CSIS Discussion Paper No. 145

A method for comparing and classifying point distributions

Yukio Sadahiro* and Yan Liu**

September 2016

*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@csis.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 ranging from local to global. 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 χ-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.

Spatial scale plays a key role in the comparison and classification of point distributions, because evaluation of similarity depends on the scale at which one compares the point distributions. Suppose there are eight point patterns on a one-dimensional space (Figure 1). Points in Γ_1, Γ_2, and Γ_3 are uniformly distributed, but with different starting points. Points in Γ_1 are more closely located with those in Γ_2 than in Γ_3, and thus we regard distribution Γ_1 is more similar to Γ_2 than Γ_3. Distributions from Γ_1 to Γ_3 are similar on a global scale with a slight variation on a local scale. Distributions Γ_4 and Γ_5 both consists of a set of uniform distribution and a point cluster on the right. We may say that they are partially similar to Γ_1 but hesitate to say that they are globally similar to Γ_1. Distributions Γ_7 and Γ_8 also contain point clusters. Though they are similar on a global scale, they are different on a local scale because of the difference in

the two points on the left. Points in distribution Γ_9 are clustered on a global scale and hence this pattern is different from distributions from Γ_1 to Γ_3. Evaluation of similarity is scale-dependent, and hence we need to consider explicitly the spatial scale in the comparison and classification of point distributions.

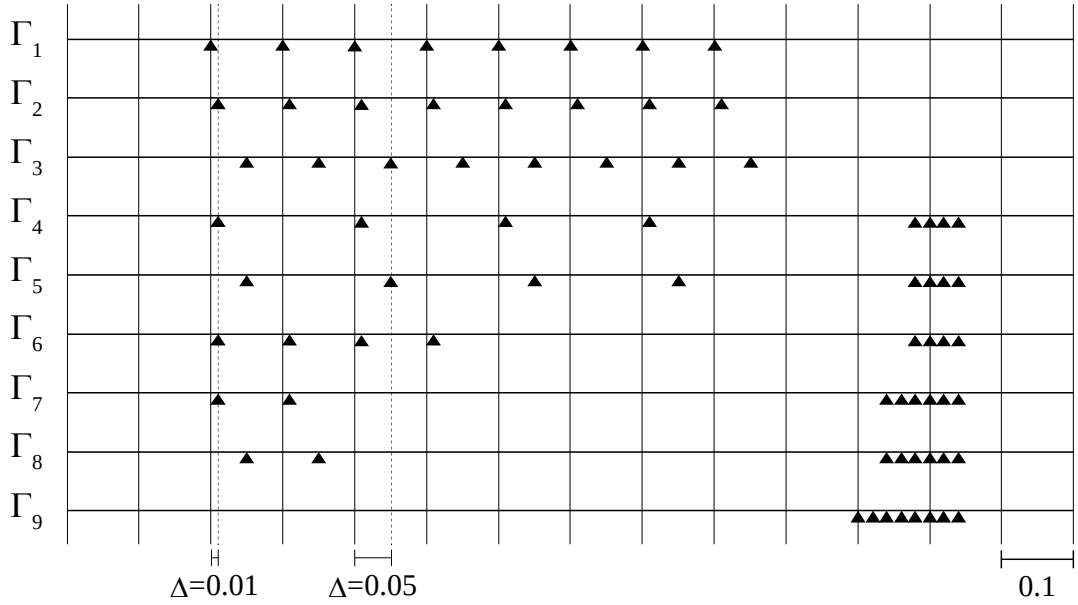


Figure 1 Point distributions on a one-dimensional space.

The nearest neighbor spatial association measure *R**, unfortunately, does not involve the concept of spatial scale. The quadrat method and cross *K*-function, on the other hand, consider the scale of analysis. In the quadrat method, the size of cells used in point counting represents the spatial scale. Large cells conceal local difference between points and thus the distributions tend to look similar. Small cells emphasizes the difference so that the distributions look different. The cross *K*-function represents the similarity as a function of spatial scale, and thus evaluates the similarity at various scales from local to global. The quadrat method and cross *K*-function, however, are sensitive to errors in positional data because they are based on the number of points contained in regions with crisp boundaries. A slight error in positional information may drastically change the evaluation outcome of similarity. Such instability in evaluation is critical since spatial data are generally inaccurate to some extent due to noise and measurement errors. In addition, the assessment of similarity depends on not only the spatial scale but also the location of lattice in the quadrat method. Since the effects of scale and location are inseparable, we cannot evaluate the role of spatial scale explicitly in the comparison and classification of point distributions. The cross *K*-function is free from the location problem, but is not appropriate for classification since it does not provide a single measure.

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

distributions. The method considers the spatial scale explicitly, and is more robust to positional error. 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), Wu and Li (2009), and Zhang et al. (2014)). Among the four scale definitions discussed in Zhang et al. (2014), analysis of point distributions is primarily related to geographical and measurement scales. Geographical scale refers to the spatial extent in which spatial analysis is performed, while measurement scale is the smallest distinguishable unit in data collection, which is often called observation scale or resolution. Let us consider the relationship between these two scales in terms of similarity between point distributions in Figure 1. We use the terms scale and resolution to refer to geographical and measurement scale, respectively.

Distributions from Γ_1 to Γ_3 are all similar on a global scale. The global similarity can be observed at 0.01 or coarser resolutions because a slight difference in the location of points between the distributions is not distinguished at these resolutions. Distributions Γ_1 and Γ_3, on the other hand, requires 0.05 or coarser resolutions to find their global similarity. Distributions Γ_9 requires a very coarse resolution to detect a global similarity between distribution Γ_1, because a finer resolution distinguishes the location of points in Γ_1 and that of points in Γ_9. Distribution Γ_6 consists of four points distributed at the same intervals and four points that are tightly clustered. We often say that Γ_6 is similar to Γ_1 on a local scale but different from Γ_1 on a global scale, or that Γ_6 is partially similar to Γ_1. The local similarity can be detected by 0.01 resolution, while a very coarse resolution is necessary to evaluate the global difference.

The above observation implies that the relationship between scale and resolution is rather complicated and sometimes confusing. Similarity between point distributions detected at a fine resolution is connected with the similarity on both global and local scales, while coarse resolution is primarily related to the similarity on a local scale. Though the terms scale and resolution are often used interchangeably, this paper distinguish these terms in the following, i.e., scale refers to geographical scale while resolution refers to measurement scale.

Scale and resolution of spatial analysis depends on the location or spatial unit of analysis, and the way in which we interpret spatial phenomena. The quadrat method, for instance, counts the number of points in each cell and compares them between different types of points. The cell serves as the basic spatial unit of analysis and the points in each cell are treated equally independent of their relative location. The cross *K*-function counts the number of points located within a certain distance from each reference point. Analysis is performed at each location of reference point, and points are equally evaluated independent of the distance from the reference point.

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 $\boldsymbol{\Gamma}=\{\Gamma_1, \Gamma_2, ..., \Gamma_N\}$. Point distribution Γ_i consists of n_i points (i=1, ..., N). Let P_{ij} and $\mathbf{z}_{ij}$ be the jth point and its location in set Γ_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 Γ_i is mathematically represented as a function of location **x**:

$$F_i\left(\mathbf{x}, h_C; \delta_C\right) = \sum_j \delta_C\left(\mathbf{x}, \mathbf{z}_{ij}, h_C\right), \tag{1}$$

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

$$\delta_C\left(\mathbf{x}, \mathbf{z}_{ij}, h_C\right) = \begin{cases} 1 & \text{if } \left|\mathbf{x}-\mathbf{z}_{ij}\right| \le h_C \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

We call $F_i(\mathbf{x}, h_C; \delta_C)$ the *interpreted surface* of points Γ_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 $F_i(\mathbf{x}, h_C; \delta_C)$ calculated based on distributions Γ_1, Γ_4, Γ_6 and Γ_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 resolution, we call h_C a *resolution 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 $F_i(\mathbf{x}, h_C; \delta_C)$ 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 higher resolutions is concealed by the similarity between distributions observed at coarser resolutions.

Given a certain **x**, interpreted surface $F_i(\mathbf{x}, h_C; \delta_C)$ represents the accumulation of spatial phenomena interpreted at resolutions finer than h_C. Consequently, $F_i(\mathbf{x}, h_C; \delta_C)$ as a function of h_C increases monotonically with h_C. The change in value of $F_i(\mathbf{x}, h_C; \delta_C)$, on the other hand, indicates the point distribution interpreted exactly at h_C. $F_i(\mathbf{x}, h_C; \delta_C)$ 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)).

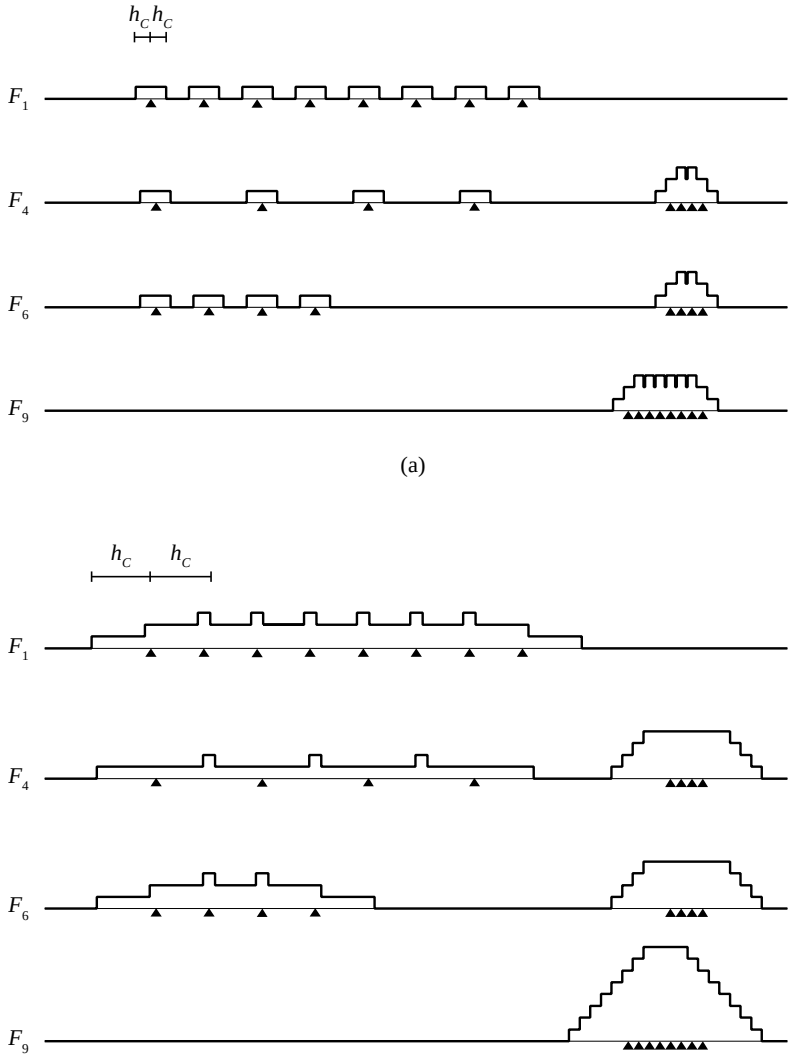


Figure 2 Interpreted surfaces calculated based on the point distributions in Figure 1. (a) Interpreted surfaces with a small h_C. (b) Interpreted surfaces with a large h_C. Points in each distribution are labelled in ascending order from left to right.

Given that $F_i(\mathbf{x}, h_C; \delta_C)$ 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 $\delta(|\mathbf{x}\text{-}\mathbf{z}_{ij}|, h)$. This gives another definition of the interpreted surface:

$$F_i(\mathbf{x}, h; \delta) = \sum_j \delta\left(\left|\mathbf{x} - \mathbf{z}_{ij}\right|, h\right). \tag{3}$$

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

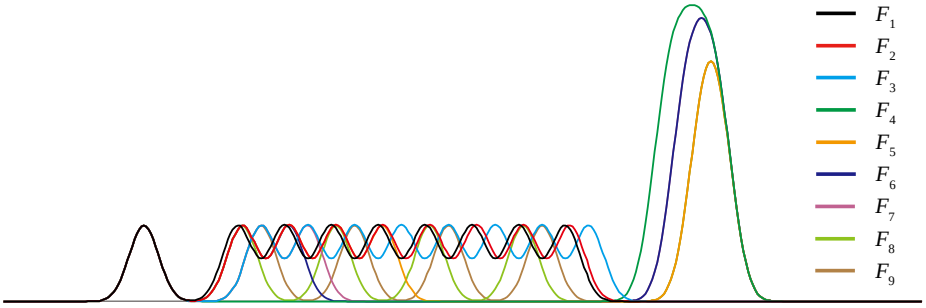
The interpreted surface $F_i(\mathbf{x}, h; \delta)$ is equivalent to kernel smoothing (Silverman 1986). Kernel smoothing puts a small bump called a *kernel* at each observation and sums the kernels up to obtain a surface function. Each kernel and its summation corresponds to $\delta(|\mathbf{x}\text{-}\mathbf{z}_{ij}|,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

$$\delta\left(\left|\mathbf{x} - \mathbf{z}_{ij}\right|, h\right) = e^{-\frac{\left|\mathbf{x}-\mathbf{z}_{ij}\right|^2}{2h^2}}. \tag{4}$$

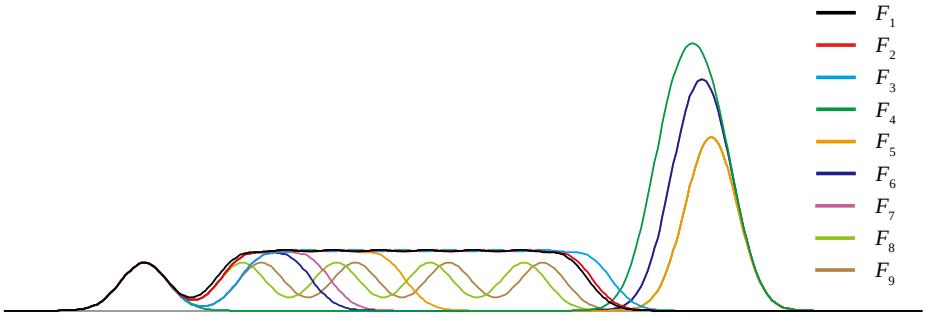
The interpreted surface becomes

$$F_i(\mathbf{x}, h; \delta) = \sum_j e^{-\frac{\left|\mathbf{x}-\mathbf{z}_{ij}\right|^2}{2h^2}}. \tag{5}$$

Figure 3 shows the interpreted surfaces $F_i(\mathbf{x}, h; \delta)$ 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 $F_i(\mathbf{x}, h; \delta)$ are more stable than those on $F_i(\mathbf{x}, h_C; \delta_C)$, 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 resolution conceals the local variation between the distributions in Figure 3b, which supports our earlier observation that the distributions are similarly uniform on a global scale with a slight variation on a local scale.



(a)



(b)

Figure 3 Interpreted surfaces calculated based on the point distributions in Figure 1. (a) Interpreted surfaces with a small resolution parameter h. (b) Interpreted surfaces with a large resolution parameter h.

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

$$\begin{aligned} F_i'(\mathbf{x}, h; \delta) &= \frac{\mathrm{d}}{\mathrm{d}h} F_i(\mathbf{x}, h; \delta) \\ &= \sum_j \frac{\mathrm{d}}{\mathrm{d}h} e^{-\frac{|\mathbf{x}-\mathbf{z}_{ij}|^2}{2h^2}} \\ &= \sum_j \frac{|\mathbf{x}-\mathbf{z}_{ij}|^2}{h^3} e^{-\frac{|\mathbf{x}-\mathbf{z}_{ij}|^2}{2h^2}} \end{aligned} . \tag{6}$$

$F'_i(\mathbf{x}, h; \delta)$ is the derivative of $F_i(\mathbf{x}, h; \delta)$ with respect to h, which we call the *interpreted density* of points Γ_i at resolution h. It represents the point distribution interpreted exactly at resolution h, i.e., the distribution interpreted on the ring of radius h centered at $\mathbf{x}$. Thus, the interpreted surface $F_i(\mathbf{x}, h_C; \delta_C)$ can be obtained

as:

$$F_i\left(\mathbf{x}, h; \delta\right) = \int_{h_0=0}^{h} F_i'\left(\mathbf{x}, h_0; \delta\right) \mathrm{d}h_0 \ . \tag{7}$$

Figure 4 shows the interpreted densities $F'_i(\mathbf{x}, h; \delta)$ calculated based on distributions Γ_1, Γ_4, Γ_6 and Γ_9 in Figure 1. $F'_i(\mathbf{x}, h; \delta)$ increases around points, the degree of which depends on the distance from the origin and that from points. $F'_i(\mathbf{x}, h; \delta)$ increases drastically near the origin when the both distances are small as seen in $F'_1(\mathbf{x}, h; \delta)$ and $F'_4(\mathbf{x}, h; \delta)$ in Figure 4. $F'_i(\mathbf{x}, h; \delta)$ becomes long-tailed in Γ_4, Γ_6 and Γ_9 that contain point clusters located far from the origin.

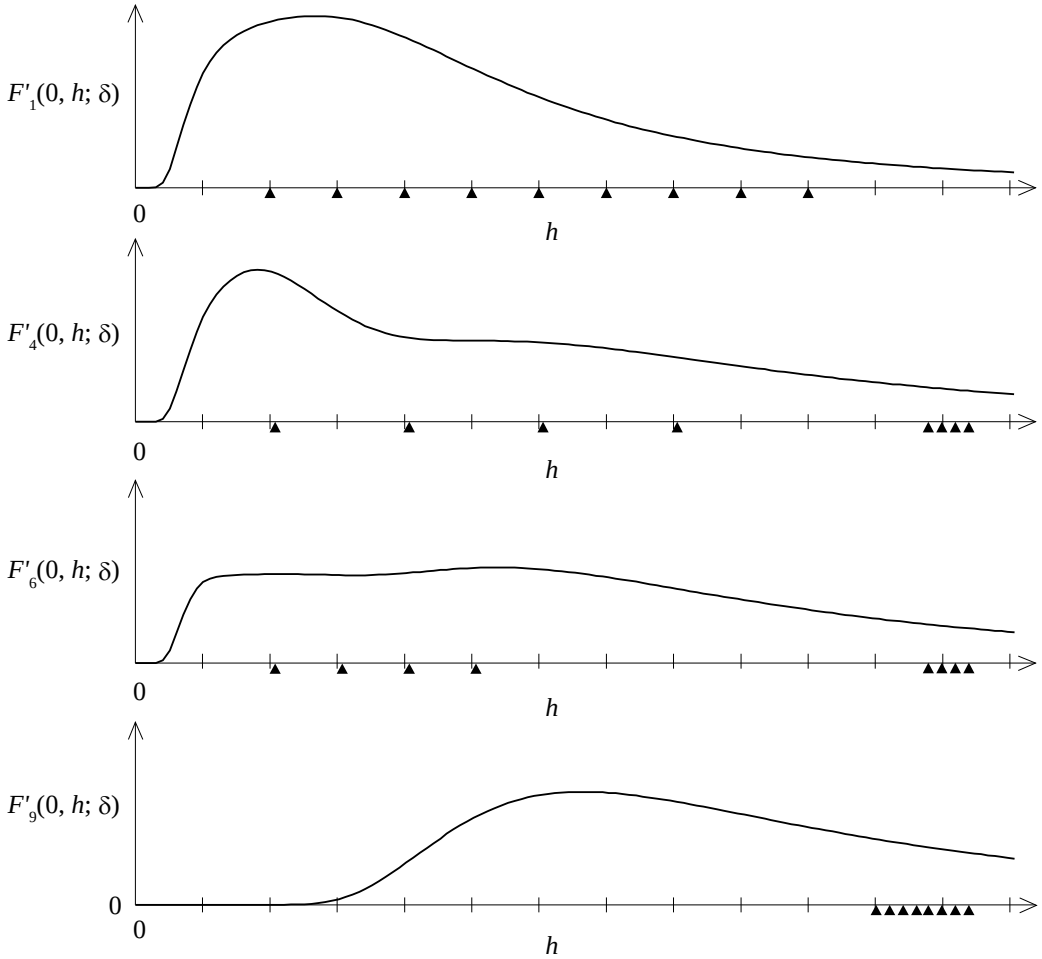


Figure 4 Interpreted density $F'_i(0, h; \delta)$ as a function of h calculated based on the point distributions in Figure 1, where $\mathbf{x}$=0 is the starting point of the horizontal axis on which points are distributed.

2.2 Comparison of point distributions

Using the interpreted surface, we evaluate the similarity between point distributions. *The*

similarity measure between Γ_i and Γ_k is defined as

$$S_{ik}(h) = 1 - \frac{1}{2}\int \left| f_i(\mathbf{x}, h; \delta) - f_k(\mathbf{x}, h; \delta) \right| \mathrm{d}\mathbf{x},$$

(8)

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

$$f_i(\mathbf{x}, h; \delta) = \frac{1}{n_i}\sum_j \frac{\delta\left(\left|\mathbf{x} - \mathbf{z}_{ij}\right|, h\right)}{\int \delta\left(\left|\mathbf{x} - \mathbf{z}_{ij}\right|, h\right)\mathrm{d}\mathbf{x}} = \frac{1}{\sqrt{2\pi} n_i h}\sum_j e^{-\frac{\left|\mathbf{x} - \mathbf{z}_{ij}\right|^2}{2h^2}},$$

(9)

(k=1, …, N, 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

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

(10)

where m_{ik} is the number of points in Γ_i that shares the same location with a point in Γ_k. If the locations of all the points in Γ_i and Γ_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 Γ_i and Γ_k with an explicit consideration of spatial resolution. The measure $S_{ik}(h)$ evaluates the accumulation of similarity between $F_i(\mathbf{x}, h; \delta)$ and $F_k(\mathbf{x}, h; \delta)$ at resolutions finer than h, and the change of $S_{ik}(h)$ indicates the similarity exactly at resolution h. The latter is mathematically represented as

$$S'_{ik}(h) = \frac{\mathrm{d}}{\mathrm{d}h} S_{ik}(h) = -\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}h}\int \left| f_i(\mathbf{x}, h) - f_k(\mathbf{x}, h) \right| \mathrm{d}\mathbf{x},$$

(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 Γ_1 and all other distributions (Γ_2-Γ_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 resolution of observation gradually changes from high to low with an increase in h and thus the difference between distributions vanishes. Consequently, $S_{ik}(h)$ is an increasing function of h, and $S'_{ik}(h)$ is non-negative for any h with multiple peaks of different height.

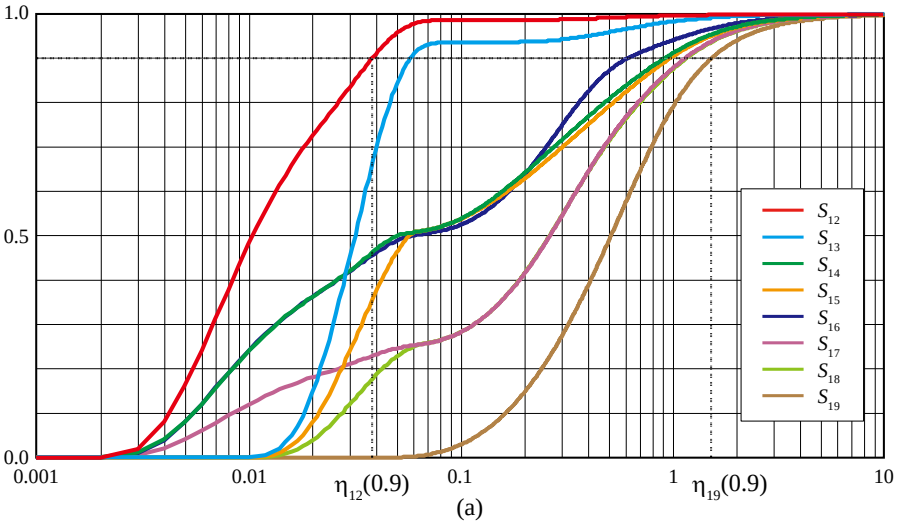
Figure 1 shows that distributions Γ_2 and Γ_3 are similar to Γ_1 on a global scale, while Γ_4 to Γ_9 are different since they contain point clusters. Measures $S'_{12}(h)$ and $S'_{13}(h)$ are characterized by their significant peaks at small h, which are much higher than their lower peaks observed in 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 on a global scale emerges as a significant peak at a fine resolution. It is consistent with our earlier discussion, i.e., a slight difference vanishes at a fine resolution when point distributions are similar on a global scale. When point distributions are not similar on a global scale, a coarse resolution is necessary to regard the distributions similar with each other.

Distributions Γ_4, Γ_6 and Γ_7 are similar to Γ_1 on a local scale because they contain several points that are closely located to those in Γ_1. This results in peaks of $S'_{ik}(h)$ at a small h as seen in Figure 5b. These distributions also have peaks at large h (Figure 5c), which implies that a coarse resolution is necessary to recognize the similarity on a global scale between these distributions and Γ_1. From this we can say that multiple peaks containing peaks at a fine resolution indicate the similarity only on a local scale between point distributions.

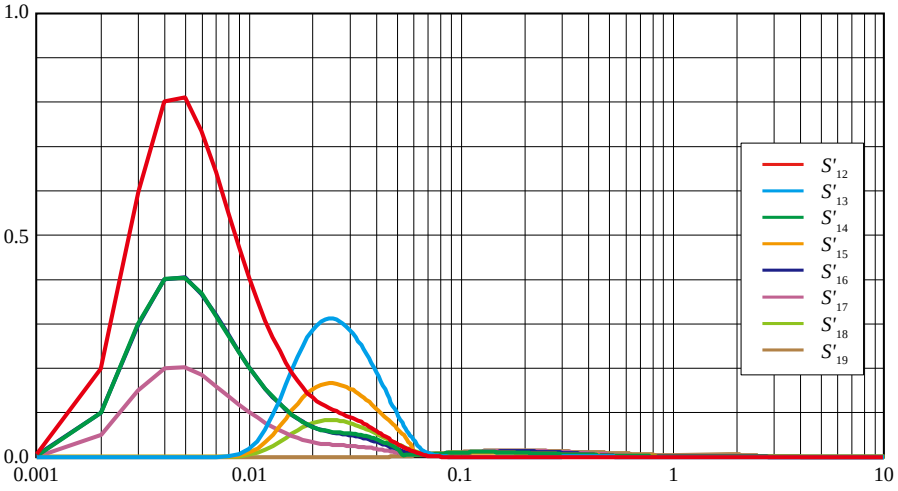
The highest peaks of Γ_4, Γ_6 and Γ_7 are observed at h=0.005, which is the half of the distance between some points in Γ_4, Γ_6, and Γ_7 and their nearest points in Γ_1. For instance, the distance between four points on the left in Γ_4 and their nearest points in Γ_1 is 0.01 as seen in Figure 1. Similarly, distributions Γ_3, Γ_5 and Γ_8 have peaks at h=0.025, which is also the half the distance between some points in Γ_3, Γ_5 and Γ_8 and their nearest points in Γ_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 Γ_9 is totally different from Γ_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 resolution and a significant peak at a coarse resolution indicate that point distributions are different on both global and local scales.

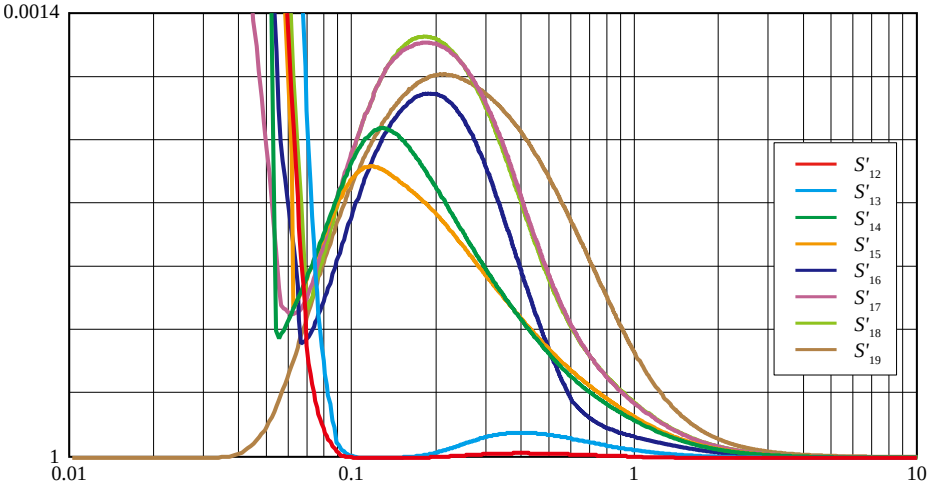
We may summarize the above observation as follows. Given two distributions, we say that they are similar on a global scale when $S'_{ik}(h)$ has a significant peak at a small h. If $S'_{ik}(h)$ has a low peak at a small h, the distributions are similar on a local scale but different on a global scale. If $S'_{ik}(h)$ does not have any peak at a small h, the distributions are different on both global and local scales.


1.0
0.5
0.0
0.001
0.01
0.1
1
10
$\eta_{12}(0.9)$
$\eta_{19}(0.9)$
S_{12}
S_{13}
S_{14}
S_{15}
S_{16}
S_{17}
S_{18}
S_{19}

(a)


1.0
0.5
0.0
0.001
0.01
0.1
1
10
S'_{12}
S'_{13}
S'_{14}
S'_{15}
S'_{16}
S'_{17}
S'_{18}
S'_{19}

(b)


0.0014
1
0.01
0.1
1
10
S'_{12}
S'_{13}
S'_{14}
S'_{15}
S'_{16}
S'_{17}
S'_{18}
S'_{19}

(c)

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 measures $S_{ik}(h)$ and $S'_{ik}(h)$ take into account the spatial resolution explicitly, and thus are useful for evaluating the similarity between Γ_i and Γ_k by resolutions. These measures, however, are not directly applicable to the classification of point distributions since most classification methods prefers to use one measure to quantify the similarity between each pair of distributions. In addition, $S_{ik}(h)$ and $S'_{ik}(h)$ are rather inconvenient to capture the entire similarity between point distributions. A single measure is useful and easy to understand the overall properties of distributions especially when we treat numerous types of points. We thus propose a single measure based on $S_{ik}(h)$ that represents the overall similarity between point distributions.

The similarity between point distributions monotonically increases by accumulating the similarity at increasing resolutions. The measure $S_{ik}(h)$ increases very rapidly when Γ_i and Γ_k are globally similar while it increases slowly when Γ_i and Γ_k are dissimilar globally. This permits us to evaluate the overall similarity of distributions by how fast they become similar with the increase in h. We propose a measure $\eta_{ik}(\alpha)$ defined by

$$\begin{aligned} \alpha &= \int_{h=0}^{\eta_{ik}(\alpha)} S'_{ik}\left(h\right)\mathrm{d}h \\ &= S_{ik}\left(\eta_{ik}\left(\alpha\right)\right) \end{aligned} . \tag{12}$$

The right side of the equation represents the accumulation of similarity at resolutions from h=0 to $\eta_{ik}(\alpha)$. $\eta_{ik}(\alpha)$ equals to h when the accumulation reaches α. This measure enables us to evaluate the overall similarity between distributions Γ_i and Γ_k. When Γ_i and Γ_k are highly similar on a global scale, $S_{ik}(h)$ increases rapidly and the accumulation reaches α at a small h. On the other hand, if Γ_i and Γ_k are not so similar, $S_{ik}(h)$ increases slowly and hence $\eta_{ik}(\alpha)$ becomes large. Therefore, $\eta_{ik}(\alpha)$ serves as a distance measure between Γ_i and Γ_j.

Parameter α 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 resolutions using a large α, as a small α implies that the comparison of distributions is only at fine resolutions. On the other hand, α=1 is meaningless as $\eta_{ik}(1)=\infty$ in any case. We thus recommend using a large α value such as α=0.9 or α=0.95 as is often adopted as the level of significance in statistical tests.

Figure 5a shows examples of $\eta_{ik}(\alpha)$ where α=0.9. The values of $\eta_{ik}(\alpha)$ reveal that distribution Γ_1 is most similar to Γ_2 but least similar to Γ_9. The values also indicate that distributions Γ_2 and Γ_3 are even more similar to Γ_1 than to Γ_4-Γ_9, which is consistent with our intuition mentioned earlier.

$\eta_{ik}(\alpha)$ works as a distance measure between Γ_i and Γ_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.

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.

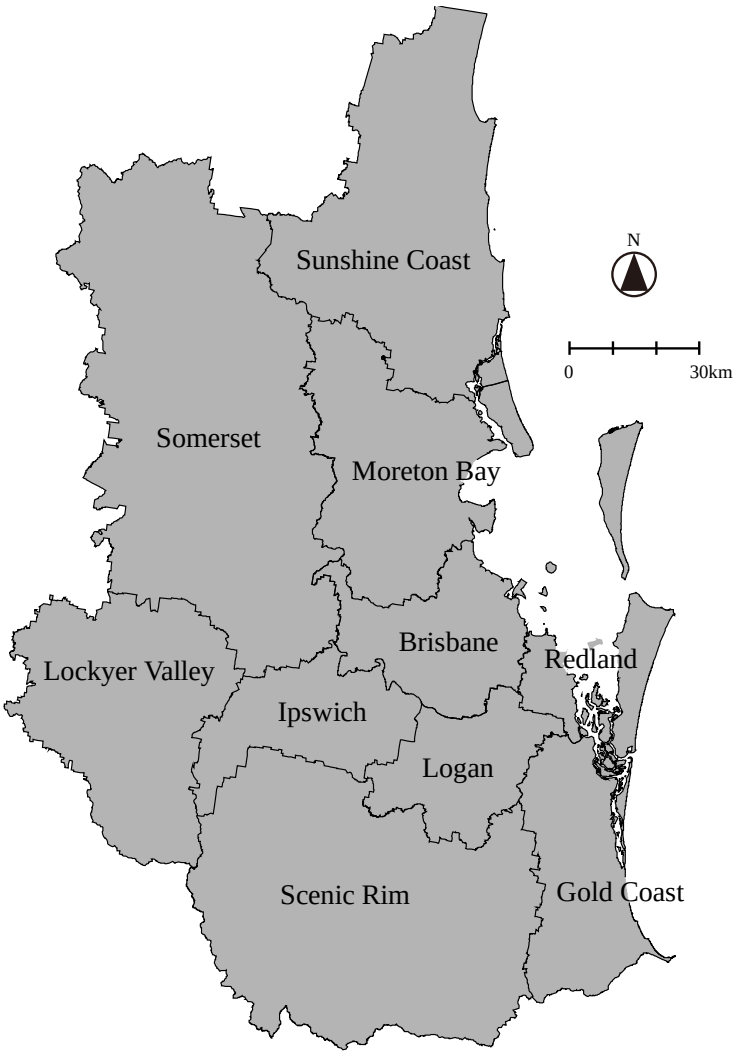


Figure 6 South East Queensland consisting Brisbane and its surrounding local government areas.

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 at global scale

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 α 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. γ in Table 1 represents the average distance between commuters and the medoid of their group and ρ 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 γ and ρ 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 γ and ρ 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. γ is the average distance between commuters and the medoid of their group, while ρ 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.

	Number of commuters	γ	ρ	G01	G02	G03	G04	G05	G06
G01	37	45.59	77.76						
G02	110	10.35	19.07	250.0					
G03	48	27.31	44.36	359.4	62.5				
G04	19	30.84	52.65	484.4	187.5	80.1			
G05	14	18.67	35.90	812.5	515.6	406.3	281.3		
G06	19	46.84	87.10	523.4	828.1	937.5	1062.5	1000.0	
G07	10	6.24	12.91	968.8	679.7	578.1	453.1	1000.0	128.9

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 ρ value in G02 (19.07m) and G07 (12.91m) but rather large ρ 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.

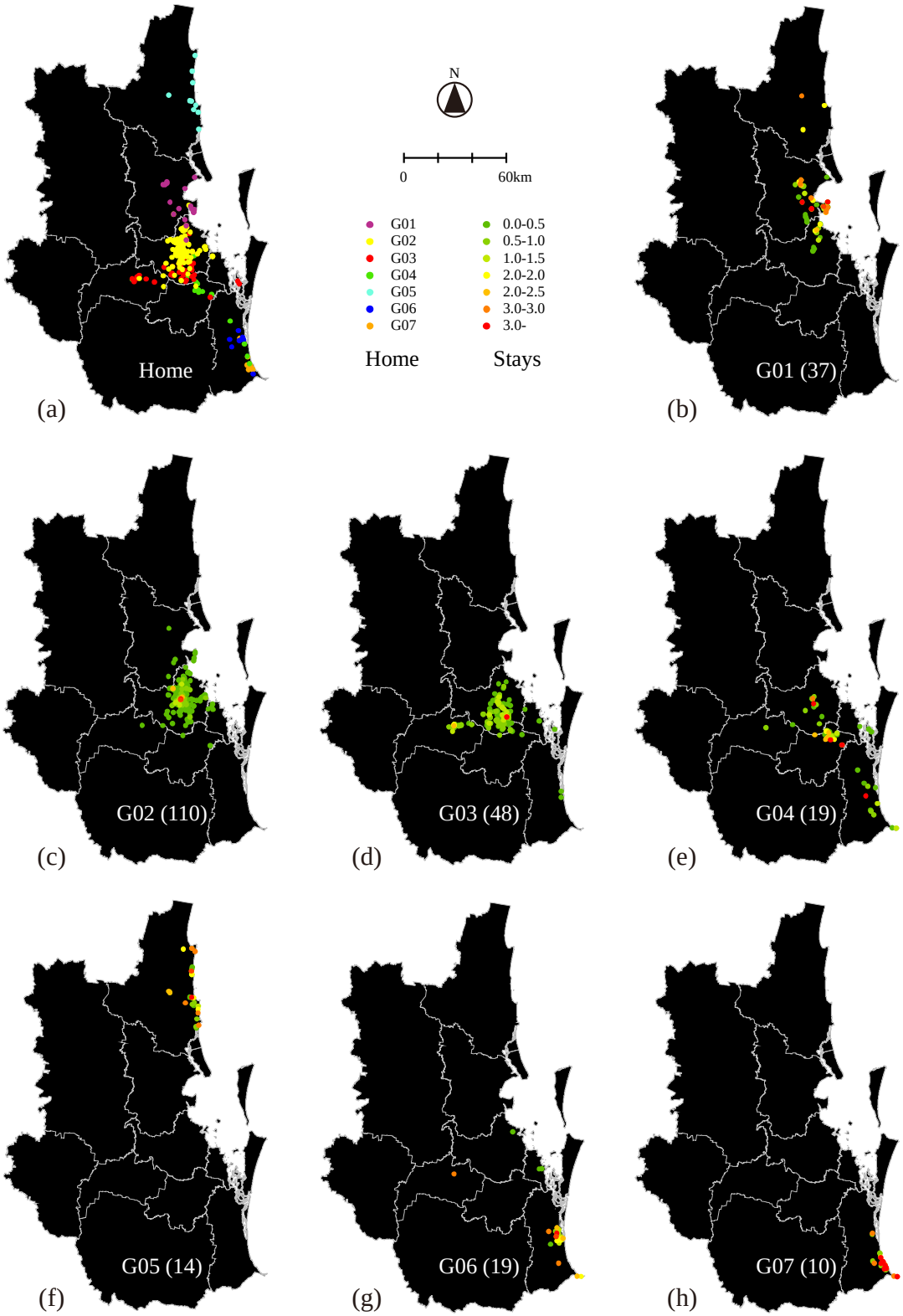


Figure 7 The location of homes (a) and stays (b-h) of the frequent commuters classified into seven groups. Color of points in G01 to G07 indicates the relative number of stays made by the frequent

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 at local scale

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. γ is the average distance between commuters and the medoid of their group, while ρ 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.

	Number of travelers	γ	ρ	G11	G12	G13	G14	G15	G16
G11	42	12.08	21.26						
G12	11	15.21	26.76	281.3					
G13	11	22.42	42.86	437.5	103.5				
G14	37	7.10	13.10	101.6	121.1	281.3			
G15	12	11.28	19.07	91.8	351.6	500.0	179.7		
G16	16	30.84	52.65	82.0	406.3	562.5	234.4	117.2	
G17	85	4.87	9.11	44.9	187.5	343.8	23.4	128.9	168.0

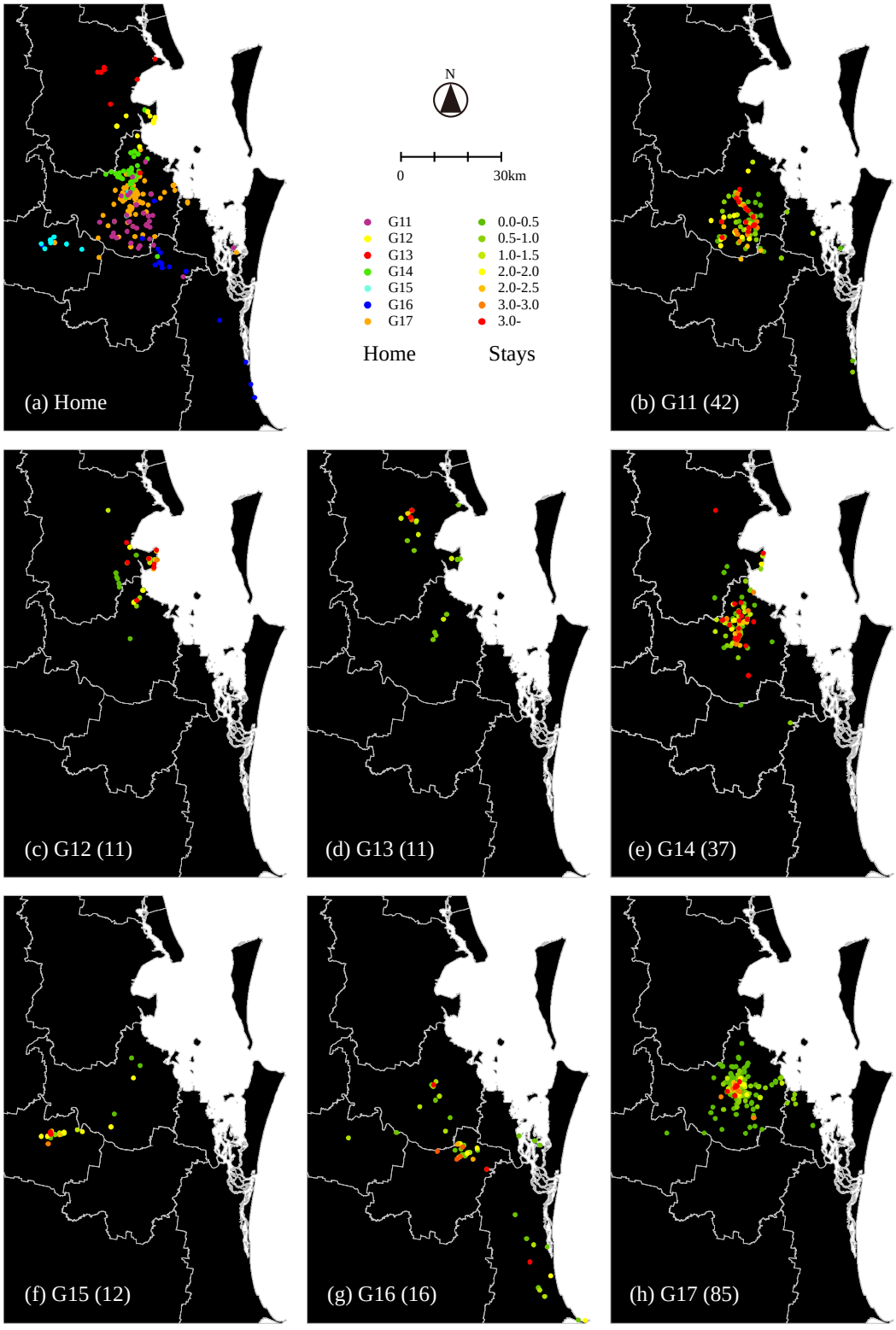


Figure 8 The location of homes (a) and stays (b-h) of frequent commuters living in and around Brisbane area. Colors of points in the maps of stays indicate the relative number of stays made by the frequent

commuters at each stop. Numbers in parentheses are the number of frequent commuters in each group.

We then review in a more local scale 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 on a global scale as discussed in the Section 2, which can also be confirmed by the relatively smaller γ and ρ values. On the other hand, large γ and ρ 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 globally different but 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 but rather different globally, 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.

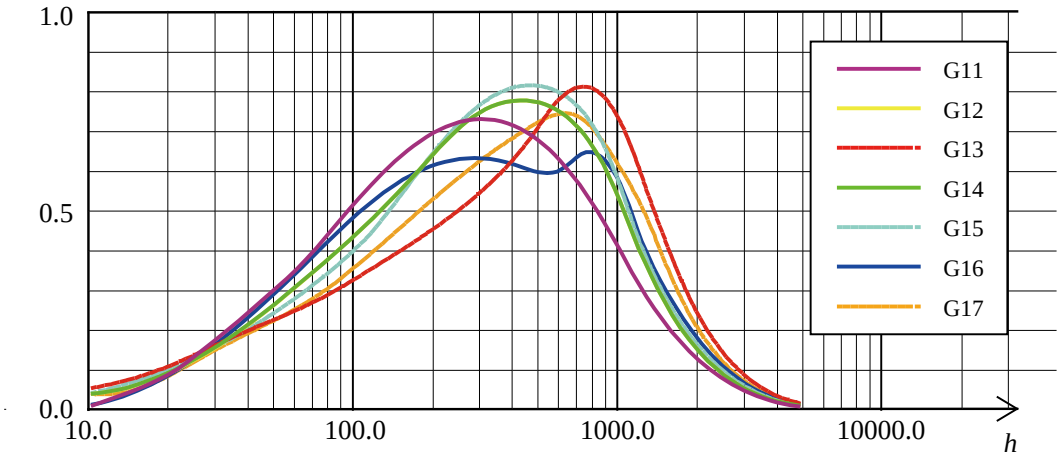


Figure 9 Average differential similarity measures between frequent commuters in each group.

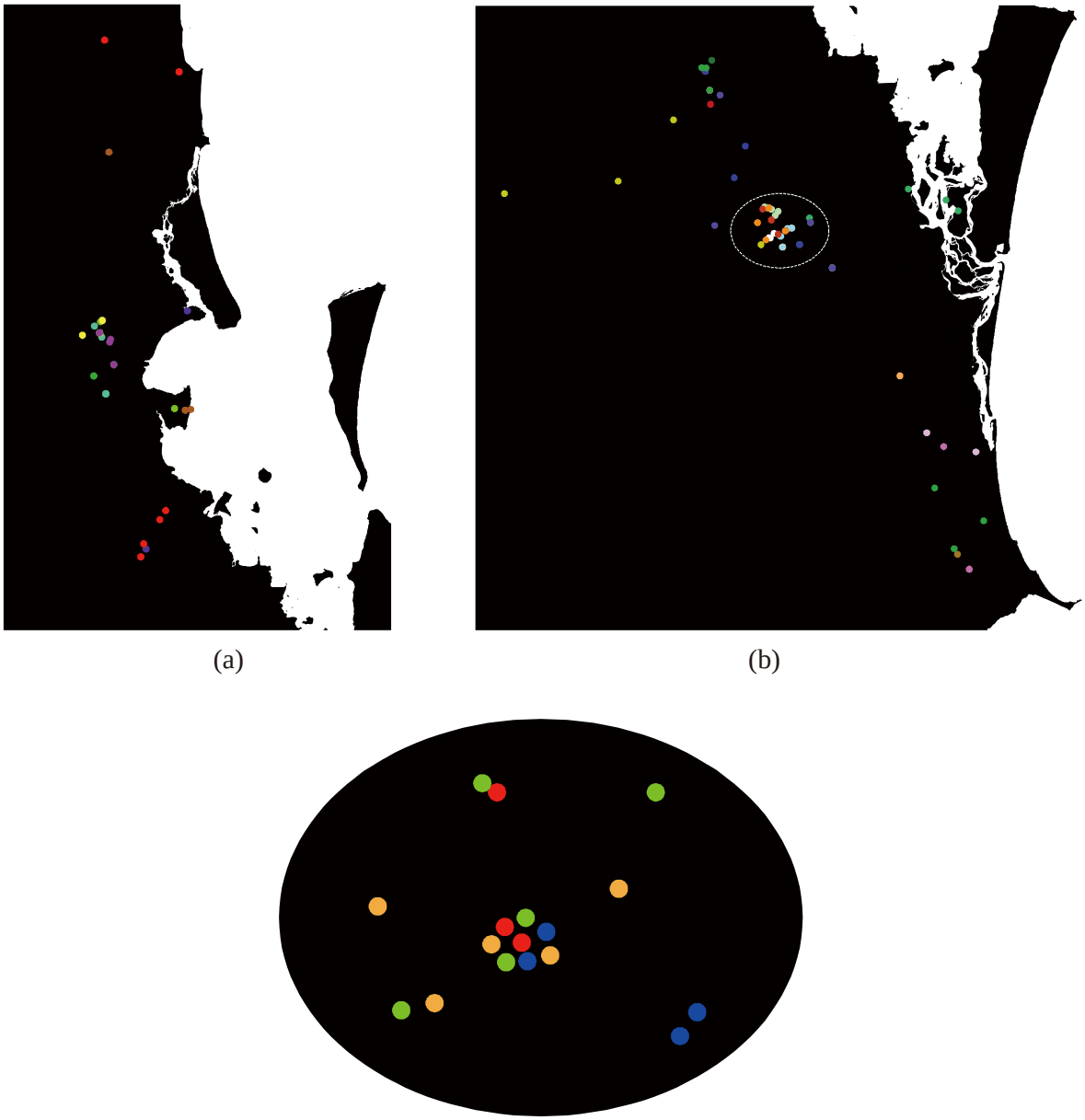


Figure 10 The location of stays of frequent commuters of (a) G13 and (b)(c) G16. (c) Stays of four commuters within the dotted elliptic area in Figure 10b at a large scale. Different colors of points indicate different commuters.

3. Concluding discussion

This paper presents a new method of analyzing point distributions. The method considers the spatial scale explicitly, and applicable to both comparison and classification of point distributions. The method is more robust to positional errors that are inevitable in spatial data. The robustness is assured by

the conversion of point distributions into continuous surfaces as discussed in Section 2. To test the validity of the proposed method, we applied the method to the analysis of trip data of the public transport users in Southeast Queensland, Australia. The results support the technical soundness of the method, and provided meaningful insights on the travel patterns of the Pensioner Concession Card holders that may not be easily observable by visual analysis and other existing methods.

The proposed approach is not without limitations. First, this proposed data clustering method is based on exploratory spatial analysis rather than statistical analysis. Though measures $S'_{ik}(h)$ and $\eta_{ik}(\alpha)$ are useful for comparing the similarity between point distributions, their statistical significance is not tested. As such, the classification of distributions is not statistically validated. Nevertheless, this exploratory spatial analysis is necessary since a strict view of statistics often conceals subtle but interesting patterns in spatial phenomena. Cluster analysis and spatial data mining are often not accompanied with statistical discussion to avoid overlooking such hidden patterns. However, statistical analysis permits us to reach more objective and persuasive results. Incorporation of statistical perspective into the proposed method should be considered in future research.

Second, the choice of parameter α 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 α should be. Theoretical approach in general statistics seems difficult to derive a unique value of α. One promising option in search of a suitable α value might be to apply the proposed method by varying the α value in a wide variety of situations. This will enable further investigation on the relationship between α 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

Cressie N (2015) *Statistics for spatial data*. John Wiley & Sons,

Diggle PJ (1983) *Statistical analysis of spatial point patterns*. Academic press,

Everitt BS, Landau S, Leese M, Stahl D (2011) *Cluster analysis, 5th edition*. Wiley,

Hennig C, Meila M, Murtagh F, Rocci R (2015) *Handbook of cluster analysis*. Chapman and Hall/CRC,

Kiskowski MA, Hancock JF, Kenworthy AK (2009) On the use of ripley's k-function and its derivatives to analyze domain size. *Biophysical Journal* 97**:** 1095-1103

Lam NS-N, Quattrochi DA (1992) On the issues of scale, resolution, and fractal analysis in the mapping sciences*. *The Professional Geographer* 44**:** 88-98

Lee Y (1979) A nearest-neighbor spatial-association measure for the analysis of firm interdependence. *Environment and Planning A* 11**:** 169-176

Okabe A, Miki F (1984) A conditional nearest-neighbor spatial-association measure for the analysis of conditional locational interdependence. *Environment and Planning A* 16**:** 163-171

Quattrochi DA, Goodchild MF (1997) *Scale in remote sensing and gis*. CRC press,

Ripley BD (1976) The second-order analysis of stationary point processes. *Journal of applied probability***:** 255-266

Ripley BD (1977) Modelling spatial patterns. *Journal of the Royal Statistical Society. Series B (Methodological)* 39**:** 172-212

Ripley BD (2005) *Spatial statistics*. John Wiley & Sons, New York

Silverman BW (1986) *Density estimation for statistics and data analysis*. CRC Press, Boca Raton

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

Wu H, Li Z-L (2009) Scale issues in remote sensing: A review on analysis, processing and modeling. *Sensors* 9**:** 1768-1793

Zhang J, Atkinson P, Goodchild MF (2014) *Scale in spatial information and analysis*. CRC Press,

Appendix A1

This appendix discusses in detail the similarity measure $S_{ik}(h)$ and the differential similarity measure $S'_{ik}(h)$ through numerical experiments. The experiments employ eleven sets of point distributions Ψ_1-Ψ_{11} on a one-dimensional space. Every set consists of multiple distributions $\{\Gamma_1, \Gamma_2, ..., \Gamma_N\}$, in which we compare distribution Γ_1 with other distributions in the same set. We move every point in Γ_1 by distance Δ to generate other distributions in sets from Ψ_1 to Ψ_9, where Δ may vary across points in each distribution. Let h_{max} and h'_{max} be the values of h that gives the highest and the second highest peaks of $S'_{1k}(h)$, respectively. This appendix omits (h) in $S_{1k}(h)$ and $S'_{1k}(h)$ for simplicity.

Set Ψ_1 consists of point distributions each of which is obtained by moving Γ_1 to the right by a constant distance Δ. The distance Δ varies from 0.01 to 0.05 between distributions as shown in Figure A1a. S_{1k} gradually moves from left to right with Δ, implying that the global similarity decreases from Γ_2 to Γ_6. The value of h_{max} seems almost proportional to Δ. It is close the half of Δ, which implies that h_{max} roughly tells us Δ when Δ is constant within the distribution. Appendix A2 gives a theoretical support of this relationship between h_{max} and Δ.

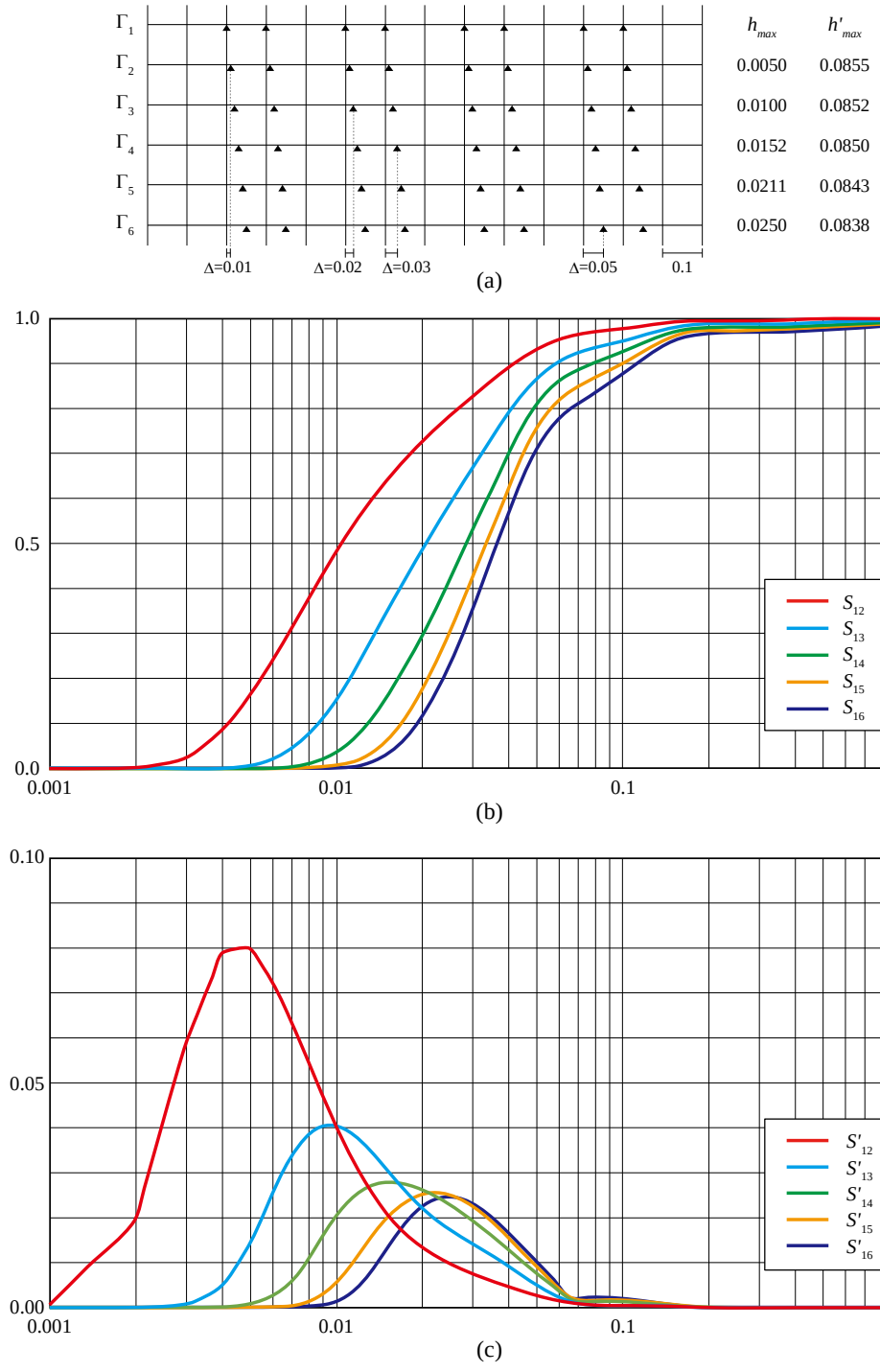


Figure A1 Measures S_{1k} and S'_{1k} in set Ψ_1.

Similar to the point distributions in Ψ_1, those in Ψ_2 are obtained by moving the points in Γ_1 by distance Δ 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 Ψ_2 are very similar

to those in set Ψ_1; S'_{1k} has a single peak whose h_{max} is again almost the half of Δ. Our earlier presumption that h_{max} roughly tells us Δ still holds when points move in different directions. A difference lies in the absence of the second highest peak in Ψ_2. This seems relevant to the distance between points in Γ_k and their neighboring but not nearest points in Γ_1.

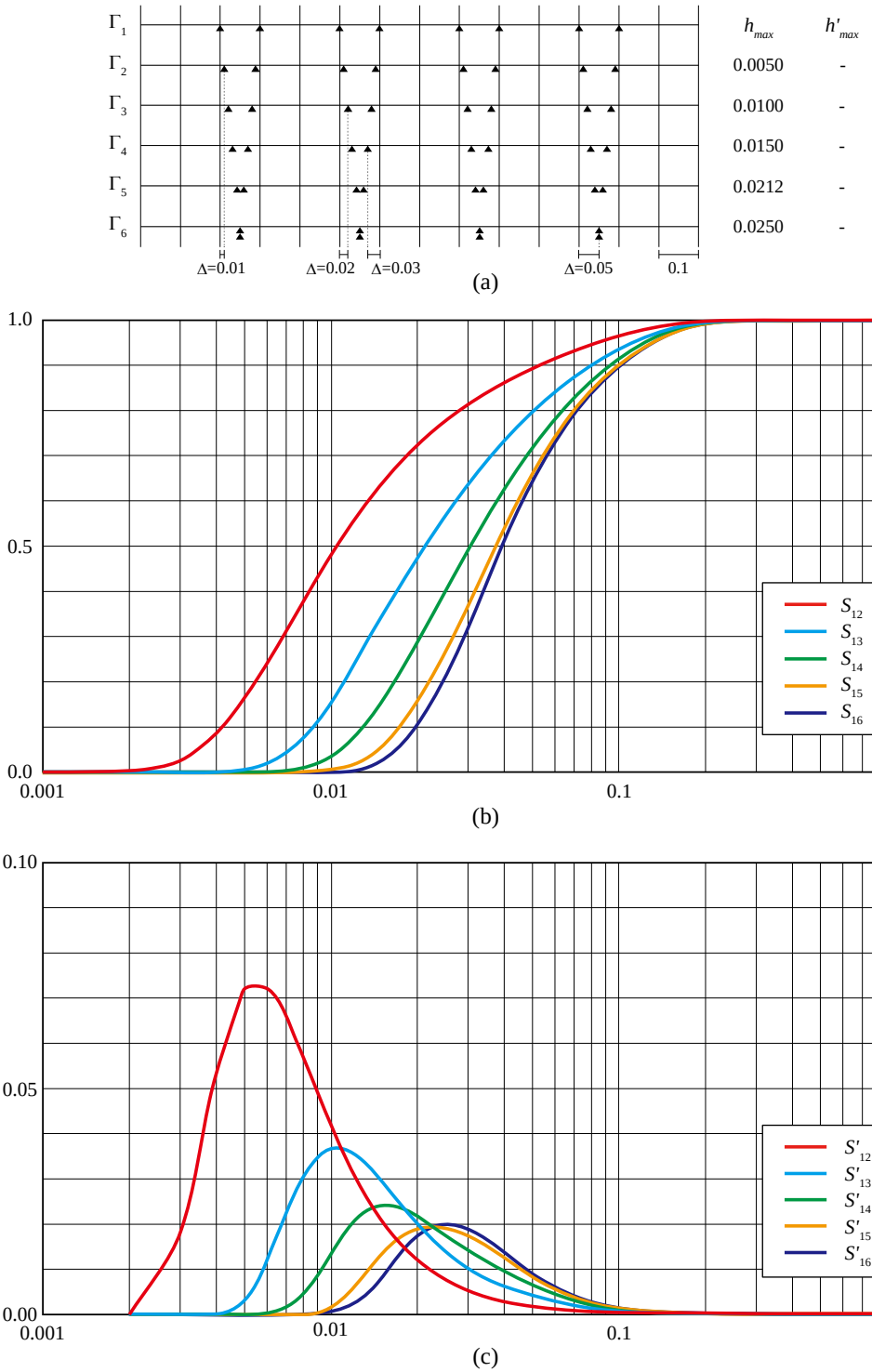


Figure A2 Measures S_{1k} and S'_{1k} in set Ψ_2.

Set Ψ_3 consists of point distributions that are partially similar to Γ_1. Four points are obtained by moving four points on the left in Γ_1 to the right by distance 0.01, while the other four are located at

distance 0.05 from their nearest points in Γ_1. The former yields a partial similarity between Γ_1 and Γ_2-Γ_6, which emerges as a gradual increase of S_{1k} in Figure A2b. The highest peaks of S'_{1k} are lower than those in Figure A1b, which is due to the difference between Γ_1 and Γ_2-Γ_6 on a global scale. The form of S'_{1k} in Figure A2c is generally similar to S'_{1k} 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 Ψ_1. A clear difference lies in the small peak of S'_{12} at h=0.0389. This reflects the difference between Γ_1 and Γ_2 on a global scale caused by the four points on the right.

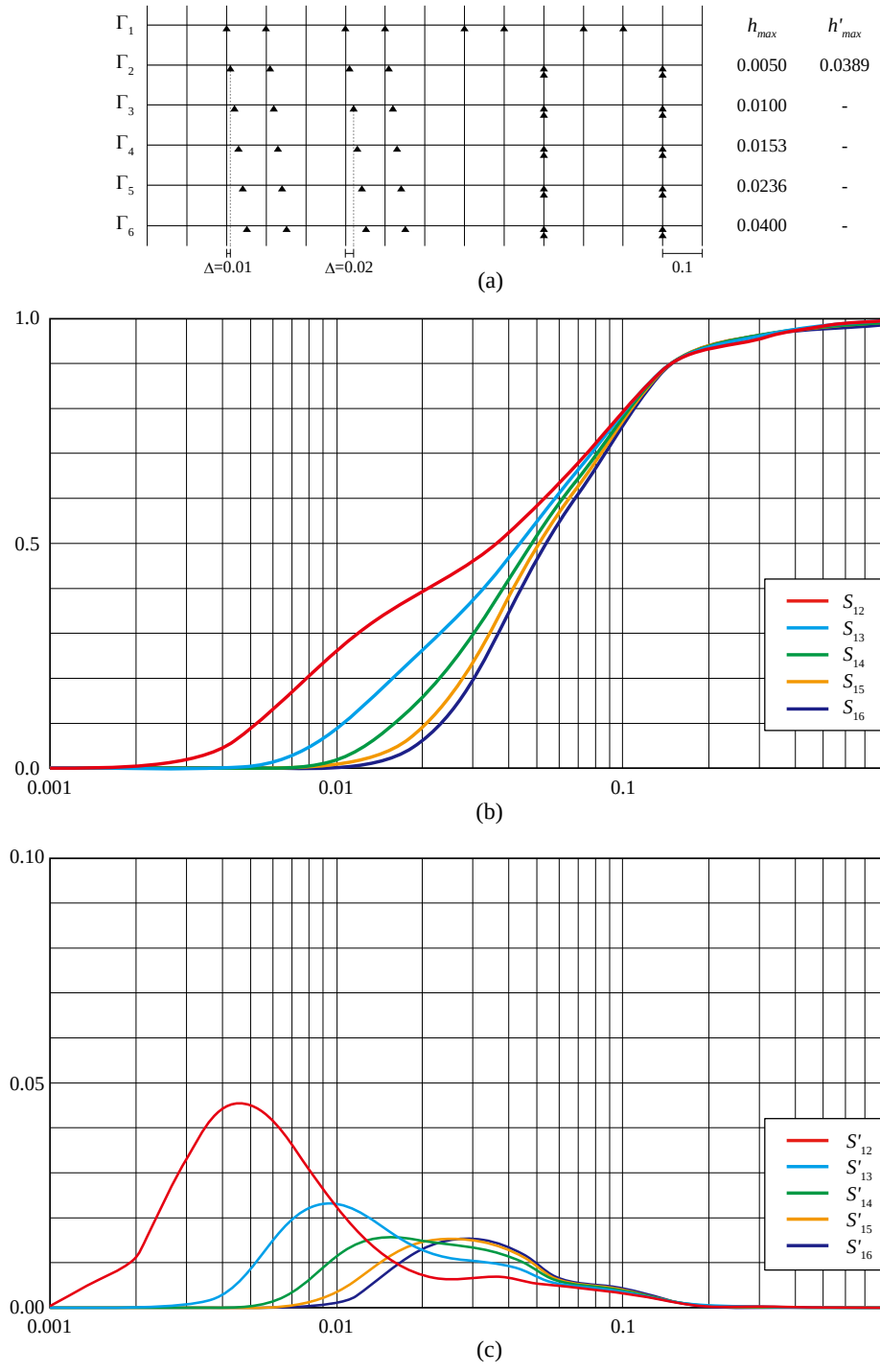


Figure A3 Measures S_{1k} and S'_{1k} in set Ψ_3.

Similar to Set Ψ_3, Set Ψ_4 contains point distributions that are partially similar to Γ_1. Two points on the left are distributed similarly with the distributions in Ψ_3, while the others are located at distance 0.05 from their nearest points in Γ_1. Distributions Γ_2-Γ_6 are less similar to Γ_1 than Γ_2-Γ_6 in Set Ψ_3. This

difference changes the form of S'_{1k} 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 Γ_2-Γ_6 are only partially similar to Γ_1.

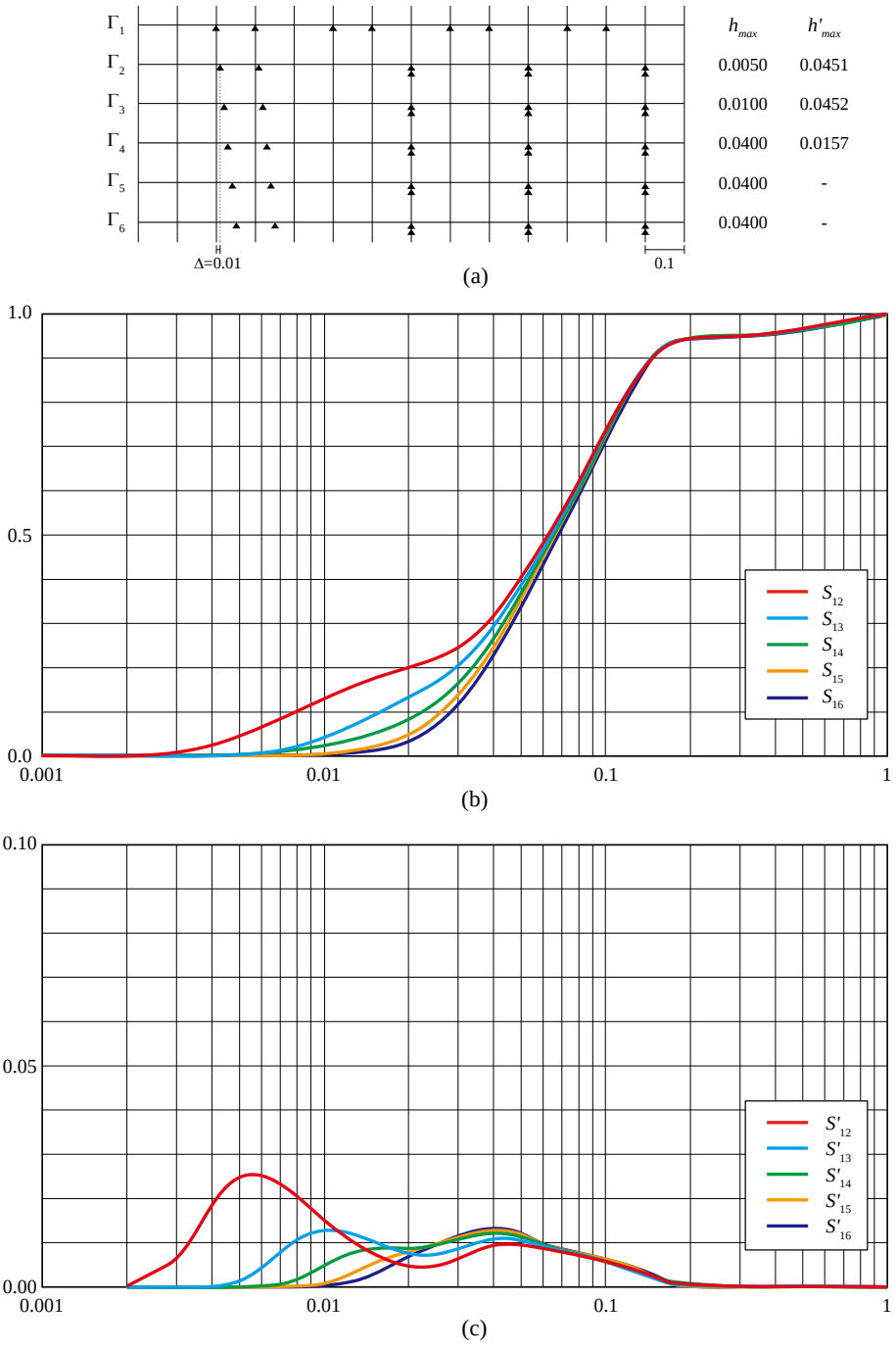


Figure A4 Measures S_{1k} and S'_{1k} in set Ψ_4.

Set Ψ_5 moves points in Γ_1 by two different distances Δ_1=0.01 and Δ_2=0.05 within each distribution. For instance, we obtain Γ_2 by moving all the points in Γ_1 by distance Δ_1, while we obtain Γ_3 by moving six points by Δ_1 and two points by Δ_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 Γ_i and its nearest point in Γ_1, indicated as Δ_1 and Δ_2 in Figure A5a. Point distribution becomes less similar with Γ_1 on a global scale from Γ_2 to Γ_6.

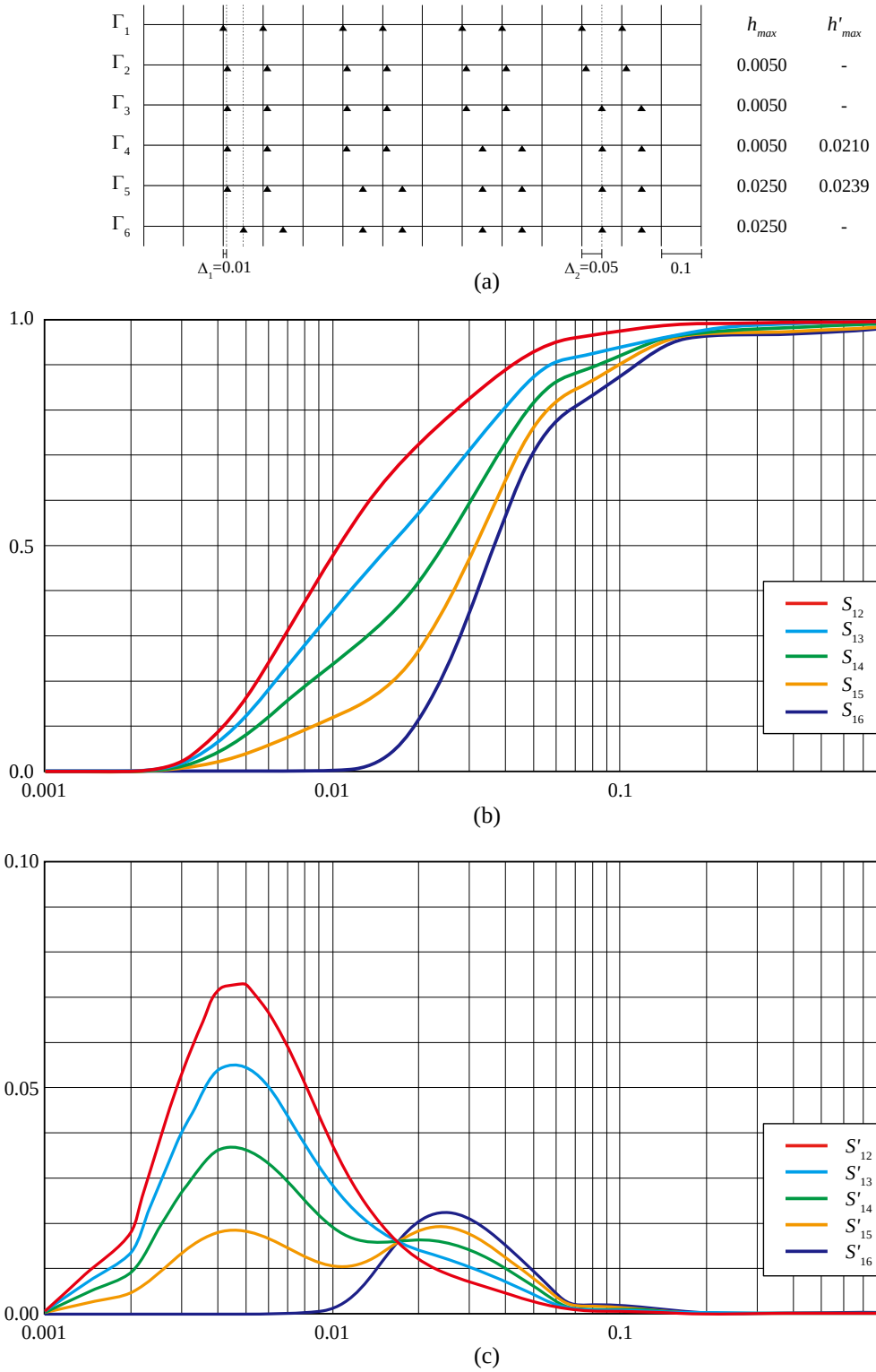
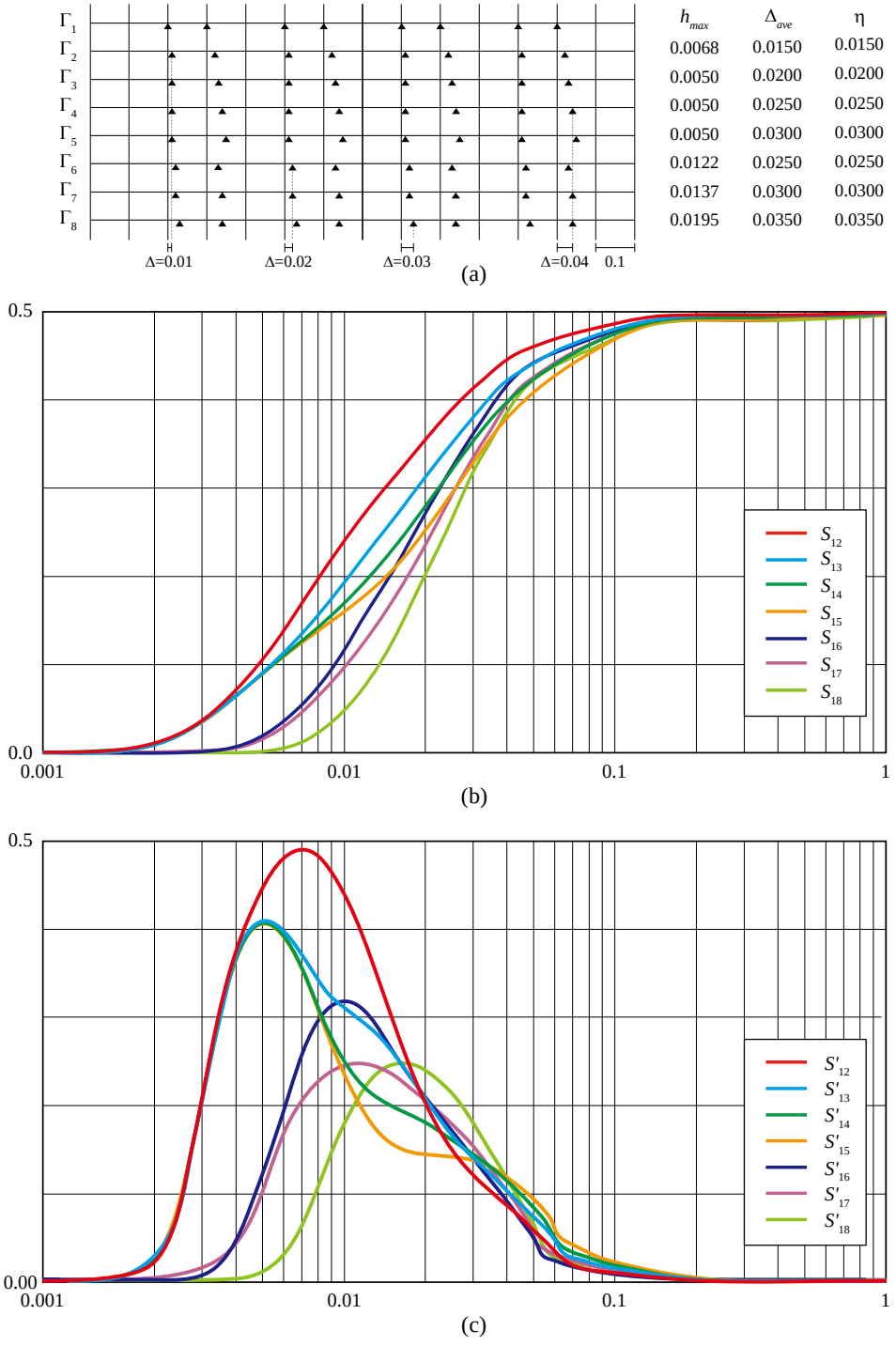


Figure A5 Measures S_{1k} and S'_{1k} in set Ψ_5.

Set Ψ_6 also moves points in Γ_1 by two different distances Δ_1 and Δ_2 within each distribution.

For instance, $(\Delta_1, \Delta_2)=(0.01, 0.02)$ in Γ_2 and $(\Delta_1, \Delta_2)=(0.01, 0.03)$ in Γ_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 Γ_1 and those in the other distributions, i.e., the nearest point from P_{21} in Γ_1 is P_{11}, and that from P_{11} in Γ_2 is P_{21}.

We newly introduce two variables Δ_{ave} and η. The former is the average of Δ_1 and Δ_2, while the latter is the average distance from points in Γ_k and their nearest points in Γ_1. The two variables are equal as long as the mutual nearest relationship holds. Figure A6a shows that h_{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_{max} is smaller than the half of Δ_{ave} in most cases, i.e., h_{max} is closer to $\Delta_1/2$ than $\Delta_2/2$. h_{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}).

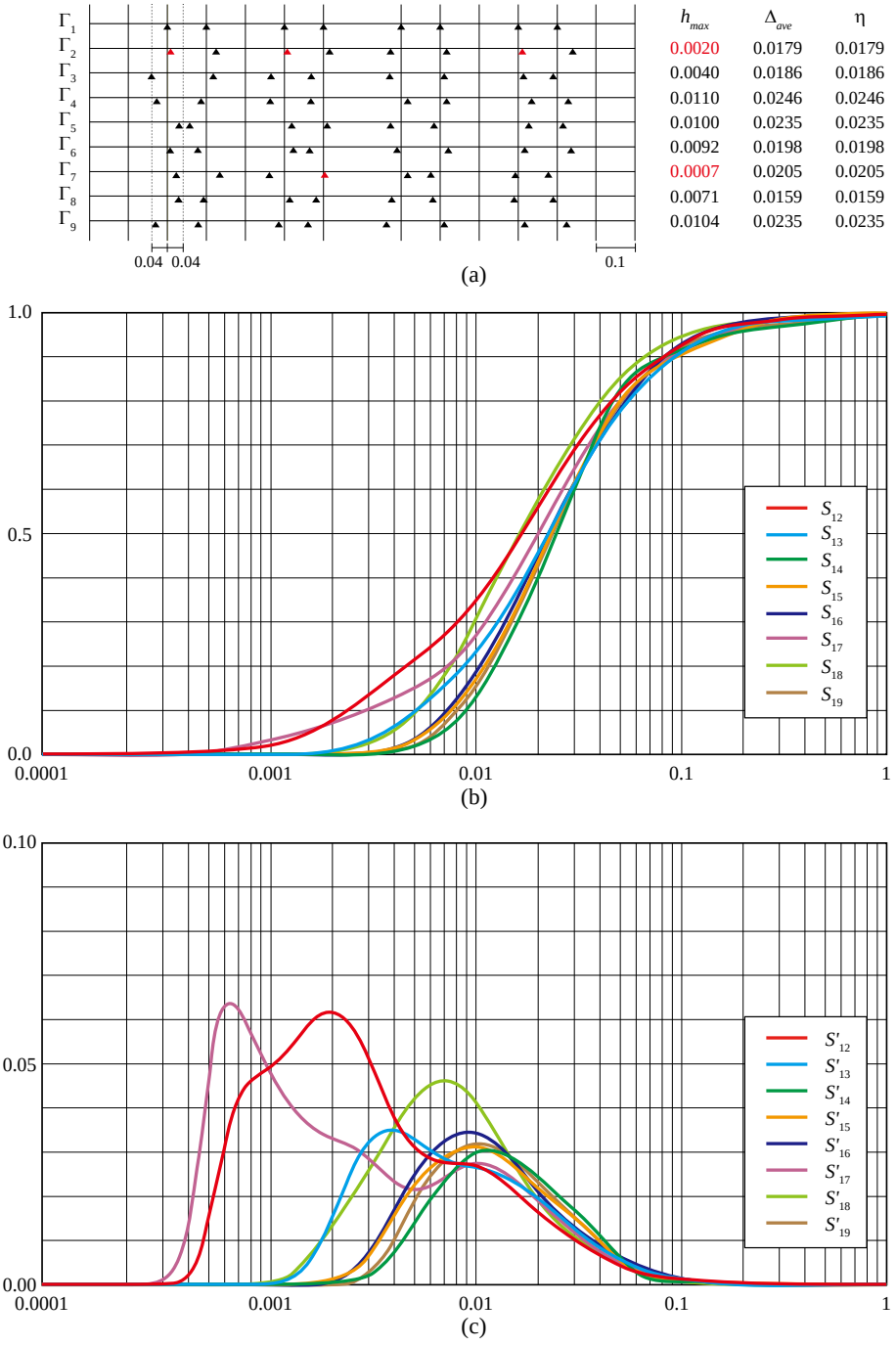


h_{max}	Δ_{ave}	η
0.0068	0.0150	0.0150
0.0050	0.0200	0.0200
0.0050	0.0250	0.0250
0.0050	0.0300	0.0300
0.0122	0.0250	0.0250
0.0137	0.0300	0.0300
0.0195	0.0350	0.0350

Figure A6 Measures S_{1k} and S'_{1k} in set Ψ_6.

The next three sets of distributions treat more general cases where points in Γ_1 are moved by a variety of distances within each distribution. Distance Δ is distributed stochastically according to a uniform distribution in each set. The direction of movement is determined randomly.

Distance Δ follows the uniform distribution between 0.00 and 0.04 in Set Ψ_7. This assures the mutual nearest relationship between the points in Γ_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 Δ_{ave} in most cases. When some points in Γ_k are located very closely to a points in Γ_1, S'_{1k} may have multiple peaks one of which is observed at a small h. Red point in Γ_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 Γ_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.

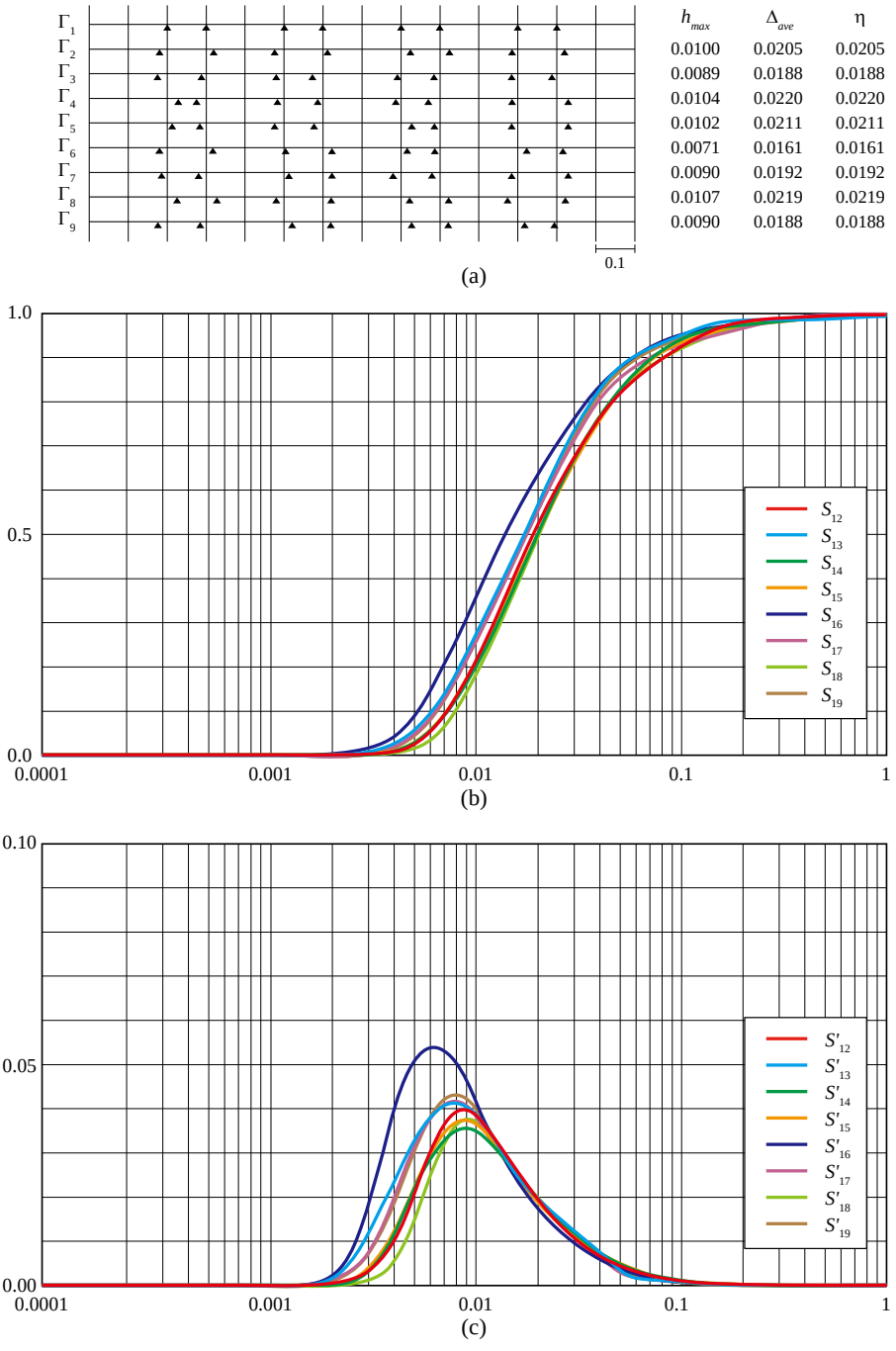


h_{max}	Δ_{ave}	η
0.0020	0.0179	0.0179
0.0040	0.0186	0.0186
0.0110	0.0246	0.0246
0.0100	0.0235	0.0235
0.0092	0.0198	0.0198
0.0007	0.0205	0.0205
0.0071	0.0159	0.0159
0.0104	0.0235	0.0235

Figure A7 Measures S_{1k} and S'_{1k} in set Ψ_7.

Set Ψ_8 consists of points distributed in a narrower range, .i.e., distance Δ follows the uniform distribution between 0.01 and 0.03. The mutual nearest relationship between the original points in Γ_1 and others still holds. This results in that the variation of Δ_{ave} is smaller than that in Set Ψ_7 as shown in Figure A8a. Measure S'_{1k} has only a single peak around h=0.01 in all the distributions. Measures S_{1k} and S'_{1k} are

more similar to those of S_{13} and S'_{13} in Set Ψ_1, where Δ=0.02 for all the points (Figure A1).

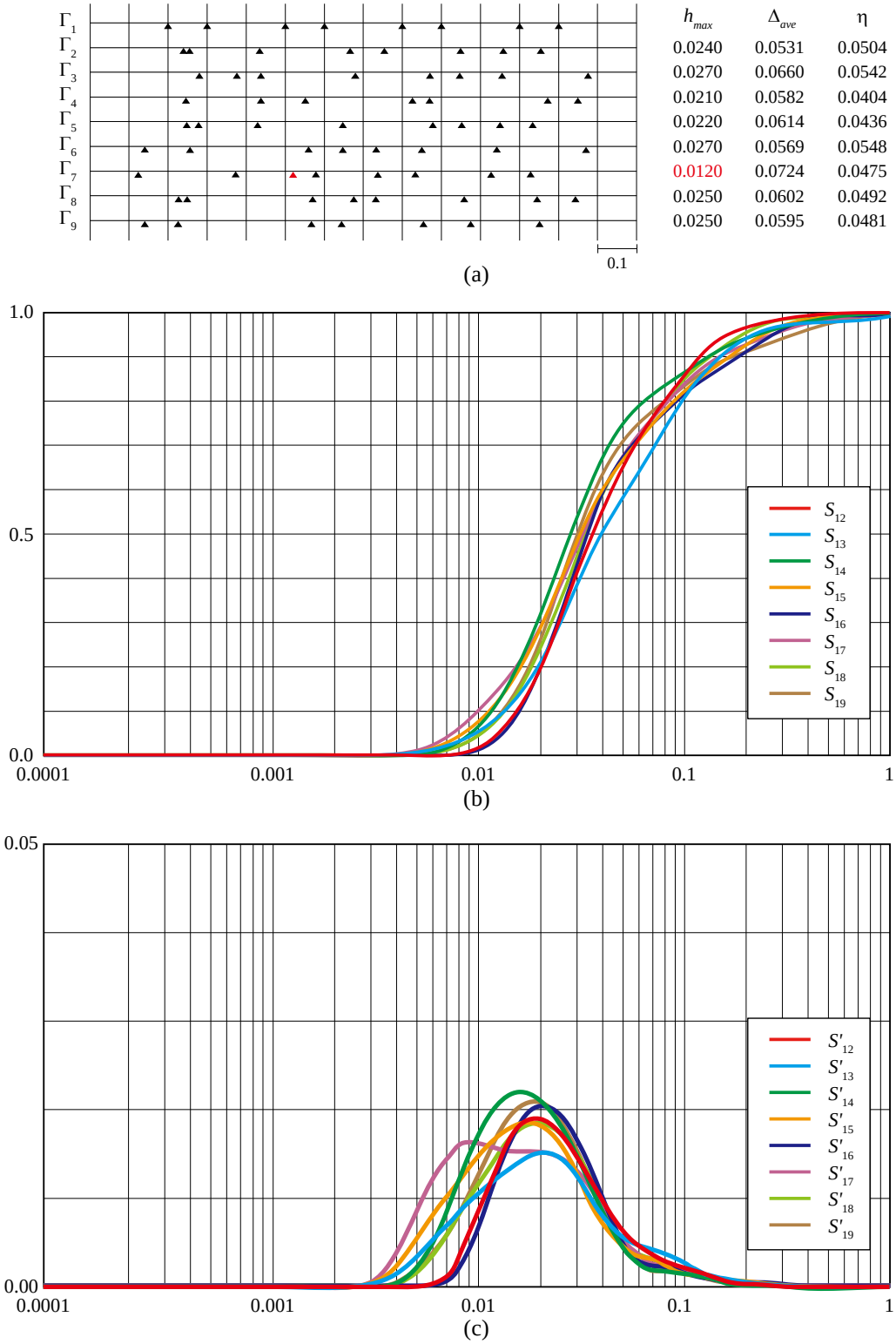


h_{max}	Δ_{ave}	η
0.0100	0.0205	0.0205
0.0089	0.0188	0.0188
0.0104	0.0220	0.0220
0.0102	0.0211	0.0211
0.0071	0.0161	0.0161
0.0090	0.0192	0.0192
0.0107	0.0219	0.0219
0.0090	0.0188	0.0188

Figure A8 Measures S_{1k} and S'_{1k} in set Ψ_8.

Set Ψ_9 consists of point distributions where distance Δ follows the uniform distribution between 0.04 and 0.08. The mutual nearest relationship may fail in this case, i.e., points in Γ_k can be closer to point

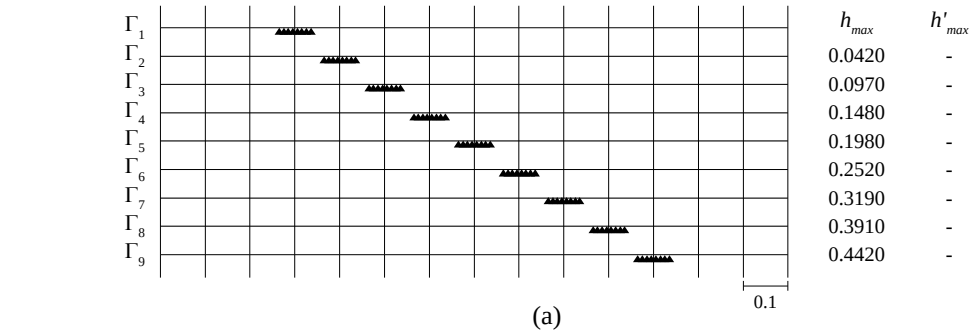
in Γ_1 other than those at their original locations. Consequently, Δ_{ave} is not always equal to η as shown in Figure A9a. Measures S'_{11} to S'_{16} and S'_{18} have a single peak, while S'_{17} has two peaks at h=0.0120 and 0.0240. The peak at h=0.0120 is due to a red point in Γ_7 that is very close to a point in Γ_1. Many highest peaks are distributed around h=0.0250, which is closer to the half of η rather than that of Δ_{ave}. This suggests that h_{max} reflects the distance between points in Γ_i and their nearest points in Γ_1.



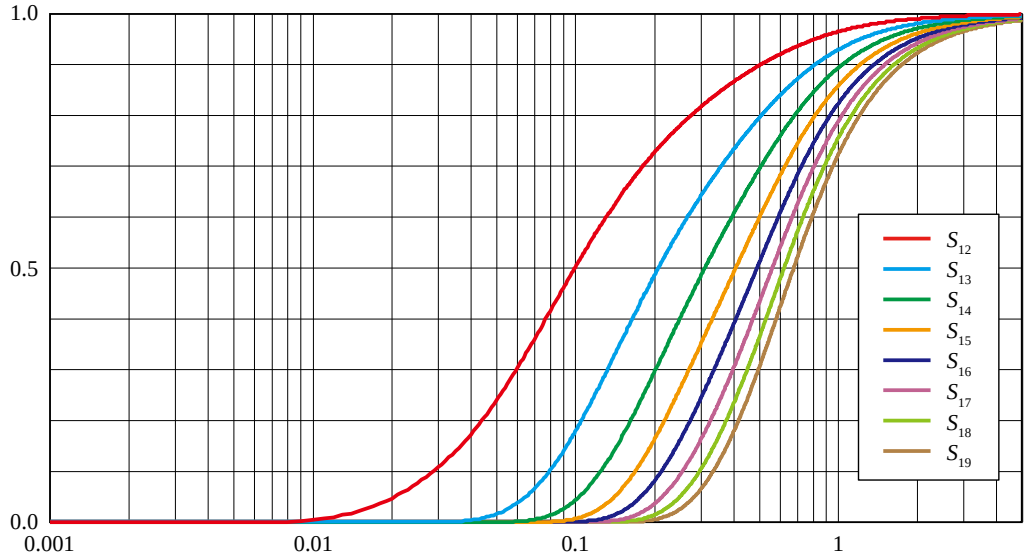
h_{max}	Δ_{ave}	η
0.0240	0.0531	0.0504
0.0270	0.0660	0.0542
0.0210	0.0582	0.0404
0.0220	0.0614	0.0436
0.0270	0.0569	0.0548
0.0120	0.0724	0.0475
0.0250	0.0602	0.0492
0.0250	0.0595	0.0481

Figure A9 Measures S_{1k} and S'_{1k} in set Ψ_9.

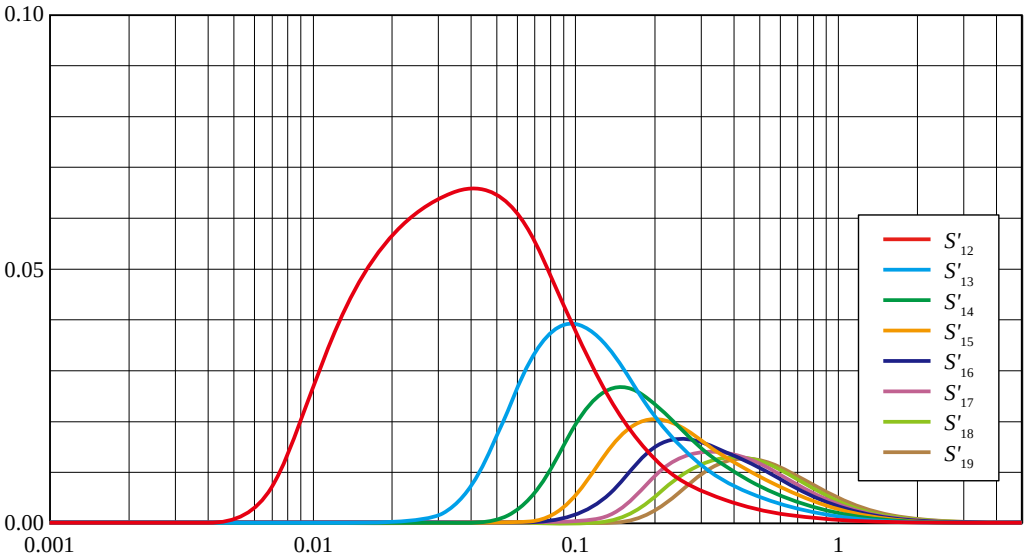
Set Ψ_{10} consists of point distributions in each of which points are clustered at intervals of 0.01. Distribution Γ_k is obtained by moving points in Γ_1 to the right by (k-1)*0.1. Measure S_{1k} gradually shifts from left to right with an increase in the distance between point clusters in Γ_k and Γ_1. Measure S'_{1k} becomes lower and its peak moves to the right. The value of h_{max} seems almost proportional to the distance between point clusters in Γ_k and Γ_1. This fails in Γ_8 and Γ_9 probably because of the boundary effect, i.e., calculation is performed only in the limited range as shown in the figure.



(a)



(b)



(c)

Figure A10 Measures S_{1k} and S'_{1k} in set Ψ_{10}.

Set Ψ_{11} analyzes the relationship between the uniformity of points and the similarity measures. Figure A11a shows distributions from Γ_1 to Γ_6. Distribution Γ_1 consists of seventeen points while distributions from Γ_2 to Γ_6 consist of sixteen points. Points in Γ_1 are distributed uniformly at intervals 0.1. Each point in Γ_2 is located at the center of neighboring points in Γ_1. Points become gradually clustered from Γ_2 to Γ_6, reducing the uniformity of points.

Similarity to Γ_1 decreases monotonically from Γ_2 to Γ_6 as shown in the figure. Measure S'_{1k} has its peak around h=0.025 for all the distributions from Γ_2 to Γ_6, which is the half of the distance between a point in Γ_k and its nearest point in Γ_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.

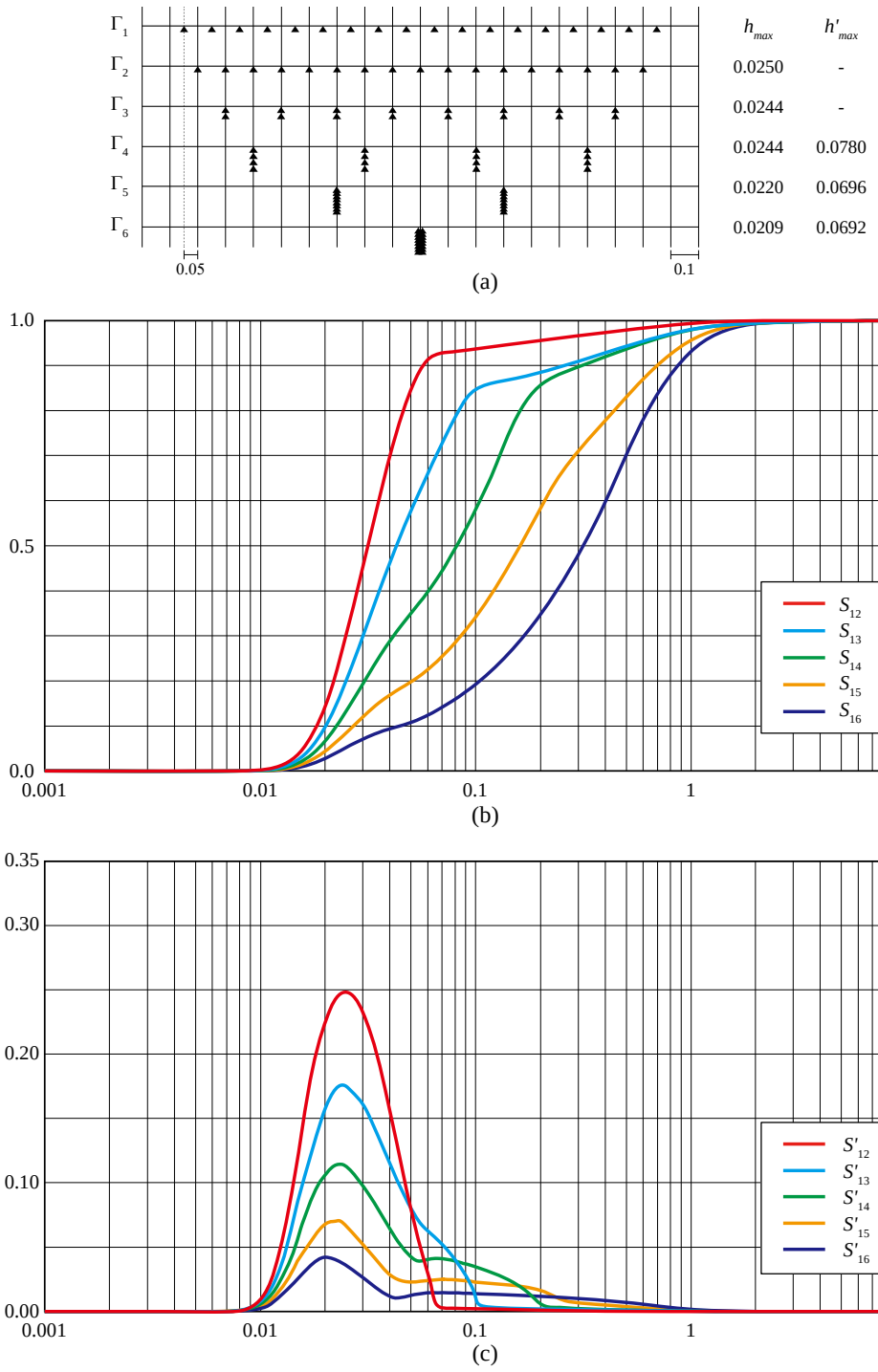


Figure A11 Measures S_{1k} and S'_{1k} in set Ψ_{11}.

Appendix A2

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 Appendix A1. Suppose point distributions Γ_P and Γ_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 jth points in distributions Γ_P and Γ_Q are denoted by P_j and Q_j, respectively. The distance between a point in Γ_1 and its nearest points in Γ_2 is ε. The locations of P_j and Q_j are represented as $2jw$-$\varepsilon/2$ and $2jw$+$\varepsilon/2$, respectively, as shown in Figure A12.

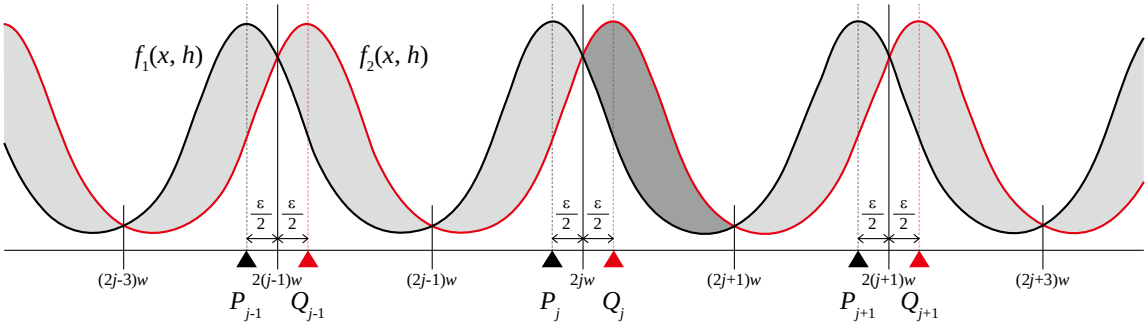


Figure A12 Distributions Γ_P and Γ_Q, and their interpreted surfaces.

The interpreted surfaces of Γ_P and Γ_Q in their standardized form are represented as

$$f_P(x,h) = \frac{1}{\sqrt{2\pi}nh}\sum_{k=1}^{\infty} e^{-\frac{\left\{x-\left(2kw-\frac{\varepsilon}{2}\right)\right\}^2}{2h^2}} \tag{13}$$

and

$$f_Q(x,h) = \frac{1}{\sqrt{2\pi}nh}\sum_{k=1}^{\infty} e^{-\frac{\left\{x-\left(2kw+\frac{\varepsilon}{2}\right)\right\}^2}{2h^2}} \tag{14}$$

respectively. The similarity measure between Γ_P and Γ_Q is

$$\begin{aligned} S_{PQ}(h) &= 1-\frac{1}{2}\int_{x=-\infty}^{\infty}\left|f_P(x,h)-f_Q(x,h)\right|dx \\ &= 1-\frac{1}{2\sqrt{2\pi}nh}\int_{x=-\infty}^{\infty}\left|\sum_{k=1}^{\infty}\left(e^{-\frac{\left\{x-\left(2kw-\frac{\varepsilon}{2}\right)\right\}^2}{2h^2}}-e^{-\frac{\left\{x-\left(2kw+\frac{\varepsilon}{2}\right)\right\}^2}{2h^2}}\right)\right|dx \end{aligned}. \tag{15}$$

The measure is represented as the area of light shades in Figure A12. Since n is very large, it is enough to consider only the dark-shaded area in the figure. The above equation then becomes

$$
\begin{aligned}
S_{PQ}(h) &= 1-\frac{2n}{2\sqrt{2\pi}nh}\int_{x=2jw}^{2jw+w}\sum_{k=1}^{\infty}\left(e^{-\frac{\left\{x-\left(2kw+\frac{\varepsilon}{2}\right)\right\}^2}{2h^2}}-e^{-\frac{\left\{x-\left(2kw-\frac{\varepsilon}{2}\right)\right\}^2}{2h^2}}\right)\mathrm{d}x \\
&= 1-\frac{1}{\sqrt{2\pi}h}\sum_{k=1}^{\infty}\int_{x=2jw}^{2jw+w}\left(e^{-\frac{\left\{x-\left(2kw-\frac{\varepsilon}{2}\right)\right\}^2}{2h^2}}-e^{-\frac{\left\{x-\left(2kw-\frac{\varepsilon}{2}\right)\right\}^2}{2h^2}}\right)\mathrm{d}x \\
&= 1-\frac{1}{\sqrt{2\pi}h}\sum_{k=1}^{\infty}\int_{x=0}^{w}\left(e^{-\frac{\left(x+2jw-2kw-\frac{\varepsilon}{2}\right)^2}{2h^2}}-e^{-\frac{\left(x+2jw-2kw+\frac{\varepsilon}{2}\right)^2}{2h^2}}\right)\mathrm{d}x \\
&= 1-\frac{1}{\sqrt{2\pi}h}\sum_{k=1}^{\infty}\int_{x=0}^{w}\left(e^{-\left(\frac{x+2jw-2kw-\frac{\varepsilon}{2}}{\sqrt{2}h}\right)^2}-e^{-\left(\frac{x+2jw-2kw+\frac{\varepsilon}{2}}{\sqrt{2}h}\right)^2}\right)\mathrm{d}x
\end{aligned}
\quad .
$$

(16)

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

$$
\begin{aligned}
\int_{x=0}^{w}e^{-x^2}\mathrm{d}x &= \frac{\sqrt{\pi}}{2}\left\{1+\frac{e^{-w^2}}{\sqrt{\pi}}\left(-\frac{1}{w}+\frac{1}{2w^3}-\frac{3}{4w^5}+\frac{15}{8w^7}+\cdots\right)\right\} \\
&\approx \frac{\sqrt{\pi}}{2}\left(1-\frac{e^{-w^2}}{\sqrt{\pi}w}\right)
\end{aligned}
$$

(17)

This approximation requires integration by substitution as follows.

$$
t = \frac{x+2jw-2kw-\frac{\varepsilon}{2}}{\sqrt{2}h}
$$

(18)

$$
dx = \sqrt{2}h\,dt
$$

(19)

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

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

(20)

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

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

(21)

Equation (16) now becomes

$$S_{PQ}(h) = 1 - \frac{\sqrt{2}h}{\sqrt{2\pi}h} \sum_{k=1}^{\infty} \left(\begin{array}{l} \int_{t=0}^{\frac{w+2jw-2kw-\frac{\varepsilon}{2}}{\sqrt{2}h}} e^{-t^2} dt - \int_{t=0}^{\frac{2jw-2kw-\frac{\varepsilon}{2}}{\sqrt{2}h}} e^{-t^2} dt \\ -\int_{t=0}^{\frac{w+2jw-2kw+\frac{\varepsilon}{2}}{\sqrt{2}h}} e^{-t^2} dt + \int_{t=0}^{\frac{2jw-2kw+\frac{\varepsilon}{2}}{\sqrt{2}h}} e^{-t^2} dt \end{array} \right)$$

$$= 1 - \frac{1}{\sqrt{\pi}} \sum_{k=1}^{\infty} \left(\begin{array}{l} \int_{t=0}^{\frac{(2j-2k+1)w-\frac{\varepsilon}{2}}{\sqrt{2}h}} e^{-t^2} dt - \int_{t=0}^{\frac{(2j-2k)w-\frac{\varepsilon}{2}}{\sqrt{2}h}} e^{-t^2} dt \\ -\int_{t=0}^{\frac{(2j-2k+1)w+\frac{\varepsilon}{2}}{\sqrt{2}h}} e^{-t^2} dt + \int_{t=0}^{\frac{(2j-2k)w+\frac{\varepsilon}{2}}{\sqrt{2}h}} e^{-t^2} dt \end{array} \right)$$

(22)

Using Approximation (17), we can rewrite each term inside the summation above as

$$\int_{t=0}^{\frac{(2j-2k+1)w+\frac{\varepsilon}{2}}{\sqrt{2}h}} e^{-t^2}dt \approx \frac{\sqrt{\pi}}{2}\left(1-\frac{e^{-\frac{\left\{(2j-2k+1)w+\frac{\varepsilon}{2}\right\}^2}{2h^2}}}{\sqrt{\pi}\,\frac{(2j-2k+1)w+\frac{\varepsilon}{2}}{\sqrt{2}h}}\right),$$

$$= \frac{\sqrt{\pi}}{2} - \frac{h}{\sqrt{2}\left\{(2j-2k+1)w+\frac{\varepsilon}{2}\right\}} e^{-\frac{\left\{(2j-2k+1)w+\frac{\varepsilon}{2}\right\}^2}{2h^2}} \tag{23}$$

$$\int_{t=0}^{\frac{(2j-2k)w-\frac{\varepsilon}{2}}{\sqrt{2}h}} e^{-t^2}dt \approx \frac{\sqrt{\pi}}{2} - \frac{h}{\sqrt{2}\left\{(2j-2k)w-\frac{\varepsilon}{2}\right\}} e^{-\frac{\left\{(2j-2k)w-\frac{\varepsilon}{2}\right\}^2}{2h^2}}, \tag{24}$$

$$\int_{t=0}^{\frac{(2j-2k+1)w+\frac{\varepsilon}{2}}{\sqrt{2}h}} e^{-t^2}dt \approx \frac{\sqrt{\pi}}{2} - \frac{h}{\sqrt{2}\left\{(2j-2k+1)w+\frac{\varepsilon}{2}\right\}} e^{-\frac{\left\{(2j-2k+1)w+\frac{\varepsilon}{2}\right\}^2}{2h^2}}, \tag{25}$$

and

$$\int_{t=0}^{\frac{(2j-2k)w+\frac{\varepsilon}{2}}{\sqrt{2}h}} e^{-t^2}dt \approx \frac{\sqrt{\pi}}{2} - \frac{h}{\sqrt{2}\left\{(2j-2k)w+\frac{\varepsilon}{2}\right\}} e^{-\frac{\left\{(2j-2k)w+\frac{\varepsilon}{2}\right\}^2}{2h^2}}. \tag{26}$$

Equation (22) then becomes

$$S_{PQ}(h) \approx 1 - \frac{h}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \left(\begin{array}{l} \dfrac{1}{(2j-2k+1)w-\frac{\varepsilon}{2}} e^{-\frac{\left\{(2j-2k+1)w-\frac{\varepsilon}{2}\right\}^2}{2h^2}} - \dfrac{1}{(2j-2k)w-\frac{\varepsilon}{2}} e^{-\frac{\left\{(2j-2k)w-\frac{\varepsilon}{2}\right\}^2}{2h^2}} \\ - \dfrac{1}{(2j-2k+1)w+\frac{\varepsilon}{2}} e^{-\frac{\left\{(2j-2k+1)w+\frac{\varepsilon}{2}\right\}^2}{2h^2}} + \dfrac{1}{(2j-2k)w+\frac{\varepsilon}{2}} e^{-\frac{\left\{(2j-2k)w+\frac{\varepsilon}{2}\right\}^2}{2h^2}} \end{array} \right) \tag{27}$$

We define τ_k^1, τ_k^2, τ_k^3, τ_k^4, and τ_k as

$$\tau_k^1 = \frac{1}{(2j-2k+1)w-\frac{\varepsilon}{2}} e^{-\frac{\left\{(2j-2k+1)w-\frac{\varepsilon}{2}\right\}^2}{2h^2}}$$

$$\tau_k^2 = \frac{1}{(2j-2k)w-\frac{\varepsilon}{2}} e^{-\frac{\left\{(2j-2k)w-\frac{\varepsilon}{2}\right\}^2}{2h^2}}$$

$$\tau_k^3 = \frac{1}{(2j-2k+1)w+\frac{\varepsilon}{2}} e^{-\frac{\left\{(2j-2k+1)w+\frac{\varepsilon}{2}\right\}^2}{2h^2}}$$

$$\tau_k^4 = \frac{1}{(2j-2k)w+\frac{\varepsilon}{2}} e^{-\frac{\left\{(2j-2k)w+\frac{\varepsilon}{2}\right\}^2}{2h^2}}$$

(28)

and

$$\tau_k = \tau_k^1 - \tau_k^2 - \tau_k^3 + \tau_k^4$$

(29)

Equation (27) becomes

$$S_{PQ}(h) \approx 1 - \frac{h}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \tau_k$$

(30)

We evaluate τ_k^1, τ_k^2, τ_k^3, and τ_k^4 from k=j-2 to j+2 as follows:

$$
\begin{aligned}
\tau_{j-2} &= \frac{e^{-\frac{\left(5w-\frac{\varepsilon}{2}\right)^2}{2h^2}}}{5-\varepsilon} - \frac{e^{-\frac{\left(4w-\frac{\varepsilon}{2}\right)^2}{2h^2}}}{4-\varepsilon} - \frac{e^{-\frac{\left(5w+\frac{\varepsilon}{2}\right)^2}{2h^2}}}{5+\varepsilon} + \frac{e^{-\frac{\left(4w+\frac{\varepsilon}{2}\right)^2}{2h^2}}}{4+\varepsilon} \\
\tau_{j-1} &= \frac{e^{-\frac{\left(3w-\frac{\varepsilon}{2}\right)^2}{2h^2}}}{3-\varepsilon} - \frac{e^{-\frac{\left(2w-\frac{\varepsilon}{2}\right)^2}{2h^2}}}{2-\varepsilon} - \frac{e^{-\frac{\left(3w+\frac{\varepsilon}{2}\right)^2}{2h^2}}}{3+\varepsilon} + \frac{e^{-\frac{\left(2w+\frac{\varepsilon}{2}\right)^2}{2h^2}}}{2+\varepsilon} \\
\tau_{j} &= \frac{e^{-\frac{\left(w-\frac{\varepsilon}{2}\right)^2}{2h^2}}}{1-\varepsilon} - \frac{e^{-\frac{\frac{\varepsilon^2}{4}}{2h^2}}}{-\varepsilon} - \frac{e^{-\frac{\left(w+\frac{\varepsilon}{2}\right)^2}{2h^2}}}{1+\varepsilon} + \frac{e^{-\frac{\frac{\varepsilon^2}{4}}{2h^2}}}{\varepsilon} \\
\tau_{j+1} &= \frac{e^{-\frac{\left(-w-\frac{\varepsilon}{2}\right)^2}{2h^2}}}{-1-\varepsilon} - \frac{e^{-\frac{\left(-2w-\frac{\varepsilon}{2}\right)^2}{2h^2}}}{-2-\varepsilon} - \frac{e^{-\frac{\left(-w+\frac{\varepsilon}{2}\right)^2}{2h^2}}}{-1+\varepsilon} + \frac{e^{-\frac{\left(-2w+\frac{\varepsilon}{2}\right)^2}{2h^2}}}{-2+\varepsilon} \\
\tau_{j+2} &= \frac{e^{-\frac{\left(-3w-\frac{\varepsilon}{2}\right)^2}{2h^2}}}{-3-\varepsilon} - \frac{e^{-\frac{\left(-4w-\frac{\varepsilon}{2}\right)^2}{2h^2}}}{-4-\varepsilon} - \frac{e^{-\frac{\left(-4w+\frac{\varepsilon}{2}\right)^2}{2h^2}}}{-3+\varepsilon} + \frac{e^{-\frac{\left(-3w+\frac{\varepsilon}{2}\right)^2}{2h^2}}}{-4+\varepsilon}
\end{aligned}
\tag{31}
$$

The above equations indicate

$$
\begin{aligned}
\tau^1_{j+k'} &= \frac{e^{-\frac{\left(-w-\frac{\varepsilon}{2}\right)^2}{2h^2}}}{-1-\varepsilon} = \frac{e^{-\frac{\left(w+\frac{\varepsilon}{2}\right)^2}{2h^2}}}{1+\varepsilon}, \\
&= \tau^3_{j+k'-1}
\end{aligned}
\tag{32}
$$

and similarly,

$$
\begin{aligned}
\tau^3_{j+k'} &= \tau^1_{j+k'-1} \\
\tau^2_{j+k'} &= \tau^4_{j+k'-2} \\
\tau^4_{j+k'} &= \tau^2_{j+k'-2}
\end{aligned}
\tag{33}
$$

Using the above equations, we rewrite Equation (30) as

$$
\begin{aligned}
S_{PQ}(h) &\approx 1 - \frac{h}{\sqrt{2\pi}w}\sum_{k=1}^{\infty}\tau_k \\
&= 1 - \frac{h}{\sqrt{2\pi}}\left\{\frac{2}{\varepsilon}e^{-\frac{\varepsilon^2}{8h^2}} - \tau^2_{2j} + \tau^4_{2j} + \sum_{k=2j+1}^{\infty}\tau_k\right\}
\end{aligned}
$$

(34)

Since variables τ_k^2, τ_k^4, and τ_k rapidly decrease with an increase of k, we can approximate the above equation as

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

(35)

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

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

(36)

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

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

(37)

Since ε, w, and h are all positive, the above equation holds only when

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

(38)

This equation indicates that $S'_{PQ}(h)$ shows its maximum when h is the half of the distance between points in Γ_P and their nearest points in Γ_Q.

Extending the above discussion, we can show that the value of h_{max} is proportional to the distance between points in Γ_Q and their nearest points in Γ_P in the case where point distributions exhibits a regular but non-uniform pattern such as seen in Set Ψ_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 A3

This appendix extends the discussion in the previous appendix to a more general case. Suppose distributions Γ_P and Γ_Q, each of which consists of n points that are arranged alternately as shown in Figure A13. The distributions are in the mutual nearest relationship such as seen in set Ψ_6 in Appendix A1, i.e., the nearest point in Γ_Q from P_i is Q_i, while the nearest point in Γ_P from Q_i is P_i. The locations of ith point in Γ_P and jth point in Γ_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 A13.

Let Φ- and Φ+ 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 Γ_P and Γ_Q in interval Φ:

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

(39)

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

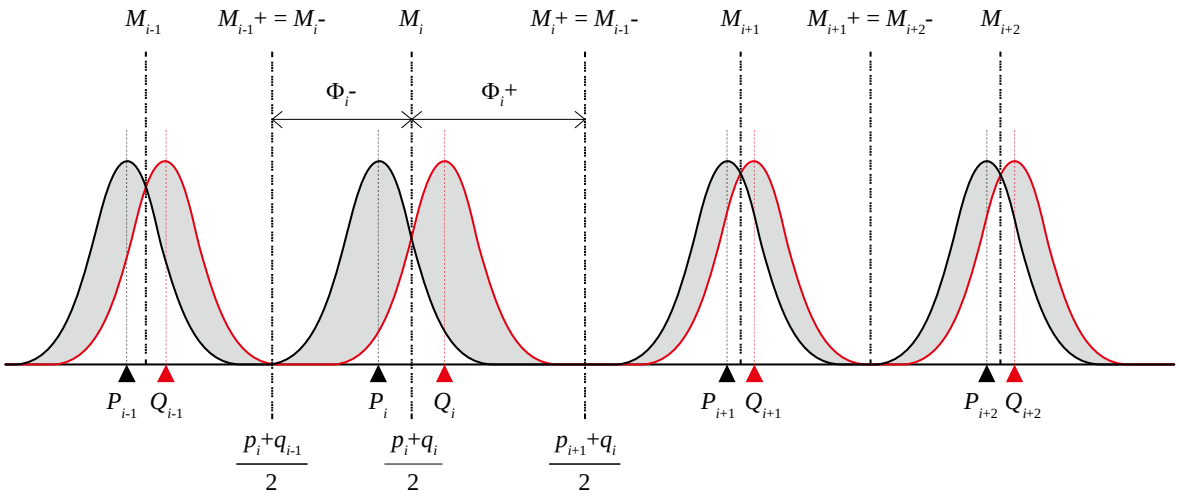


Figure A13 Distributions Γ_P and Γ_Q, and their interpreted surfaces.

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

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

(40)

and

$$x_U = \frac{p_i + q_i}{2},$$

(41)

respectively.

Equations from (13) to (15) and the above simulations that the similarity between Γ_P and Γ_Q evaluated at location x is primarily based on the points in the neighborhood of x. Equation (27) calculates

the summation of τ_k^1, τ_k^2, τ_k^3, and τ_k^4, each of which is a negative exponential function of $(j\text{-}k)^2 w$. Since they rapidly decrease with an increase of $(j\text{-}k)^2$, we can almost evaluate their summation from k=1 to infinity by the term when $j\text{-}k$=0. This implies that we can approximate $\gamma(h; \Phi\text{-})$ by the interpreted surfaces generated from P_i and Q_i:

$$\begin{aligned}
\gamma\left(h;\Phi_i-\right) &= \frac{1}{2}\int_{x=x_L}^{x_U}\left(f_P\left(x,h\right)-f_Q\left(x,h\right)\right)\mathrm{d}x \\
&= \frac{1}{2}\int_{x=x_L}^{x_U}\left\{\frac{1}{\sqrt{2\pi}nh}\sum_{i=1}^{n_P}e^{-\frac{(x-p_i)^2}{2h^2}}-\frac{1}{\sqrt{2\pi}nh}\sum_{j=1}^{n_Q}e^{-\frac{(x-q_i)^2}{2h^2}}\right\}\mathrm{d}x \\
&= \frac{1}{2\sqrt{2\pi}nh}\int_{x=x_L}^{x_U}\left\{\sum_{i=1}^{n}e^{-\frac{(x-p_i)^2}{2h^2}}-\sum_{j=1}^{n}e^{-\frac{(x-q_i)^2}{2h^2}}\right\}\mathrm{d}x \\
&\approx \frac{1}{2\sqrt{2\pi}nh}\int_{x=x_L}^{x_U}\left\{e^{-\frac{(x-p_i)^2}{2h^2}}-e^{-\frac{(x-q_i)^2}{2h^2}}\right\}\mathrm{d}x
\end{aligned} \tag{42}$$

Using approximation (17), we obtain

$$\begin{aligned}
\int_{x=x_L}^{x_U}e^{-\frac{(x-p_i)^2}{2h^2}}\mathrm{d}x &= \int_{x=0}^{x_U}e^{-\left(\frac{x-p_i}{\sqrt{2}h}\right)^2}\mathrm{d}x-\int_{x=0}^{x_L}e^{-\left(\frac{x-p_i}{\sqrt{2}h}\right)^2}\mathrm{d}x \\
&= \frac{\sqrt{2\pi}h}{2}\left(1-\frac{1}{\sqrt{\pi}\,\frac{x_U-p_i}{\sqrt{2}h}}e^{-\left(\frac{x_U-p_i}{\sqrt{2}h}\right)^2}\right)-\frac{\sqrt{2\pi}h}{2}\left(1-\frac{1}{\sqrt{\pi}\,\frac{x_L-p_i}{\sqrt{2}h}}e^{-\left(\frac{x_L-p_i}{\sqrt{2}h}\right)^2}\right). \\
&= h^2\left\{-\frac{1}{x_U-p_i}e^{-\left(\frac{x_U-p_i}{\sqrt{2}h}\right)^2}+\frac{1}{x_L-p_i}e^{-\left(\frac{x_L-p_i}{\sqrt{2}h}\right)^2}\right\}
\end{aligned} \tag{43}$$

and

$$\begin{aligned}
\int_{x=x_L}^{x_U}e^{-\frac{(x-q_i)^2}{2h^2}}\mathrm{d}x &= \int_{x=0}^{x_U}e^{-\left(\frac{x-q_i}{\sqrt{2}h}\right)^2}\mathrm{d}x-\int_{x=0}^{x_L}e^{-\left(\frac{x-q_i}{\sqrt{2}h}\right)^2}\mathrm{d}x \\
&= \frac{\sqrt{2\pi}h}{2}\left(1-\frac{1}{\sqrt{\pi}\,\frac{x_U-q_i}{\sqrt{2}h}}e^{-\left(\frac{x_U-q_i}{\sqrt{2}h}\right)^2}\right)-\frac{\sqrt{2\pi}h}{2}\left(1-\frac{1}{\sqrt{\pi}\,\frac{x_L-q_i}{\sqrt{2}h}}e^{-\left(\frac{x_L-q_i}{\sqrt{2}h}\right)^2}\right). \\
&= h^2\left\{-\frac{1}{x_U-q_i}e^{-\left(\frac{x_U-q_i}{\sqrt{2}h}\right)^2}+\frac{1}{x_L-q_i}e^{-\left(\frac{x_L-q_i}{\sqrt{2}h}\right)^2}\right\}
\end{aligned}$$

(44)

Since P_i and Q_i are both closer to the middle point between the boundaries of Φ than the boundaries, we can say

$$\left(\frac{x_L - p_i}{\sqrt{2}h}\right)^2 \gg \left(\frac{x_U - p_i}{\sqrt{2}h}\right)^2$$

(45)

and

$$\left(\frac{x_L - q_i}{\sqrt{2}h}\right)^2 \gg \left(\frac{x_U - q_i}{\sqrt{2}h}\right)^2.$$

(46)

This yields

$$\int_{x=x_L}^{x_U} e^{-\frac{(x-p_i)^2}{2h^2}} \mathrm{d}x = h^2 \left\{ -\frac{1}{x_U - p_i} e^{-\left(\frac{x_U - p_i}{\sqrt{2}h}\right)^2} + \frac{1}{x_L - p_i} e^{-\left(\frac{x_L - p_i}{\sqrt{2}h}\right)^2} \right\}.$$
$$\approx -\frac{h^2}{x_U - p_i} e^{-\left(\frac{x_U - p_i}{\sqrt{2}h}\right)^2}$$

(47)

and

$$\int_{x=x_L}^{x_U} e^{-\frac{(x-q_i)^2}{2h^2}} \mathrm{d}x = h^2 \left\{ -\frac{1}{x_U - q_i} e^{-\left(\frac{x_U - q_i}{\sqrt{2}h}\right)^2} + \frac{1}{x_L - q_i} e^{-\left(\frac{x_L - q_i}{\sqrt{2}h}\right)^2} \right\}.$$
$$\approx -\frac{h^2}{x_U - q_i} e^{-\left(\frac{x_U - q_i}{\sqrt{2}h}\right)^2}$$

(48)

Consequently,

$$\int_{x=x_L}^{x_U} e^{-\frac{(x-p_i)^2}{2h^2}} \mathrm{d}x - \int_{x=x_L}^{x_U} e^{-\frac{(x-q_i)^2}{2h^2}} \mathrm{d}x \approx h^2 \left\{ -\frac{1}{x_U - p_i} e^{-\left(\frac{x_U - p_i}{\sqrt{2}h}\right)^2} + \frac{1}{x_U - q_i} e^{-\left(\frac{x_U - q_i}{\sqrt{2}h}\right)^2} \right\}.$$

(49)

and

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

(50)

Since

$$x_U - p_i = -x_U + q_i \,,$$

(51)

We rewrite Equation (50) as

$$\begin{aligned}
\gamma\left(h;\Phi_i -\right) &\approx \frac{h}{2\sqrt{2\pi}n}\left[\frac{1}{x_U - p_i}e^{-\left(\frac{x_U - p_i}{\sqrt{2}h}\right)^2} - \frac{1}{x_U - q_i}e^{-\left(\frac{x_U - q_i}{\sqrt{2}h}\right)^2}\right] \\
&= \frac{h}{\sqrt{2\pi}n\left(x_U - p_i\right)}e^{-\left(\frac{x_U - p_i}{\sqrt{2}h}\right)^2} \\
&= \frac{h}{\sqrt{2\pi}n\left(\frac{p_i + q_i}{2} - p_i\right)}e^{-\left(\frac{\frac{p_i + q_i}{2} - p_i}{\sqrt{2}h}\right)^2} \\
&= \frac{h}{\sqrt{2\pi}n\varepsilon_i}e^{-\left(\frac{\varepsilon_i}{2\sqrt{2}h}\right)^2}
\end{aligned} \,,$$

(52)

where ε_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

$$x_L = \frac{p_i + q_i}{2} \,.$$

(53)

and

$$x_U = \frac{p_{i+1} + q_i}{2} \,,$$

(54)

respectively.

We can employ Equations from (42) to (44) with a slight modification. Inequalities (45) and (46) now become

$$\left(\frac{x_L - p_i}{\sqrt{2}h}\right)^2 \ll \left(\frac{x_U - p_i}{\sqrt{2}h}\right)^2$$

(55)

and

$$\left(\frac{x_L - q_i}{\sqrt{2}h}\right)^2 \ll \left(\frac{x_U - q_i}{\sqrt{2}h}\right)^2 .$$

(56)

Using the above inequalities, we obtain

$$\int_{x=x_L}^{x_U} e^{-\frac{(x-p_i)^2}{2h^2}} \mathrm{d}x \approx \frac{h^2}{x_L - p_i} e^{-\left(\frac{x_L - p_i}{\sqrt{2}h}\right)^2},$$

(57)

$$\int_{x=x_L}^{x_U} e^{-\frac{(x-q_i)^2}{2h^2}} \mathrm{d}x \approx \frac{h^2}{x_L - q_i} e^{-\left(\frac{x_L - q_i}{\sqrt{2}h}\right)^2},$$

(58)

and

$$-\int_{x=x_L}^{x_U} e^{-\frac{(x-p_i)^2}{2h^2}} \mathrm{d}x + \int_{x=x_L}^{x_U} e^{-\frac{(x-q_i)^2}{2h^2}} \mathrm{d}x \approx h^2 \left\{ -\frac{1}{x_L - p_i} e^{-\left(\frac{x_L - p_i}{\sqrt{2}h}\right)^2} + \frac{1}{x_L - q_i} e^{-\left(\frac{x_L - q_i}{\sqrt{2}h}\right)^2} \right\}.$$

(59)

Consequently,

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

(60)

Adding Equation (52) to (60), we obtain the difference between Γ_P and Γ_Q in interval $\Phi(M(i, j\text{-}1), M(i\text{+}1, j))$:

$$\gamma\left(h;\Phi_i-\right)+\gamma\left(h;\Phi_i+\right)=\frac{\sqrt{2}h}{2\sqrt{\pi}n\varepsilon_i}e^{-\left(\frac{\varepsilon_i}{2\sqrt{2}h}\right)^2}\ .$$

(61)

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

$$\begin{aligned}S_{PQ}\left(h\right)&=1-\sum_{i=1}^{n}\left\{\gamma\left(h;\Phi_i-\right)+\gamma\left(h;\Phi_i+\right)\right\}\\&=1-\sum_{i=1}^{n}e^{-\left(\frac{\varepsilon_i}{\sqrt{2}h}\right)^2}\\&=1-\frac{2\sqrt{2}h}{\sqrt{\pi}n}\sum_{i=1}^{n}\frac{1}{\varepsilon_i}e^{-\left(\frac{\varepsilon_i}{2\sqrt{2}h}\right)^2}\end{aligned}\ .$$

(62)

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

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

(63)

and

$$S''_{PQ}\left(h\right)=\frac{2\sqrt{2}}{\sqrt{\pi}nh^5}\sum_{i=1}^{n}\varepsilon_i\left(\frac{\varepsilon_i^2}{4}-h^2\right)e^{-\left(\frac{\varepsilon_i}{2\sqrt{2}h}\right)^2}\ ,$$

(64)

respectively.

If ε_i is equal for any i, $S''_{PQ}(h)$ becomes zero when $h=\varepsilon_i/2$, i.e., a peak of $S'_{PQ}(h)$ is observed at $h=\varepsilon_i/2$. This holds not only when points in Γ_P and Γ_Q are both uniformly distributed but also when Γ_P and Γ_Q show a regular but non-uniform pattern such as seen in Set Ψ_1.

Using Equations from (62) to (64), we evaluate the location of peaks of $S'_{PQ}(h)$ when ε_i takes one of two different values as seen in Set Ψ_6. We assume that $\varepsilon_i=0.01$ for half of the points while $\varepsilon_i=\varepsilon$ for the other half. Solving $S''_{PQ}(h)=0$ numerically, we obtain h_{max} and h'_{max} as shown in Figure A14. The figure also shows two functions $h=\varepsilon/2$ and $h=(0.01+\varepsilon)/4$ for comparison purposes. The result is supportive of our earlier observation and presumption in Set Ψ_6. The highest peak of $S'_{PQ}(h)$ is observed between $h=0.005$ and $h=\varepsilon/2$ in any case. Our presumption of potential peaks at $h=0.005$ and $h=\varepsilon/2$ still seems to hold since the former is represented as h'_{max} while the latter corresponds to h_{max} in Figure A14. When ε is smaller than to 0.036, only a single peak appears at $h=h'_{max}$. Two differences from the results obtained for Set Ψ_6 are that peaks at small h found in Set Ψ_6 do not appear at $h=0.005$ and that h_{max} is larger than

(0.01+ε)/4. 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.

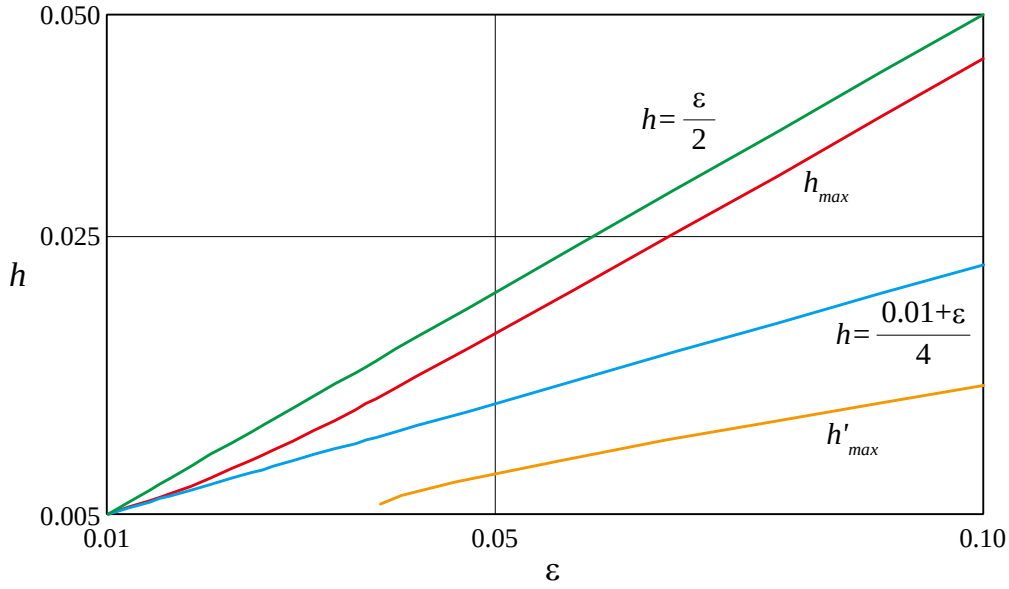


Figure A14 The relationship between ε and the value of h that gives the peaks of $S'_{PQ}(h)$.