one day. We define a card holder who makes four stays or more in at least one day during a week a frequent commuter. This section focuses on the trip patterns of the frequent commuters of the PCC holders in SEQ. We extracted 257 frequent commuters from the 3026 card holders we initially sampled.

3.2 Classification results of commuter groups

All transaction data by the frequent commuters were classified into three to seven groups based on the location of their stays using the K-medoids method. The threshold $\alpha$ was set to 0.9. Table 1 shows the distances between groups when we classified the data into seven groups (G01-G07). Other number of groups (ranging from three to six) show similar classification results. $\gamma$ in Table 1 represents the average distance between commuters and the medoid of their group and $\rho$ is the average distance between commuters within each group, both of which indicate the variation in the location of stays in each group. The 7 by 7 matrix on the right in Table 1 is the distance matrix between the medoids of groups.

G01-G03 are larger in size than other groups, accounting for 75.9% of all the frequent commuters sampled. Group G02 shows small $\gamma$ and $\rho$ values despite the largest number of commuters, which implies that commuters in this group made stays at very similar locations. On the other hand, the large $\gamma$ and $\rho$ values in G06 indicate a wide spread in the location of stays by this group. The distance matrix also shows that there is a large distance between Groups G01-G03 and Groups G05-G07 (i.e., all over 400m). Groups G01-G03 are relatively closer to each other; G06 is separated from G05 and G07 (1000m) while G05 and G07 are rather close (128.9m). Group G04 sits in between G01-G03 and G05-G07; G04 is close to G02, G03 and G05 while separated from G01, G06, and G07.

Table 1 Distances between commuters and groups. $\gamma$ is the average distance between commuters and the medoid of their group, while $\rho$ is the average distance between commuters within each group. The 7 by 7 matrix on the right is the distance matrix between groups, i.e., the distances between the medoids of groups. All distances are measured in meters.
Figure 7 shows the location of homes and stays of the frequent commuters, demonstrating a spatial closeness between the home locations and the locations they travelled to. Commuters in general made stays in the same area of their residence. For instance, commuters in group G01 primarily live and made stays in the Moreton Bay region, while G05 in the Sunshine Coast and G06 and G07 in the Gold Coast regions. Groups G01-G04 made stays in and around the Brisbane area, which are rather separated from G05-G07. This is consistent with our observation in Table 1, i.e., groups G01-G04 share similar pattern of stays that yield small between-group distances. Stays of G02 and G07 are tightly clustered (Figure 7c and h), while those of G01, G04 and G06 are rather scattered (Figure 7b, e and g), which is also confirmed by the small $\rho$ value in G02 (19.07m) and G07 (12.91m) but rather large $\rho$ value in G01 (77.76m), G04 (52.65m) and G06 (87.10m) (Table 1). It is speculated that the former groups (G02 and G07) made trips for working and shopping around their homes, while commuters in the latter (G01, G04 and G06) travelled within their home area as well as to distant places for work or other purposes.

<table>
<thead>
<tr>
<th>Number of commuters</th>
<th>$\gamma$</th>
<th>$\rho$</th>
<th>G01</th>
<th>G02</th>
<th>G03</th>
<th>G04</th>
<th>G05</th>
<th>G06</th>
</tr>
</thead>
<tbody>
<tr>
<td>G01</td>
<td>37</td>
<td>45.59</td>
<td>77.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G02</td>
<td>110</td>
<td>10.35</td>
<td>19.07</td>
<td>250.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G03</td>
<td>48</td>
<td>27.31</td>
<td>44.36</td>
<td>359.4</td>
<td>62.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G04</td>
<td>19</td>
<td>30.84</td>
<td>52.65</td>
<td>484.4</td>
<td>187.5</td>
<td>80.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G05</td>
<td>14</td>
<td>18.67</td>
<td>35.90</td>
<td>812.5</td>
<td>515.6</td>
<td>406.3</td>
<td>281.3</td>
<td></td>
</tr>
<tr>
<td>G06</td>
<td>19</td>
<td>46.84</td>
<td>87.10</td>
<td>523.4</td>
<td>828.1</td>
<td>937.5</td>
<td>1062.5</td>
<td>1000.0</td>
</tr>
<tr>
<td>G07</td>
<td>10</td>
<td>6.24</td>
<td>12.91</td>
<td>968.8</td>
<td>679.7</td>
<td>578.1</td>
<td>453.1</td>
<td>1000.0</td>
</tr>
</tbody>
</table>
Figure 7 The location of homes (a) and stays (b-h) of the frequent commuters classified into seven groups. Color of points in G01 to G07 indicates the relative number of stays made by the frequent...
commuters at each stop, which ranges from 0 to 3 or more stays, with the reddish points indicating more frequent while greenish less frequent stay locations. Numbers in parentheses are the number of frequent commuters in each group.

Commuters can be broadly classified into two groups, i.e., one in G01-G04 and the other in G05-G07. Members in the latter broad group are distinctive to each other with each group sharing similar travel features, that is, the commuters mostly travel within their respective local areas. Groups G01-G04, on the other hand, share a common feature of less number of stays and smaller between-group distances, as shown in Table 1 and Figure 7, however, it is difficult to uncover other differences or similarities amongst these groups.

3.3 Reclassification

We thus regrouped the 214 commuters in G01-G04 together and reclassified them into seven subgroups (G11 to G17) using the same approach described in Section 2. The result is shown in Table 2 and Figure 8. Again, commuters made stays in and around the area of their residence. G12 and G13 consist of commuters living in the Moreton Bay region (Figure 8a), and their stay locations were also largely in this region (Figure 8c and d). These two subgroups have similar distribution of stays featured by a small distance between these groups in the distance matrix (Table 2). G15 and G16 consists of commuters living in the Ipswich and the Logan-Gold Coast regions, respectively; their travel stays were also mainly clustered within their own regions, with some living in Logan but travelling to Brisbane (Figure 8f and g). For G11, G14, and G17, commuters made stays in their home region as well as the surrounding areas. Commuters of G11 and G14 live in the south and north suburbs in Brisbane, respectively, and they made stays in their residential area as well as commuting to the inner city of Brisbane (Figure 8b and e). Commuters of G17 live in the inner city of Brisbane but travel to all parts in Brisbane as well as to the coastal area in the Redlands (Figure 8h). Clearly, Brisbane is a central location that is attractive to many commuters in the surrounding regions. These three subgroups (G11, G14 and G17) form the majority of the frequent commuters, i.e., 63.8% of all the commuters, with similar pattern of stays featured by the small values in the distance matrix (Table 2).

Table 2 Distances between commuters and groups. $\gamma$ is the average distance between commuters and the medoid of their group, while $\rho$ is the average distance between commuters within each group. The 7 by 7 matrix on the right is the distance matrix between groups, i.e., the distances between the medoids of groups. All distances are measured in meters.
<table>
<thead>
<tr>
<th>Number of travelers</th>
<th>$\gamma$</th>
<th>$\rho$</th>
<th>G11</th>
<th>G12</th>
<th>G13</th>
<th>G14</th>
<th>G15</th>
<th>G16</th>
</tr>
</thead>
<tbody>
<tr>
<td>G11</td>
<td>42</td>
<td>12.08</td>
<td>21.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G12</td>
<td>11</td>
<td>15.21</td>
<td>26.76</td>
<td>281.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G13</td>
<td>11</td>
<td>22.42</td>
<td>42.86</td>
<td>437.5</td>
<td>103.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G14</td>
<td>37</td>
<td>7.10</td>
<td>13.10</td>
<td>101.6</td>
<td>121.1</td>
<td>281.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G15</td>
<td>12</td>
<td>11.28</td>
<td>19.07</td>
<td>91.8</td>
<td>351.6</td>
<td>500.0</td>
<td>179.7</td>
<td></td>
</tr>
<tr>
<td>G16</td>
<td>16</td>
<td>30.84</td>
<td>52.65</td>
<td>82.0</td>
<td>406.3</td>
<td>562.5</td>
<td>234.4</td>
<td>117.2</td>
</tr>
<tr>
<td>G17</td>
<td>85</td>
<td>4.87</td>
<td>9.11</td>
<td>44.9</td>
<td>187.5</td>
<td>343.8</td>
<td>23.4</td>
<td>128.9</td>
</tr>
</tbody>
</table>
Figure 8 The location of homes (a) and stays (b-h) of frequent commuters living in and around Brisbane area. Colors of points in the maps of stays indicate the relative number of stays made by the frequent
commuters at each stop. Numbers in parentheses are the number of frequent commuters in each group.

We then review the distribution of stays in the seven subgroups. Figure 9 illustrates the average differential similarity measures between the frequent commuters in each group. Groups G11, G12, G14, G15, and G17 have single peaks at $h=288.4, 416.9, 426.6, 478.6,$ and 631.0, respectively. These relatively smaller values imply that the distributions of stays are similar as a whole as discussed in the Section 2, which can also be confirmed by the relatively smaller $\gamma$ and $\rho$ values. On the other hand, large $\gamma$ and $\rho$ values are observed in G13 and G16; the spatial distribution of stays for these two subgroups are shown in Figure 10 (a and b). G13 has a single peak at $h=741.3$, indicating a great variation in the distribution of stays in the group shown in Figure 10a. Group G16 has two peaks (Figure 9), one at $h=275.4$ and the other at $h=776.2$. Multiple peaks suggest that the point distributions within this group are partially similar as seen in Section 2 ($S'_{17}-S'_{19}$ in Figure 5). Figure 10c shows a zoom-in view of the locations of stays of four commuters within the dotted elliptic area in Figure 10b at a large scale, illustrating such partial similarity at different locations. All four commuters made stays at the center of the elliptic area, while each commuter made its own stays separately in other locations. The distribution of stays are partially similar at the center, which yields two peaks (Figure 9, G16). Such partial similarity between point distributions cannot be easily detected by visual analysis, confirming the effectiveness of differential similarity measure.

![Figure 9](image-url)  
Figure 9 Average differential similarity measures between frequent commuters in each group.
3. Concluding discussion

This paper presents a new method of comparing and classifying point distributions. The proposed measures quantify the degree and scale that point distributions may be similar to or different from each other. The similarity characteristic $S_{ik}$ is effective in comparing similarity at a given scale.
between different distributions, and can quantify roughly the distance between points of different distributions. The relative similarity characteristic $T_k$ enables the comparison of similarity between different scales within the same distributions, and also the evaluation of the scale at which the similarity is predominant. In addition, these measures aid in detecting partial similarity between distributions, as seen in the numerical experiments. Using the $s_k(x, h)$ map, we can easily find the location and degree of such partial similarity amongst point distributions. To test the validity of the proposed method, we performed numerical experiments and analyzed the journey data of the public transport users in Southeast Queensland, Australia. The results support the technical soundness of the method in both data comparison and the classification of point distributions. The application to real data also provided meaningful insights on the travel patterns of public transport commuters in Southeast Queensland that may not be easily observable by visual observation or other existing methods.

The proposed approach is not without limitations. First, the paper does not discuss the statistical significance of the similarity between point distributions. Statistical test requires the probability distributions of similarity measures where points are randomly distributed. Given certain $h$ values, we can derive the probability distributions in a reasonable time by performing Monte Carlo simulation. This provides the threshold values of similarity measures at a certain significance level. Interpolation of the threshold values permit us to approximate them at any scale, and consequently, to test the significance of similarity measures. Probability distributions in analytical form, however, are more convenient and efficient than Monte Carlo simulation. Analytical representation is indispensable for statistical test when treating numerous distributions of points. The explicit incorporation of a statistical perspective to the proposed method should be considered in subsequent studies.

Multiscale classification of point distributions is another important topic. The computational complexity is unacceptable when point distributions are classified independently at each scale as mentioned in Section 2. If a faster clustering algorithm is developed, multiscale classification may run in a reasonable time. Otherwise, an efficient method is necessary that performs classification and scaling simultaneously. Multiscale clustering algorithms developed in image processing may be a clue to multiscale classification of point distributions (Kushnir et al. (2006); Poole et al. (2017)).

The choice of parameter $\alpha$ needs further consideration. Though a large value is clearly necessary for the evaluation of overall similarity between point distributions, it is still unknown how large $\alpha$ should be. Theoretical approach in general statistics seems difficult to derive a unique value of $\alpha$. One promising option in search of a suitable $\alpha$ value might be to apply the proposed method by varying the $\alpha$ value in a wide variety of situations. This will enable further investigation on the relationship between $\alpha$ values and the effectiveness of the results obtained.

Future research should also consider incorporating the temporal dimension in the analysis of the point distributions. An extension of the spatial dimension from two to three is rather straightforward since it is possible to redefine the kernel function in the three-dimensional space. The addition of the
temporal dimension, on the other hand, requires the comparison of two different types of dimensions, i.e., the spatial and temporal dimensions, in the evaluation of similarity between point distributions. This is an important extension since spatial data are often accompanied with temporal information, as seen in the empirical study in Section 3. The proposed method should be extended to treat temporal dimension appropriately.

References
Diggle PJ (1983) Statistical analysis of spatial point patterns. Academic press,
Quattrochi DA, Goodchild MF (1997) Scale in remote sensing and gis. CRC press,
Center for Spatial Information Science, The University of Tokyo,
Appendix A1

This appendix discusses in detail the similarity measure $S_{ik}(h)$ and the differential similarity measure $S'_{ik}(h)$ through numerical experiments on a one-dimensional space. The experiments employ eleven sets of point distributions $\Psi_1$-$\Psi_{11}$ on a one-dimensional space. Every set consists of multiple distributions $\{\Gamma_1, \Gamma_2, ..., \Gamma_N\}$, in which we compare distribution $\Gamma_1$ with other distributions in the same set. We move every point in $\Gamma_1$ by distance $\Delta$ to generate other distributions in sets from $\Psi_1$ to $\Psi_9$, where $\Delta$ may vary across points in each distribution.

Let $h_{\text{max}}$ and $h'_{\text{max}}$ be the values of $h$ that gives the highest and the second highest peaks of $S'_{ik}(h)$, respectively. This appendix omits $(h)$ in $S_{ik}(h)$ and $S'_{ik}(h)$ for simplicity.

Set $\Psi_1$ consists of point distributions each of which is obtained by moving $\Gamma_1$ to the right by a constant distance $\Delta$. The distance $\Delta$ varies from 0.01 to 0.05 between distributions as shown in Figure A1a. $S_{ik}$ gradually moves from left to right with $\Delta$, implying that the similarity decreases from $\Gamma_2$ to $\Gamma_6$. The value of $h_{\text{max}}$ seems almost proportional to $\Delta$. It is close the half of $\Delta$, which implies that $h_{\text{max}}$ roughly tells us $\Delta$ when $\Delta$ is constant within the distribution. Appendix A3 gives a theoretical support of this relationship between $h_{\text{max}}$ and $\Delta$. 


Upton G, Fingleton B (1985) *Spatial data analysis by example. Volume 1: Point pattern and quantitative data*. John Wiley & Sons Ltd.,


Similar to the point distributions in $\Psi_1$, those in $\Psi_2$ are obtained by moving the points in $\Gamma_1$ by distance $\Delta$ varying from 0.01 to 0.05 between distributions. The direction of movement changes in turn from right to left in each distribution as seen in Figure A2a. Measures $S_{1k}$ and $S'_{1k}$ in set $\Psi_2$ are very similar.
to those in set $\Psi_1$; $S'_1$ has a single peak whose $h_{\text{max}}$ is again almost the half of $\Delta$. Our earlier presumption that $h_{\text{max}}$ roughly tells us $\Delta$ still holds when points move in different directions. A difference lies in the absence of the second highest peak in $\Psi_2$. This seems relevant to the distance between points in $\Gamma_2$ and their neighboring but not nearest points in $\Gamma_1$. 
Figure A2 Measures $S_{1k}$ and $S'_{1k}$ in set $\Psi_2$.

Set $\Psi_3$ consists of point distributions that are partially similar to $\Gamma_1$. Four points are obtained by moving four points on the left in $\Gamma_1$ to the right by distance 0.01, while the other four are located at
distance 0.05 from their nearest points in \( \Gamma_1 \). The former yields a partial similarity between \( \Gamma_1 \) and \( \Gamma_2-\Gamma_6 \), which emerges as a gradual increase of \( S_{16} \) in Figure A2b. The highest peaks of \( S_{16} \) are lower than those in Figure A1b, which is due to the difference between \( \Gamma_1 \) and \( \Gamma_2-\Gamma_6 \) as a whole. The form of \( S_{16} \) in Figure A2c is generally similar to \( S_{16} \) in Figure A1c. The value of \( h_{\max} \) almost increases by 0.05 from 0.05 to 0.15 in \( S_{12}-S_{16} \), which is the same as observed in Set \( \Psi_1 \). A clear difference lies in the small peak of \( S_{12} \) at \( h=0.0389 \). This reflects the difference between \( \Gamma_1 \) and \( \Gamma_2 \) caused by the four points on the right.
Figure A3 Measures $S_{1k}$ and $S'_{1k}$ in set $\Psi_3$.

Similar to Set $\Psi_3$, Set $\Psi_4$ contains point distributions that are partially similar to $\Gamma_1$. Two points on the left are distributed similarly with the distributions in $\Psi_3$, while the others are located at distance 0.05 from their nearest points in $\Gamma_1$. Distributions $\Gamma_2$-$\Gamma_6$ are less similar to $\Gamma_1$ than $\Gamma_2$-$\Gamma_6$ in Set $\Psi_3$. This
difference changes the form of \( S_{16} \) rather drastically. Measures from \( S'_{12} \) to \( S'_{14} \) have two peaks, while \( S'_{15} \) and \( S'_{16} \) have only a single peak. Though the highest peaks of \( S'_{12} \) to \( S'_{14} \) observed in Figure A3 still stay at \( h=0.05, 0.10, \) and \( 0.15 \), that of \( S'_{14} \) is the second highest peak in Figure A4. The highest peaks of \( S'_{14} \) to \( S'_{16} \) are now found around \( h=0.04 \). This reflects that distributions \( \Gamma_2-\Gamma_6 \) are only partially similar to \( \Gamma_1 \).

Figure A4 Measures \( S_{1k} \) and \( S'_{1k} \) in set \( \Psi_4 \).
Set $\Psi_5$ moves points in $\Gamma_1$ by two different distances $\Delta_1=0.01$ and $\Delta_2=0.05$ within each distribution. For instance, we obtain $\Gamma_2$ by moving all the points in $\Gamma_1$ by distance $\Delta_1$, while we obtain $\Gamma_3$ by moving six points by $\Delta_1$ and two points by $\Delta_2$ as seen in Figure A5. Measures from $S_{12}$ to $S_{16}$ in Figure A5b looks similar with those in Figure A1b. On the other hand, $S'_{12}$ to $S'_{16}$ are different, i.e., they are lower and measures $S'_{14}$ to $S'_{16}$ have two peaks. Peaks at $h=0.05$ gradually become lower while those from $h=0.020$ to $h=0.025$ appear and grow higher. The latter of $S'_{16}$ finally becomes the highest peak. These values are the half of the distance between a point in $\Gamma_i$ and its nearest point in $\Gamma_1$, indicated as $\Delta_1$ and $\Delta_2$ in Figure A5a. Point distribution becomes less similar with $\Gamma_1$ from $\Gamma_2$ to $\Gamma_6$. 
Figure A5 Measures $S_{ik}$ and $S'_{ik}$ in set $\Psi_5$.

Set $\Psi_6$ also moves points in $\Gamma_1$ by two different distances $\Delta_1$ and $\Delta_2$ within each distribution.
For instance, \((\Delta_1, \Delta_2) = (0.01, 0.02)\) in \(\Gamma_2\) and \((\Delta_1, \Delta_2) = (0.01, 0.03)\) in \(\Gamma_3\) as shown in Figure A6a. We keep the distance values in such a way that they assure the mutual nearest relationship between the original points in \(\Gamma_1\) and those in the other distributions, i.e., the nearest point from \(P_{21}\) in \(\Gamma_1\) is \(P_{11}\), and that from \(P_{11}\) in \(\Gamma_2\) is \(P_{21}\).

We newly introduce two variables \(\Delta_{\text{ave}}\) and \(\eta\). The former is the average of \(\Delta_1\) and \(\Delta_2\), while the latter is the average distance from points in \(\Gamma_k\) and their nearest points in \(\Gamma_1\). The two variables are equal as long as the mutual nearest relationship holds. Figure A6a shows that \(h_{\text{max}}\) is between \(\Delta_1/2\) and \(\Delta_2/2\) in any case, such as \(0.0050 < 0.0068 < 0.0100\) in \(S'_{12}\), and \(0.0100 < 0.0122 < 0.0150\) in \(S'_{16}\). We also find that \(h_{\text{max}}\) is smaller than the half of \(\Delta_{\text{ave}}\) in most cases, i.e., \(h_{\text{max}}\) is closer to \(\Delta_1/2\) than \(\Delta_2/2\). \(h_{\text{max}}\) is exactly at \(\Delta_1/2\) in \(S'_{13}, S'_{14},\) and \(S'_{15}\). We may interpret these results by assuming potential peaks at both \(\Delta_1/2\) and \(\Delta_2/2\). When the similarity at \(h = \Delta_1/2\) is predominant, only the peak at \(\Delta_1/2\) emerges while that at \(\Delta_2/2\) is concealed as seen in \(S'_{13}, S'_{14},\) and \(S'_{15}\). When the similarity at \(h = \Delta_1/2\) is significant but not predominant compared with the similarity at \(h = \Delta_2/2\), the two peaks behave as a single peak and appear at \(h\) which is closer to \(\Delta_1/2\) than \(\Delta_2/2\) \((S'_{12}, S'_{16},\) and \(S'_{17}\)). When both peaks are not significant, the peaks emerge as a single peak at \(h\) which is closer to \(\Delta_2/2\) than \(\Delta_1/2\) \((S'_{18})\).
Figure A6 Measures $S_{1k}$ and $S'_{1k}$ in set $\Psi_6$.

The next three sets of distributions treat more general cases where points in $\Gamma_1$ are moved by a variety of distances within each distribution. Distance $\Delta$ is distributed stochastically according to a uniform distribution in each set. The direction of movement is determined randomly.
Distance \( \Delta \) follows the uniform distribution between 0.00 and 0.04 in Set \( \Psi_7 \). This assures the mutual nearest relationship between the points in \( \Gamma_1 \) and those obtained by the random movement. Figure A7 shows the obtained distributions and their measures. Many peaks exist around \( h=0.0100 \), which is about the half of \( \Delta_{\text{ave}} \) in most cases. When some points in \( \Gamma_k \) are located very closely to points in \( \Gamma_1 \), \( S_{1k} \) may have multiple peaks one of which is observed at a small \( h \). Red point in \( \Gamma_7 \) causes the highest peak of \( S_{17} \) at \( h=0.0007 \), while \( S_{17} \) has the highest peak at \( h=0.0020 \) due to the three red points in \( \Gamma_2 \). Only \( S_{17} \) has two peaks out of eight measures. Our potential peak assumption interprets this as the fusion of multiple peaks at various \( h \) in the absence of a predominant peak at small \( h \).
Figure A7 Measures $S_{1k}$ and $S'_{1k}$ in set $\Psi_7$.

Set $\Psi_8$ consists of points distributed in a narrower range, i.e., distance $\Delta$ follows the uniform distribution between 0.01 and 0.03. The mutual nearest relationship between the original points in $\Gamma_1$ and others still holds. This results in that the variation of $\Delta_{\text{ave}}$ is smaller than that in Set $\Psi_7$ as shown in Figure A8a. Measure $S'_{1k}$ has only a single peak around $h=0.01$ in all the distributions.
more similar to those of $S_{13}$ and $S'_{13}$ in Set $\Psi_1$, where $\Delta=0.02$ for all the points (Figure A1).

Set $\Psi_9$ consists of point distributions where distance $\Delta$ follows the uniform distribution between 0.04 and 0.08. The mutual nearest relationship may fail in this case, i.e., points in $\Gamma_k$ can be closer to point
in $\Gamma_1$ other than those at their original locations. Consequently, $\Delta_{\text{ave}}$ is not always equal to $\eta$ as shown in Figure A9a. Measures $S_{11}$ to $S_{16}$ and $S_{18}$ have a single peak, while $S_{17}$ has two peaks at $h=0.0120$ and 0.0240. The peak at $h=0.0120$ is due to a red point in $\Gamma_7$ that is very close to a point in $\Gamma_1$. Many highest peaks are distributed around $h=0.0250$, which is closer to the half of $\eta$ rather than that of $\Delta_{\text{ave}}$. This suggests that $h_{\text{max}}$ reflects the distance between points in $\Gamma_i$ and their nearest points in $\Gamma_1$. 
Figure A9 Measures $S_{1k}$ and $S'_{1k}$ in set $\Psi_9$. 

- **Measurements Table:**
  
<table>
<thead>
<tr>
<th>$\Gamma_1$</th>
<th>$\Gamma_2$</th>
<th>$\Gamma_3$</th>
<th>$\Gamma_4$</th>
<th>$\Gamma_5$</th>
<th>$\Gamma_6$</th>
<th>$\Gamma_7$</th>
<th>$\Gamma_8$</th>
<th>$\Gamma_9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_{\text{max}}$</td>
<td>$\Delta_{\text{ave}}$</td>
<td>$\eta$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0240</td>
<td>0.0551</td>
<td>0.0504</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0270</td>
<td>0.0660</td>
<td>0.0542</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0210</td>
<td>0.0582</td>
<td>0.0404</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0220</td>
<td>0.0614</td>
<td>0.0436</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0270</td>
<td>0.0569</td>
<td>0.0548</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>0.0120</strong></td>
<td><strong>0.0724</strong></td>
<td><strong>0.0475</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0250</td>
<td>0.0602</td>
<td>0.0492</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0250</td>
<td>0.0595</td>
<td>0.0481</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Graphs:**
  
  (a) Graph showing the distribution of $S_{1k}$ and $S'_{1k}$.
  
  (b) Graph showing the cumulative distribution of $S_{1k}$.
  
  (c) Graph showing the cumulative distribution of $S'_{1k}$.
Set $\Psi_{10}$ consists of point distributions in each of which points are clustered at intervals of 0.01. Distribution $\Gamma_k$ is obtained by moving points in $\Gamma_1$ to the right by $(k-1) \times 0.1$. Measure $S_{1k}$ gradually shifts from left to right with an increase in the distance between point clusters in $\Gamma_k$ and $\Gamma_1$. Measure $S'_{1k}$ becomes lower and its peak moves to the right. The value of $h_{\text{max}}$ seems almost proportional to the distance between point clusters in $\Gamma_k$ and $\Gamma_1$. This fails in $\Gamma_8$ and $\Gamma_9$ probably because of the boundary effect, i.e., calculation is performed only in the limited range as shown in the figure.
Figure A10 Measures $S_{1k}$ and $S'_{1k}$ in set $\Psi_{10}$. 
Set $\Psi_{11}$ analyzes the relationship between the uniformity of points and the similarity measures. Figure A11a shows distributions from $\Gamma_1$ to $\Gamma_6$. Distribution $\Gamma_1$ consists of seventeen points while distributions from $\Gamma_2$ to $\Gamma_6$ consist of sixteen points. Points in $\Gamma_1$ are distributed uniformly at intervals 0.1. Each point in $\Gamma_2$ is located at the center of neighboring points in $\Gamma_1$. Points become gradually clustered from $\Gamma_2$ to $\Gamma_6$, reducing the uniformity of points.

Similarity to $\Gamma_1$ decreases monotonically from $\Gamma_2$ to $\Gamma_6$ as shown in the figure. Measure $S'_{14}$ has its peak around $h=0.025$ for all the distributions from $\Gamma_2$ to $\Gamma_6$, which is the half of the distance between a point in $\Gamma_k$ and its nearest point in $\Gamma_1$. The peak becomes lower with the reduction of uniformity. Measures $S'_{14}, S'_{15}$ and $S'_{16}$ have another lower peaks at $h=0.0780, 0.0696$ and $0.0692$, respectively.
Sets $\Psi_{12}$-$\Psi_{14}$ consist of points distributed stochastically according to the uniform distribution. They, in a sense, represent very general cases. Figure A12-Figure A14 are similar to Figure A7 in that $S_{1k}$ gradually increases and $S'_{1k}$ sometimes shows several peaks.
Figure A12 Measures $S_{1k}$ and $S'_{1k}$ in set $\Psi_{12}$. 
Figure A13 Measures $S_{1k}$ and $S'_{1k}$ in set $\Psi_{13}$. 
So far we have discussed point distributions each of which basically consists of eight points. Sets $\Psi_{15}$-$\Psi_{17}$, on the other hand, treat the cases where 30 points are distributed stochastically according to the uniform distribution. Comparing Figure A15-Figure A17 with Figure A12-Figure A14, we notice...
that $S_{1k}$ is more similar with each other in each figure and $S'_{1k}$ has fewer peaks. The former is due to the reduction of the difference between point distributions by an increase of points, while the latter occurs because regularity decreases that generates multiple peaks.

Figure A15 Measures $S_{1k}$ and $S'_{1k}$ in set $\Psi_{15}$. 

\begin{figure}
\centering
(a) 
\includegraphics[width=\textwidth]{figure_a15a.png}

(b) 
\includegraphics[width=\textwidth]{figure_a15b.png}

(c) 
\includegraphics[width=\textwidth]{figure_a15c.png}

\caption{Figure A15 Measures $S_{1k}$ and $S'_{1k}$ in set $\Psi_{15}$.}
\end{figure}
Figure A16 Measures $S_{1k}$ and $S'_{1k}$ in set $\Psi_{16}$. 
Appendix A2

This appendix tests the performance of the proposed method through numerical experiments on a two-dimensional space. The experiments employ three sets of point distributions $\Psi_1$-$\Psi_3$ in a square.
region $R$ of sides 100. Every set consists of multiple distributions $\{\Gamma_1, \Gamma_2, \ldots, \Gamma_N\}$, each of which consists of 25 points. We compare distribution $\Gamma_1$ with other distributions in each set by using a program written in C++. We again omit $(h)$ of $S_{a}(h)$, $S'_{a}(h)$ and $T_{a}(h)$ in the following.

The three sets are designed to evaluate the behavior of the proposed method from different perspectives. The sets $\Psi_1$ and $\Psi_2$ consider the case when point distributions are partially similar at specific scales. The sets represent the different levels of partial similarity by the number of points and the distance between points that are closely located, respectively. The set $\Psi_3$ treats the case where point distributions are similar at various scales. The distance between points in distribution $\Gamma_1$ and their nearest points in $\Gamma_i$ varies among points.

Figure A12a indicates set $\Psi_1$ that consists of seven distributions $\{\Gamma_1, \Gamma_2, \ldots, \Gamma_7\}$. A 10 by 10 cells grid is added as a reference to illustrate the relative locations of points between distributions. $\Gamma_1$ points are shown in black color and located in the lower-left corner of region $R$ at intervals of 5.0 in all grids. We obtain $\Gamma_2$ by moving every $\Gamma_1$ point 1.0 up and 1.0 right. We then move five points of $\Gamma_2$ to the upper-right corner in turn to generate distribution $\Gamma_3$, and 5 points more each time to generate $\Gamma_4$ to $\Gamma_7$. Distribution $\Gamma_2$ is similar to $\Gamma_1$, while $\Gamma_7$ is quite different from $\Gamma_1$. Distributions $\Gamma_3$-$\Gamma_6$ are partially similar to $\Gamma_1$. 

52
Figure A18 Comparison of point distributions. (a) Point distributions from $\Gamma_1$ to $\Gamma_7$. (b) Distribution function of $\Gamma_1$. (c) Distribution function of $\Gamma_2$. (d) Similarity between distribution functions $s_{12}$. Brighter colors indicate larger values.

Figure A12b and Figure A12c show the distribution functions of $\Gamma_1$ and $\Gamma_2$, respectively, where $h=35.0$. We can confirm the similarity between $\Gamma_1$ and $\Gamma_2$ with a difference on the upper-right corner. A more effective means of comparing the distributions is to visualize the spatial distribution of their similarity weighted by the point density:
\[ s_{ik}(x,h) = \left\{ f_i(x,h) + f_k(x,h) \right\} \left\{ 1 - \frac{1}{2} \left| f_i(x,h) - f_k(x,h) \right| \right\} . \]

Figure A12d shows \( s_{12}(x,h) \), which clearly reveals the location where \( \Gamma_1 \) and \( \Gamma_2 \) are similar.

Figure A13 shows the similarity measures \( S_{1k} \), \( S'_{1k} \), and \( T_{1k} \). The measure \( S_{1k} \) increases rapidly twice in most cases in Figure A13a, which emerges as higher peaks of \( S'_{1k} \) at \( h_1=0.5 \) and lower ones at \( h_2=36.0-37.0 \) in Figure A13b, though the latter are unrecognizably low. The peaks of \( S'_{12} \) and \( S'_{13} \) at \( h_1=0.5 \) are higher than those of \( S'_{16} \) and \( S'_{17} \), which implies that \( \Gamma_2 \) and \( \Gamma_3 \) are more similar to \( \Gamma_1 \) at fine scales than \( \Gamma_6 \) and \( \Gamma_7 \). Figure A13c shows the measure \( T_{1k} \) that indicates the relative degree of similarity between different scales. For instance, the peak of \( T_{13} \) at a small \( h \) is higher than that of a large \( h \). This implies that the similarity observed at fine scales is dominant compared with that of coarse scales. This occurs because \( \Gamma_3 \) is fairly similar to \( \Gamma_1 \), i.e., 80% of \( \Gamma_3 \) points are closely located to \( \Gamma_1 \) points. On the other hand, the peak of \( T_{16} \) at a small \( h \) is lower than that of a large \( h \). The similarity is primarily observed at coarse scales because \( \Gamma_6 \) is only partially similar to \( \Gamma_1 \).

Similarity at a fine scale is represented as the scale that gives the highest peak in \( T_{1k} \). If the scale is small as seen in \( T_{12} \) and \( T_{13} \), we say that the distributions are similar to a considerable extent. The degree of similarity is represented as the height of the peak. If \( T_{1k} \) has the highest peak at a large \( h \), the distributions are totally different. The distributions, however, are partially similar if \( T_{1k} \) has another peak at a small \( h \). Examples include \( \Gamma_5 \) and \( \Gamma_6 \) that are only partially similar to \( \Gamma_1 \).

The value of \( h \) that gives a local maximum in \( S'_{ik} \) represents the distance between points of different distributions. Let \( \Delta_1(=1.0) \) be the horizontal distance between a point of \( \Gamma_1 \) in the lower-left corner and its nearest point of \( \Gamma_1 \). It is approximated by

\[ \Delta_1 \approx 2h_1. \]  

(14)

Similarly, the horizontal distance between the point cluster in the lower-left corner and that in the upper-right corner, which is denoted by \( \Delta_2(\approx 71.5-73.5) \), is approximately given by

\[ \Delta_2 \approx 2h_2. \]  

(15)

Sadahiro and Liu (2016) gives a proof of the above approximations.
Figure A19 The relationship between $h$ and similarity measures in set $\Psi_\ell$. (a) Similarity function $S_{ik}$. (b) Similarity characteristic $S'_{ik}$. (c) Relative similarity characteristic $T'_{ik}$. 
Figure A14a shows set $\Psi_2$ that consists of six distributions $\Gamma_1$-$\Gamma_6$. Points of $\Gamma_1$ are all arranged regularly at intervals of 20.0. Distributions from $\Gamma_2$ to $\Gamma_6$ contain twenty points that are also regularly arranged at distance $\Delta_1=5.0$ in both horizontal and vertical directions from points in $\Gamma_1$. Five points on the left in $\Gamma_2$-$\Gamma_6$ gradually become separated from their nearest points in $\Gamma_1$. The distance $\Delta_2$ increases linearly from 1.0 to 5.0. The five points represent the different levels of partial similarity between $\Gamma_1$ and $\Gamma_2$-$\Gamma_6$.

The measure $S_{1k}$ shown in Figure A14b increases rapidly at a small $h$, which is represented by the peaks of $S_{1k}$ at $h=2.5$ in Figure A14c. Their correspondent peaks of $T_{1k}$ in Figure A14d indicate the similarity at a fine scale between $\Gamma_1$ and $\Gamma_2$-$\Gamma_6$. The measure $S_{12}$ has another peak at $h=0.5$ in Figure A14c. The two peaks of $S_{12}$ confirm approximations (13) and (14), where $\Delta_1=5.0$ and $\Delta_2=1.0$. Measures $S_{13}$-$S_{16}$ do not have peaks at small $h$ as expected by the approximations. This is because $S_{1k}$ distinguishes the distance $\Delta_1=5.0$ from $\Delta_2=1.0$ while regards $\Delta_1=5.0$ and $\Delta_2=2.0$ as very close. The measure presents two peaks at different scales in the former case while summarizes the partial similarity at different scales as a single peak in the latter.
Figure A20 The relationship between $h$ and similarity measures in set $\Psi_2$. (a) Point distributions from $\Gamma_1$ to $\Gamma_6$. (b) Similarity function $S_{ik}$. (c) Similarity characteristic $S'_{ik}$. (d) Relative similarity characteristic $T'_{ik}$.

In the above two sets of point distributions, the points of $\Gamma_k$ are located at certain distances from...
those of $\Gamma_1$ so that the points are arranged rather regularly. Set $\Psi_3$, on the other hand, considers the cases where the distance varies among points. Twenty-one points of $\Gamma_1$ are located in the upper-left corner while twenty-one points of $\Gamma_2-\Gamma_5$ are located in the lower-right corner (Figure A15a). Four of $\Gamma_1$ points form a square of side 10.0 in the lower-left corner. Four of $\Gamma_2-\Gamma_5$ points are distributed around the four of $\Gamma_1$ points in the lower-left corner. Each of $\Gamma_2-\Gamma_5$ points is separated by a different distance from that of $\Gamma_1$ points.

Distributions $\Gamma_2-\Gamma_5$ are partially similar to $\Gamma_1$ in that they all include four points in the lower-left corner. The four points are distributed stochastically according to the uniform distribution in the circles centered at each $\Gamma_1$ point in the lower-left corner. The radius of the circles ranges from 1.0 ($\Gamma_2$) to 4.0 ($\Gamma_5$), which represents the different levels of similarity. We have tried 100 different patterns of these points and present five typical results in each case from $\Gamma_2$ to $\Gamma_5$ in Figure A15. The measure $S_{ik}$ shown in Figure A15b increases very slowly, which is represented as the highest peak of $T_{ik}$ at a large $h$ in Figure A15d. This reflects the difference between $\Gamma_1$ and $\Gamma_2-\Gamma_5$ at a fine scale. The partial similarity between $\Gamma_1$ and $\Gamma_2-\Gamma_5$ in the lower-left corner is represented as peaks at small $h$ in Figure A15c and Figure A15d. Though the distance between $\Gamma_1$ points and their nearest $\Gamma_1$ points varies in each distribution, $S'_{ik}$ and $T_{ik}$ have only a single peak at a small $h$ in 17 out of 20 (=$5\times4$) cases. The measures again summarize the similarity at different scales as fewer peaks.
Figure A21 The relationship between $h$ and similarity measures in set $\Psi$. (a) Point distributions from $\Gamma_1$ to $\Gamma_5$. (b) Similarity function $S_{ik}$. (c) Similarity characteristic $S'_{ik}$. (d) Relative similarity characteristic $T'_{ik}$.

The three measures proposed above quantifies the degree, resolution, and hence the scale that the distributions are similar with each other. The similarity characteristic $S'_{ik}$ is effective in comparing the similarity at a given scale between different distributions, and also quantify roughly the distance between

59
points of different distributions. The relative similarity characteristic $T_{ik}$ enables comparison of similarity between different scales within the same distributions, and evaluation of the scale at which the similarity is predominant. The measures also help us to detect partial similarity between distributions as seen in the numerical experiments. With the help of $s_{ik}(x, h)$ map, we can easily find the location and degree of such partial similarity.

Appendix A3

This appendix discusses in detail the structure of measure $S_{ik}(h)$ with a focus is on the highest peaks at small $h$, which are often observed in Appendices A1 and A2. Suppose point distributions $\Gamma_P$ and $\Gamma_Q$, each of which consists of $n$ points that are located at intervals $2w$ on a one dimensional space. The points are distributed almost infinitely, i.e., $n$ is extremely large. The $j$th points in distributions $\Gamma_P$ and $\Gamma_Q$ are denoted by $P_j$ and $Q_j$, respectively. The distance between a point in $\Gamma_1$ and its nearest points in $\Gamma_2$ is $\varepsilon$. The locations of $P_j$ and $Q_j$ are represented as $2jw - \varepsilon/2$ and $2jw + \varepsilon/2$, respectively, as shown in Figure A22.

![Figure A22 Distributions $\Gamma_P$ and $\Gamma_Q$, and their interpreted surfaces.](image)

The interpreted surfaces of $\Gamma_P$ and $\Gamma_Q$ in their standardized form are represented as

$$f_P(x, h) = \frac{1}{\sqrt{2\pi nh}} \sum_{k=1}^{n} e^{-\frac{[x - (2hw - \varepsilon/2)]^2}{2h'}}$$

and

$$f_Q(x, h) = \frac{1}{\sqrt{2\pi nh}} \sum_{k=1}^{n} e^{-\frac{[x - (2hw + \varepsilon/2)]^2}{2h'}}$$

respectively. The similarity measure between $\Gamma_P$ and $\Gamma_Q$ is
\[
S_{PQ}(h) = 1 - \frac{1}{2} \int_{x=-\infty}^{\infty} \left| f_P(x, h) - f_Q(x, h) \right| \, dx \\
= 1 - \frac{1}{2\sqrt{2\pi} nh} \int_{x=-\infty}^{\infty} \sum_{k=1}^{\infty} \left[ e^{\frac{\left(x - \left(\frac{2kw + \frac{\epsilon}{2}\right)^2}{2h^2}\right)}{2h^2}} - e^{\frac{\left(x - \left(\frac{2kw + \frac{\epsilon}{2}\right)^2}{2h^2}\right)}{2h^2}} \right] \, dx.
\]

(18)

The measure is represented as the area of light shades in Figure A22. Since \( n \) is very large, it is enough to consider only the dark-shaded area in the figure. The above equation then becomes

\[
S_{PQ}(h) = 1 - \frac{2n}{2\sqrt{2\pi} nh} \int_{x=2hw}^{\infty} \sum_{k=1}^{\infty} \left[ e^{\frac{\left(x - \left(\frac{2kw + \frac{\epsilon}{2}\right)^2}{2h^2}\right)}{2h^2}} - e^{\frac{\left(x - \left(\frac{2kw + \frac{\epsilon}{2}\right)^2}{2h^2}\right)}{2h^2}} \right] \, dx \\
= 1 - \frac{1}{\sqrt{2\pi} h} \sum_{k=1}^{\infty} \int_{x=2hw}^{\infty} \left[ e^{\frac{\left(x - \left(\frac{2kw + \frac{\epsilon}{2}\right)^2}{2h^2}\right)}{2h^2}} - e^{\frac{\left(x - \left(\frac{2kw + \frac{\epsilon}{2}\right)^2}{2h^2}\right)}{2h^2}} \right] \, dx \\
= 1 - \frac{1}{\sqrt{2\pi} h} \sum_{k=1}^{\infty} \int_{x=0}^{\infty} \left[ e^{\frac{\left(x + 2hw - \frac{k\epsilon}{2}\right)^2}{2h^2}} - e^{\frac{\left(x + 2hw - \frac{k\epsilon}{2}\right)^2}{2h^2}} \right] \, dx \\
= 1 - \frac{1}{\sqrt{2\pi} h} \sum_{k=1}^{\infty} \int_{x=0}^{\infty} \left[ e^{\frac{\left(x + 2hw - \frac{k\epsilon}{2}\right)^2}{2\sqrt{2}h}} - e^{\frac{\left(x + 2hw - \frac{k\epsilon}{2}\right)^2}{2\sqrt{2}h}} \right] \, dx.
\]

(19)

To evaluate the integral term in the above equation, we employ an approximation of error function:

\[
\int_{x=0}^{\infty} e^{-x^2} \, dx = \frac{\sqrt{\pi}}{2} \left\{ 1 + e^{-w^2} \left( -\frac{1}{w} + \frac{1}{2w^3} - \frac{3}{4w^5} + \frac{15}{8w^7} + \text{L} \right) \right\} \\
\approx \frac{\sqrt{\pi}}{2} \left( 1 - \frac{e^{-w^2}}{\sqrt{\pi} w} \right).
\]

(20)

This approximation requires integration by substitution as follows.
\[ t = \frac{x + 2jw - 2kw - \frac{E}{2}}{\sqrt{2h}} \]  

\[ dx = \sqrt{2h} dt \]  

\[ \int_{x=0}^{w} e^{\left(\frac{x + 2jw - 2kw - \frac{E}{2}}{\sqrt{2h}}\right)^{2}} dx = \sqrt{2h} \int_{t=0}^{\frac{w+2jw - 2kw - \frac{E}{2}}{\sqrt{2h}}} e^{-t^{2}} dt \]

\[ = \sqrt{2h} \left( \int_{t=0}^{\frac{w+2jw - 2kw + \frac{E}{2}}{\sqrt{2h}}} e^{-t^{2}} dt - \int_{t=0}^{\frac{w+2jw - 2kw - \frac{E}{2}}{\sqrt{2h}}} e^{-t^{2}} dt \right) \]  

\[ \int_{x=0}^{w} e^{\left(\frac{x + 2jw - 2kw + \frac{E}{2}}{\sqrt{2h}}\right)^{2}} dx = \sqrt{2h} \int_{t=0}^{\frac{w+2jw - 2kw + \frac{E}{2}}{\sqrt{2h}}} e^{-t^{2}} dt \]

\[ = \sqrt{2h} \left( \int_{t=0}^{\frac{w+2jw - 2kw - \frac{E}{2}}{\sqrt{2h}}} e^{-t^{2}} dt - \int_{t=0}^{\frac{w+2jw + \frac{E}{2}}{\sqrt{2h}}} e^{-t^{2}} dt \right) \]  

Equation (18) now becomes

\[ S_{PQ}(h) = 1 - \frac{\sqrt{2h}}{\sqrt{2\pi h}} \sum_{k=1}^{\infty} \left( \int_{t=0}^{\frac{w+2jw - 2kw - \frac{E}{2}}{\sqrt{2h}}} e^{-t^{2}} dt - \int_{t=0}^{\frac{w+2jw - 2kw + \frac{E}{2}}{\sqrt{2h}}} e^{-t^{2}} dt \right) \]

\[ = 1 - \frac{1}{\sqrt{\pi}} \sum_{k=1}^{\infty} \left( \int_{t=0}^{\frac{(2j-2k+1)w - \frac{E}{2}}{\sqrt{2h}}} e^{-t^{2}} dt - \int_{t=0}^{\frac{(2j-2k)w - \frac{E}{2}}{\sqrt{2h}}} e^{-t^{2}} dt \right) \]

Using Approximation (19), we can rewrite each term inside the summation above as
\[
\int_{t=0}^{(2j-2k+1)w+\frac{\varepsilon}{2}} e^{-t^2} dt \approx \frac{\sqrt{\pi}}{2} \left( 1 - e^{-\frac{(2j-2k+1)w+\varepsilon}{2}} \right)
\]
\[
\frac{h}{\sqrt{2} (2j-2k+1) w + \frac{\varepsilon}{2}} e^{\frac{(2j-2k+1)w+\varepsilon}{2}} \]
\[
\tau_k^1 = \frac{1}{(2j-2k+1)w - \frac{\epsilon}{2}} e^{\frac{(2j-2k+1)w - \frac{\epsilon}{2}}{2h^2}}
\]
\[
\tau_k^2 = \frac{1}{(2j-2k)w - \frac{\epsilon}{2}} e^{\frac{(2j-2k)w - \frac{\epsilon}{2}}{2h^2}}
\]
\[
\tau_k^3 = \frac{1}{(2j-2k+1)w + \frac{\epsilon}{2}} e^{\frac{(2j-2k+1)w + \frac{\epsilon}{2}}{2h^2}}
\]
\[
\tau_k^4 = \frac{1}{(2j-2k)w + \frac{\epsilon}{2}} e^{\frac{(2j-2k)w + \frac{\epsilon}{2}}{2h^2}}
\]

and
\[
\tau_k = \tau_k^1 - \tau_k^2 - \tau_k^3 + \tau_k^4
\]

Equation (29) becomes
\[
S_{PQ}(h) \approx 1 - \frac{h}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \tau_k
\]

We evaluate \(\tau_k^1, \tau_k^2, \tau_k^3,\) and \(\tau_k^4\) from \(k=j-2\) to \(j+2\) as follows:
The above equations indicate

\[
\tau^1_{j+k'} = \frac{e^{-\frac{w-e^2}{2h^2}}}{1-e^{-1+\varepsilon}} = \frac{e^{-\frac{w+e^2}{2h^2}}}{1-e^{-1+\varepsilon}},
\]

\[= \tau^3_{j+k'-1}
\]

and similarly,

\[
\tau^3_{j+k'} = \tau^1_{j+k'-1}
\]
\[
\tau^2_{j+k'} = \tau^4_{j+k'-2}
\]
\[
\tau^4_{j+k'} = \tau^2_{j+k'-2}
\]

Using the above equations, we rewrite Equation (32) as

\[
S_{pq}(h) \approx 1 - \frac{h}{\sqrt{2\pi}w} \sum_{k=1}^{\infty} \tau_k
\]
\[
= 1 - \frac{h}{\sqrt{2\pi}} \left\{ \frac{2}{e} \frac{e^{\frac{e^2}{8h^2}} - \tau^2_j + \tau^4_j + \sum_{k=2j+1}^{\infty} \tau_k}{\sqrt{2h^2}} \right\}
\]
Since variables $\tau^2_k$, $\tau^4_k$, and $\tau_k$ rapidly decrease with an increase of $k$, we can approximate the above equation as

$$S_{PQ}(h) \approx 1 - \frac{2\sqrt{2}h}{\sqrt{\pi \varepsilon}} e^{-\frac{\varepsilon^2}{8h^2}}$$

(37)

Differentiation of $S_{PQ}(h)$ with respect to $h$ yields $S'(h)$:

$$S'_{PQ}(h) = -\frac{2\sqrt{2}}{\sqrt{\pi \varepsilon h^2}} \left( h^2 + \frac{\varepsilon^2}{4} \right) e^{-\frac{\varepsilon^2}{8h^2}}$$

(38)

Measure $S'_{PQ}(h)$ reaches its maximum when

$$S''_{PQ}(h) = -\frac{\sqrt{2} e \left( \frac{\varepsilon^2}{4} - h^2 \right)}{2h^5} e^{-\frac{\varepsilon^2}{8h^2}} = 0$$

(39)

Since $\varepsilon$, $w$, and $h$ are all positive, the above equation holds only when

$$h = \frac{\varepsilon}{2}$$

(40)

This equation indicates that $S_{PQ}(h)$ shows its maximum when $h$ is the half of the distance between points in $\Gamma_P$ and their nearest points in $\Gamma_Q$.

Extending the above discussion, we can show that the value of $h_{\text{max}}$ is proportional to the distance between points in $\Gamma_Q$ and their nearest points in $\Gamma_P$ in the case where point distributions exhibits a regular but non-uniform pattern such as seen in Set $\Psi_1$. Though we omit the details of the proof due to space limitations, the following Appendix gives its outline in a rather simple setting.

**Appendix A4**

This appendix extends the discussion in the previous appendix to a more general case. Suppose distributions $\Gamma_P$ and $\Gamma_Q$, each of which consists of $n$ points that are arranged alternately as shown in Figure A23. The distributions are in the mutual nearest relationship such as seen in set $\Psi_6$ in Appendix A1, i.e., the nearest point in $\Gamma_Q$ from $P_i$ is $Q_i$, while the nearest point in $\Gamma_P$ from $Q_i$ is $P_i$. The locations of $i$th point in $\Gamma_P$ and $j$th point in $\Gamma_Q$ are denoted by $p_i$ and $q_j$, respectively. Let $M_i$ be the middle point between $P_i$ and $Q_i$. The middle point between $Q_{i+1}$ and $P_i$ is denoted by $M'_i$, and that between $Q_i$ and $P_{i+1}$ is $M''_i$. Point $M'_i$
is equivalent to $M_{i+1}^+$, while $M_i^+$ is equivalent to $M_{i-1}^+$ as seen in Figure A23.

Let $\Phi^-$ and $\Phi^+$ be the interval bounded by $M_{i-}$ and $M_i$, and that by $M_i$ and $M_{i+}$, respectively. $\gamma(h; \Phi)$ as the difference between $\Gamma_P$ and $\Gamma_Q$ in interval $\Phi$:

$$\gamma(h; \Phi) = \frac{1}{2} \int_{x=x_L(\Phi)}^{x_U(\Phi)} \left[f_P(x, h) - f_Q(x, h)\right] dx,$$

where $x_L(\Phi)$ and $x_U(\Phi)$ are the lower and upper bound of interval $\Phi$. We can calculate $S_{PQ}(h)$ by summing up $\gamma(h; \Phi)$ for all the intervals divided by the middle points in $\Gamma_P$ and $\Gamma_Q$, and substitute the summation from 1.

Let us first focus on the calculation of $\gamma(h; \Phi^-)$. The coordinates of the lower and upper boundaries are given by

$$x_L = \frac{p_i + q_{i-1}}{2},$$

and

$$x_U = \frac{p_i + q_{i}}{2},$$

respectively.

Equations from (15) to (17) and the above simulations that the similarity between $\Gamma_P$ and $\Gamma_Q$ evaluated at location $x$ is primarily based on the points in the neighborhood of $x$. Equation (29) calculates

Figure A23 Distributions $\Gamma_P$ and $\Gamma_Q$, and their interpreted surfaces.
the summation of $\tau_1, \tau_2, \tau_3$, and $\tau_4$, each of which is a negative exponential function of $(j-k)^2 w$. Since they rapidly decrease with an increase of $(j-k)^2$, we can almost evaluate their summation from $k=1$ to infinity by the term when $j-k=0$. This implies that we can approximate $\gamma(h; \Phi(-)$ by the interpreted surfaces generated from $P_i$ and $Q_i$:

$$\gamma(h; \Phi(-) = \frac{1}{2} \int_{x=x_L}^{x_U} \left( f_p(x, h) - f_Q(x, h) \right) dx$$

$$= \frac{1}{2} \int_{x=x_L}^{x_U} \left\{ \frac{1}{\sqrt{2\pi n_h}} \sum_{i=1}^{n_p} e^{-\frac{(x-p_i)^2}{2n_h}} - \frac{1}{\sqrt{2\pi n_h}} \sum_{j=1}^{n_Q} e^{-\frac{(x-q_j)^2}{2n_h}} \right\} dx$$

$$= \frac{1}{2\sqrt{2\pi n_h}} \int_{x=x_L}^{x_U} \left\{ \sum_{i=1}^{n_p} e^{-\frac{(x-p_i)^2}{2n_h}} - \sum_{j=1}^{n_Q} e^{-\frac{(x-q_j)^2}{2n_h}} \right\} dx$$

$$\approx \frac{1}{2\sqrt{2\pi n_h}} \int_{x=x_L}^{x_U} \left\{ e^{-\frac{(x-p_i)^2}{2n_h}} - e^{-\frac{(x-q_j)^2}{2n_h}} \right\} dx$$

(45)

Using approximation (19), we obtain

$$\int_{x=x_L}^{x_U} e^{-\frac{(x-p_i)^2}{2n_h}} dx = \int_{x=0}^{x_U} e^{-\frac{(x-p_i)^2}{2n_h}} dx - \int_{x=0}^{x_L} e^{-\frac{(x-p_i)^2}{2n_h}} dx$$

$$= \frac{\sqrt{2\pi n_h}}{2} \left\{ 1 - \frac{1}{\sqrt{\pi}} \frac{x_U - p_i}{\sqrt{2n_h}} e^{-\frac{(x_U-p_i)^2}{2n_h}} \right\} - \frac{\sqrt{2\pi n_h}}{2} \left\{ 1 - \frac{1}{\sqrt{\pi}} \frac{x_L - p_i}{\sqrt{2n_h}} e^{-\frac{(x_L-p_i)^2}{2n_h}} \right\}$$

$$= h^2 \left\{ -\frac{1}{x_U - p_i} e^{-\frac{(x_U-p_i)^2}{2n_h}} + \frac{1}{x_L - p_i} e^{-\frac{(x_L-p_i)^2}{2n_h}} \right\}$$

(46)

and

$$\int_{x=x_L}^{x_U} e^{-\frac{(x-q_j)^2}{2n_h}} dx = \int_{x=0}^{x_U} e^{-\frac{(x-q_j)^2}{2n_h}} dx - \int_{x=0}^{x_L} e^{-\frac{(x-q_j)^2}{2n_h}} dx$$

$$= \frac{\sqrt{2\pi n_h}}{2} \left\{ 1 - \frac{1}{\sqrt{\pi}} \frac{x_U - q_j}{\sqrt{2n_h}} e^{-\frac{(x_U-q_j)^2}{2n_h}} \right\} - \frac{\sqrt{2\pi n_h}}{2} \left\{ 1 - \frac{1}{\sqrt{\pi}} \frac{x_L - q_j}{\sqrt{2n_h}} e^{-\frac{(x_L-q_j)^2}{2n_h}} \right\}$$

$$= h^2 \left\{ -\frac{1}{x_U - q_j} e^{-\frac{(x_U-q_j)^2}{2n_h}} + \frac{1}{x_L - q_j} e^{-\frac{(x_L-q_j)^2}{2n_h}} \right\}$$
Since \( P_i \) and \( Q_i \) are both closer to the middle point between the boundaries of \( \Phi \) than the boundaries, we can say

\[
\left( \frac{x_L - P_i}{\sqrt{2h}} \right)^2 \approx \left( \frac{x_U - P_i}{\sqrt{2h}} \right)^2
\]

(48)

and

\[
\left( \frac{x_L - q_i}{\sqrt{2h}} \right)^2 \approx \left( \frac{x_U - q_i}{\sqrt{2h}} \right)^2.
\]

(49)

This yields

\[
\int_{x=x_L}^{x_U} e^{-\frac{(x-P_i)^2}{2h}} \, dx = h^2 \left\{ -\frac{1}{x_U - P_i} e^{-\frac{(x_U - P_i)^2}{2h}} + \frac{1}{x_L - P_i} e^{-\frac{(x_L - P_i)^2}{2h}} \right\}
\]

\[
\approx -\frac{h^2}{x_U - P_i} e^{-\frac{(x_U - P_i)^2}{2h}}
\]

(50)

and

\[
\int_{x=x_L}^{x_U} e^{-\frac{(x-q_i)^2}{2h}} \, dx = h^2 \left\{ -\frac{1}{x_U - q_i} e^{-\frac{(x_U - q_i)^2}{2h}} + \frac{1}{x_L - q_i} e^{-\frac{(x_L - q_i)^2}{2h}} \right\}
\]

\[
\approx -\frac{h^2}{x_U - q_i} e^{-\frac{(x_U - q_i)^2}{2h}}
\]

(51)

Consequently,

\[
\int_{x=x_L}^{x_U} e^{-\frac{(x-P_i)^2}{2h}} \, dx - \int_{x=x_L}^{x_U} e^{-\frac{(x-q_i)^2}{2h}} \, dx \approx h^2 \left\{ -\frac{1}{x_U - P_i} e^{-\frac{(x_U - P_i)^2}{2h}} + \frac{1}{x_L - q_i} e^{-\frac{(x_L - q_i)^2}{2h}} \right\}
\]

(52)

and

\[
\gamma (h; \Phi_i -) \approx \frac{h}{2\sqrt{2\pi} n} \left[ -\frac{1}{x_U - P_i} e^{-\frac{(x_U - P_i)^2}{2h}} + \frac{1}{x_U - q_i} e^{-\frac{(x_U - q_i)^2}{2h}} \right].
\]

(53)
Since
\[ x_U - p_i = -x_U + q_i , \]

We rewrite Equation (52) as
\[
\gamma \left( h; \Phi_i - \right) \approx \frac{h}{2\sqrt{2\pi n}} \left[ \frac{1}{x_U - p_i} e^{-\left( \frac{x_U - p_i}{\sqrt{2h}} \right)^2} - \frac{1}{x_U - q_i} e^{-\left( \frac{x_U - q_i}{\sqrt{2h}} \right)^2} \right] \\
= \frac{h}{\sqrt{2\pi n (x_U - p_i)}} e^{-\left( \frac{x_U - p_i}{\sqrt{2h}} \right)^2} \\
= \frac{h}{\sqrt{2\pi n \left( \frac{p_i + q_i}{2} - p_i \right)}} e^{-\left( \frac{p_i + q_i - p_i}{\sqrt{2h}} \right)^2} \\
= \frac{h}{\sqrt{2\pi n \varepsilon_i}} e^{-\left( \frac{\varepsilon_i}{\sqrt{2h}} \right)^2}
\]

where \( \varepsilon_i \) is the distance between \( P_i \) and \( Q_i \).

We then turn to the calculation of \( \gamma(h; \Phi^+) \). The coordinates of the lower and upper boundaries are given by
\[
x_L = \frac{p_i + q_i}{2} .
\]

and
\[
x_U = \frac{p_{i+1} + q_i}{2} ,
\]

respectively.

We can employ Equations from (44) to (46) with a slight modification. Inequalities (47) and (48) now become
\[
\left( \frac{x_L - p_i}{\sqrt{2h}} \right)^2 = \left( \frac{x_U - p_i}{\sqrt{2h}} \right)^2
\]

and
$$\left( \frac{x_L - q_i}{\sqrt{2h}} \right)^2 = \left( \frac{x_L - q_i}{\sqrt{2h}} \right)^2.$$ (59)

Using the above inequalities, we obtain

$$\int_{x=x_L}^{x_u} e^{-\frac{(x-p_i)^2}{2\pi}} \, dx \approx \frac{h^2}{x_L - p_i} e^{-\left( \frac{x_L - p_i}{\sqrt{2h}} \right)^2},$$ (60)

$$\int_{x=x_L}^{x_u} e^{-\frac{(x-q_i)^2}{2\pi}} \, dx \approx \frac{h^2}{x_L - q_i} e^{-\left( \frac{x_L - q_i}{\sqrt{2h}} \right)^2},$$ (61)

and

$$\int_{x=x_L}^{x_u} e^{-\frac{(x-p_i)^2}{2\pi}} \, dx + \int_{x=x_L}^{x_u} e^{-\frac{(x-q_i)^2}{2\pi}} \, dx \approx h^2 \left\{ - \frac{1}{x_L - p_i} e^{-\left( \frac{x_L - p_i}{\sqrt{2h}} \right)^2} + \frac{1}{x_L - q_i} e^{-\left( \frac{x_L - q_i}{\sqrt{2h}} \right)^2} \right\}.$$ (62)

Consequently,

$$\gamma(h; \Phi_i +) \approx \frac{h}{2\sqrt{2\pi n}} \left[ \frac{1}{x_L - p_i} e^{-\left( \frac{x_L - p_i}{\sqrt{2h}} \right)^2} - \frac{1}{x_L - q_i} e^{-\left( \frac{x_L - q_i}{\sqrt{2h}} \right)^2} \right]$$

$$= \frac{h}{\sqrt{2\pi n (x_L - p_i)}} e^{-\left( \frac{x_L - p_i}{\sqrt{2h}} \right)^2}$$

$$= \frac{h}{\sqrt{2\pi n \left( \frac{p_i + q_i}{2} - p_i \right)}} e^{-\left( \frac{p_i + q_i}{2\sqrt{2h}} \right)^2}$$

$$= \frac{2h}{\sqrt{2\pi n \epsilon_i}} e^{-\left( \frac{q_i}{2\sqrt{2h}} \right)^2}.$$ (63)

Adding Equation (54) to (62), we obtain the difference between $\Gamma_P$ and $\Gamma_Q$ in interval $\Phi(M(i, j-1), M(i+1, j))$: 
\[ \gamma(h; \Phi_i -) + \gamma(h; \Phi_i +) = \frac{\sqrt{2h}}{2\sqrt{\pi n} \epsilon_i} e^{\left(\frac{\epsilon_i}{\sqrt{2h}}\right)^2}. \] (64)

Summing up the above equation for all the intervals, we can calculate \( S(h) \):

\[ S_{PQ} (h) = 1 - \sum_{i=1}^{n} \left\{ \gamma(h; \Phi_i -) + \gamma(h; \Phi_i +) \right\} \]
\[ = 1 - \sum_{i=1}^{n} e^{\left(\frac{\epsilon_i}{\sqrt{2h}}\right)^2} \]
\[ = 1 - \frac{2\sqrt{2h}}{\sqrt{\pi n}} \sum_{i=1}^{n} \frac{1}{\epsilon_i} e^{\left(\frac{\epsilon_i}{\sqrt{2h}}\right)^2} \] (65)

Measure \( S'_{PQ}(h) \) and \( S''_{PQ}(h) \) are

\[ S'_{PQ} (h) = -\frac{2\sqrt{2}}{\sqrt{\pi n}} \sum_{i=1}^{n} \frac{h^2 + \epsilon_i^2}{4h^2 \epsilon_i} e^{\left(\frac{\epsilon_i}{\sqrt{2h}}\right)^2} \] (66)

and

\[ S''_{PQ} (h) = \frac{2\sqrt{2}}{\sqrt{\pi n} h^5} \sum_{i=1}^{n} \epsilon_i \left(\frac{\epsilon_i^2}{h^2} - h^2\right) e^{\left(\frac{\epsilon_i}{\sqrt{2h}}\right)^2}, \] (67)

respectively.

If \( \epsilon_i \) is equal for any \( i \), \( S'_{PQ}(h) \) becomes zero when \( h=\epsilon/2 \), i.e., a peak of \( S'_{PQ}(h) \) is observed at \( h=\epsilon/2 \). This holds not only when points in \( \Gamma_P \) and \( \Gamma_Q \) are both uniformly distributed but also when \( \Gamma_P \) and \( \Gamma_Q \) show a regular but non-uniform pattern such as seen in Set \( \Psi_1 \).

Using Equations from (64) to (66), we evaluate the location of peaks of \( S_{PQ}(h) \) when \( \epsilon_i \) takes one of two different values as seen in Set \( \Psi_6 \). We assume that \( \epsilon_i=0.01 \) for half of the points while \( \epsilon_i=\epsilon \) for the other half. Solving \( S'_{PQ}(h)=0 \) numerically, we obtain \( h_{\text{max}} \) and \( h'_{\text{max}} \) as shown in Figure A24. The figure also shows two functions \( h=\epsilon/2 \) and \( h=(0.01+\epsilon)/4 \) for comparison purposes. The result is supportive of our earlier observation and presumption in Set \( \Psi_6 \). The highest peak of \( S_{PQ}(h) \) is observed between \( h=0.005 \) and \( h=\epsilon/2 \) in any case. Our presumption of potential peaks at \( h=0.005 \) and \( h=\epsilon/2 \) still seems to hold since the former is represented as \( h'_{\text{max}} \) while the latter corresponds to \( h_{\text{max}} \) in Figure A24. When \( \epsilon \) is smaller than to 0.036, only a single peak appears at \( h=h'_{\text{max}} \). Two differences from the results obtained for Set \( \Psi_6 \) are that peaks at small \( h \) found in Set \( \Psi_6 \) do not appear at \( h=0.005 \) and that \( h_{\text{max}} \) is larger than
Since it is not clear whether these are caused by approximation error or boundary effect, further numerical experiments are necessary to understand the properties of $S_{PQ}(h)$ and $S'_{PQ}(h)$ in more detail.

Figure A24 The relationship between $\varepsilon$ and the value of $h$ that gives the peaks of $S'_{PQ}(h)$.

**Appendix A5**

This appendix evaluates the behavior of measure $S_{ik}(h)$ and cross $K$-function against positional error. We compare two distributions of points $\Gamma_1$ and $\Gamma_2$ shown in Figure A25a. Distributions from $\Gamma_2(+1)$ to $\Gamma_2(-3)$ are the location of points imprecisely reported as that of $\Gamma_2$. Positional error increases from $\Gamma_2(+1)$ to $\Gamma_2(+3)$ in the positive direction, while increases from $\Gamma_2(-1)$ to $\Gamma_2(-3)$ in the negative direction.

Figure A25b and Figure A25c show $S_{1k}$ and cross $K$-function that compare $\Gamma_1$ and $\Gamma_2$. Measure $S_{12}$ and $K_{12}$ show the actual similarity between $\Gamma_1$ and $\Gamma_2$, while $S_{12(+1)}-S_{12(-3)}$ and $K_{12(+1)}-K_{12(-3)}$ indicate the similarity based on inaccurate data. The two figures are comparable since $S_{1k}$ and cross $K$-function are standardized such a way that they reach their maximum 1.0 when $h \to \infty$. Figure A25b indicates that error in similarity evaluation increases with positional error, positively from $\Gamma_2(+1)$ to $\Gamma_2(+3)$, and negatively from $\Gamma_2(-1)$ to $\Gamma_2(-3)$, at any scale. In $K$-function, on the other hand, whether error in similarity evaluation is positive or negative depends on the scale at which $\Gamma_1$ and $\Gamma_2$ are compared. For instance, error in $K_{12(-3)}$ is positive when $0.021 < h < 0.025$, while error is negative when $0.077 < h < 0.080$.

Figure A26 shows the error in similarity evaluation. Error in $\Delta S_{12}$ gradually increases and then decreases with an increase of $h$ while error suddenly changes in $\Delta K_{12}$. Sudden increase occurs at specific scales such as $h=0.025$, $0.075$, and $0.125$, which are determined by the locational error in point data. Let
us consider the evaluation of similarity at $h=0.025$. Error in $\Delta S_{12}$ increases from 0.0150 to 0.0439 with an increase of locational error. Error in $\Delta K_{12}$, on the other hand, increases to 0.1250 even by a slight locational error. Error in $\Delta K_{12}$ is at least three times as large as that in $\Delta S_{12}$. Error in $\Delta S_{12}$ at $h=0.075$ increases from 0.0013 to 0.0039 while error in $\Delta K_{12}$ is 0.1094. Though such sudden increases do not always occur, they are unavoidable if the true location of points is unknown. Measure $S_{ik}(h)$ is more robust against positional error than cross $K$-function since the similarity evaluated based on $S_{ik}(h)$ does not drastically change if point data are imprecise to some extent.
Figure A25 Similarity between point distributions $\Gamma_1$ and $\Gamma_2$ calculated based on imprecise spatial data. (a) $\Gamma_1$ and $\Gamma_2$: Actual location of points. $\Gamma_2(+1)$-$\Gamma_2(-3)$: Imprecisely reported location of points $\Gamma_2$. (b) Measure $S$. (c) Cross $K$-function.
Figure A26 Error in similarity evaluation by (a) $S_\delta$ and (b) cross $K$-function.