Discussion Paper Series
No. 98R

Department of Urban Engineering
University of Tokyo

Analysis of the relation among point distributions on a discrete space

Yukio Sadahiro
Department of Urban Engineering, University of Tokyo

Analysis of the relation among point distributions on a discrete space

Yukio Sadahiro
Department of Urban Engineering, University of Tokyo

Abstract

In urban areas numerous types of spatial objects are distributed in a close mutual relation. Drug stores and newspaper stalls gather around railway stations while fastfood restaurants are competing with each other in downtown area. This paper discusses such relation among spatial objects usually represented as points in GIS. A focus is on the relation among point distributions on a discrete space. Numerical measures are proposed to describe the properties of point distributions. Geographical and graphical representations visualize the relation among distributions. The latter also provides a means of classifying the point distributions. The method proposed is applied to school location planning in Japan. The result reveals the properties of the method and provides empirical findings.

1. Introduction

Spatial pattern of point distributions has been drawing much attention in geography, ecology, epidemiology, and other academic fields interested in spatial phenomena. There have been developed numerous methods of analyzing point patterns, from visual tools to statistical techniques.

A single set of points is often analyzed by statistical methods such as the nearest neighbor distance, K-function, quadrat counts, and so forth (Upton and Fingleton, 1985; Cressie, 1993; Bailey and Gatrell, 1995). They have been extended for almost half a century to treat a wider variety of situations. Examples include spatiotemporal analysis (Knox, 1964; Mantel, 1967; Openshaw, 1994; Jacquez, 1996; Fortin and Dale, 2005), local pattern analysis (Openshaw et al., 1987; Besag and Newell, 1991; Kulldorff, M., and Nagawalla, 1994; Ord & Getis, 1995; Kulldorff, 1997; Glaz et al. 2001; Lai et al., 2008), point pattern analysis on a network (Okabe and Yamada, 2001; Okabe et al., 2006; Yamada and Thill, 2007), and so forth.

One important extension of traditional point pattern analysis is to analyze the relation among multiple sets of points. Similarity in distributions suggests a close relation between the points, typically as seen in maps of John Snow’s London (Department of Epidemiology, UCLA 2009). For two sets of points statistical methods are available including cross K-function (Ripley, 1981) and quadrat counts (Bailey and

Gatrell, 1995).

On the other hand, there are few methods of analyzing the relation among more than two sets of points. A possible option is to apply a method for two sets of points to all the pairs of point sets and integrate the individual results toward a general conclusion. However, this approach is clearly inefficient and cannot treat complex relation such as spatial interaction among plants and insects in a field.

To meet the demand, this paper develops a new method of analyzing multiple distributions of points. A focus is on points distributed on a discrete space, that is, a limited number of locations. Urban facilities, both public and private ones, often fall into this category, where land blocks are firmly predetermined by a land ownership system. Since they are hard to change, new facilities choose their location from a limited set of available lots.

Methods are proposed in the following five sections. Section 2 proposes numerical measures to describe the properties of point distributions. Section 3 defines basic relations between points and those between point distributions. Using the relations, Sections 4 and 5 propose geographical and graphical representations of the relation among point distributions. Section 6 extends the method proposed to cover a wider range of situations. Section 7 applies the method to school location planning in Japan. Section 8 summarizes the conclusions with a discussion.

2. Numerical measures of the properties of point distributions

Suppose a set of locations $\Lambda=\{Q_i,\ i\in\mathfrak{M}\}$ in region S, where $\mathfrak{M}=\{1, 2, ..., M\}$. A set of point distributions $\Omega=\{\Omega_i,\ i\in\mathfrak{N}\}$ ($\mathfrak{N}=\{1, 2, ..., N\}$) are defined on Λ, each of which consists of m_i points denoted by their locations $\{Q_{i1}, Q_{i2}, \ldots, Q_{im_i}\}$.

The mean and variance of the number of points are the most basic measures that describe the properties of point distributions. Besides these statistics, *diversity* is useful to evaluate the structural difference among distributions:

$$\gamma(\Omega)=1-\frac{Nn\left(I(\Omega)\right)}{\sum_i m_i}, \tag{1}$$

where $I(\Omega)$ and $n(I(\Omega))$ are the intersection of Ω and the number of elements in $I(\Omega)$, respectively. This is the ratio of points not shared by all the distributions. It shows a small value if all the distributions consist of similar points and increases with the variation of elements in the distributions.

Since diversity is defined based on $I(\Omega)$, it may not work for numerous

distributions of points where $I(\Omega)$ often becomes empty. To treat such cases, *weighted diversity* relaxes the definition of $\gamma(\Omega)$:

$$\gamma_W(\Omega) = 1 - \frac{N}{\sum_i m_i} \sum_j w(\Omega, Q_j), \tag{2}$$

where $w(\Omega, Q_j)$ is a measure of sharing Q_j in Ω. A possible definition of $w(\Omega, Q_j)$ is

$$w(\Omega, Q_j) = \left\{ \frac{1}{N} \sum_i n(\Omega_i \cap Q_j) \right\}^2, \tag{3}$$

where the right side is the square of the ratio of distributions in Ω that contains Q_j.

3. Relation between point distributions and that between points

This section defines basic relations between point distributions and those between points. They serve as a basis for analysis and visualization of more complicated relation among point distributions.

3.1 Relation between a pair of point distributions

Two distributions Ω_i and Ω_j are called *equal* if $\Omega_i=\Omega_j$, that is, if they consist of the same set of locations (Figure 1a). Two distributions are *inclusive* (Figure 1b) if one set is a subset of the other. If $\Omega_i \subset \Omega_j$, Ω_j is a higher level distribution of Ω_i, while Ω_i is a lower level distribution of Ω_j. If $\Omega_i \cap \Omega_j = \varnothing$, they are called *exclusive* (Figure 1c). They do not share any location in their distributions. Two distributions are *overlapping* (Figure 1d) if they are neither inclusive nor overlapping. These four relations are mutually exclusive.

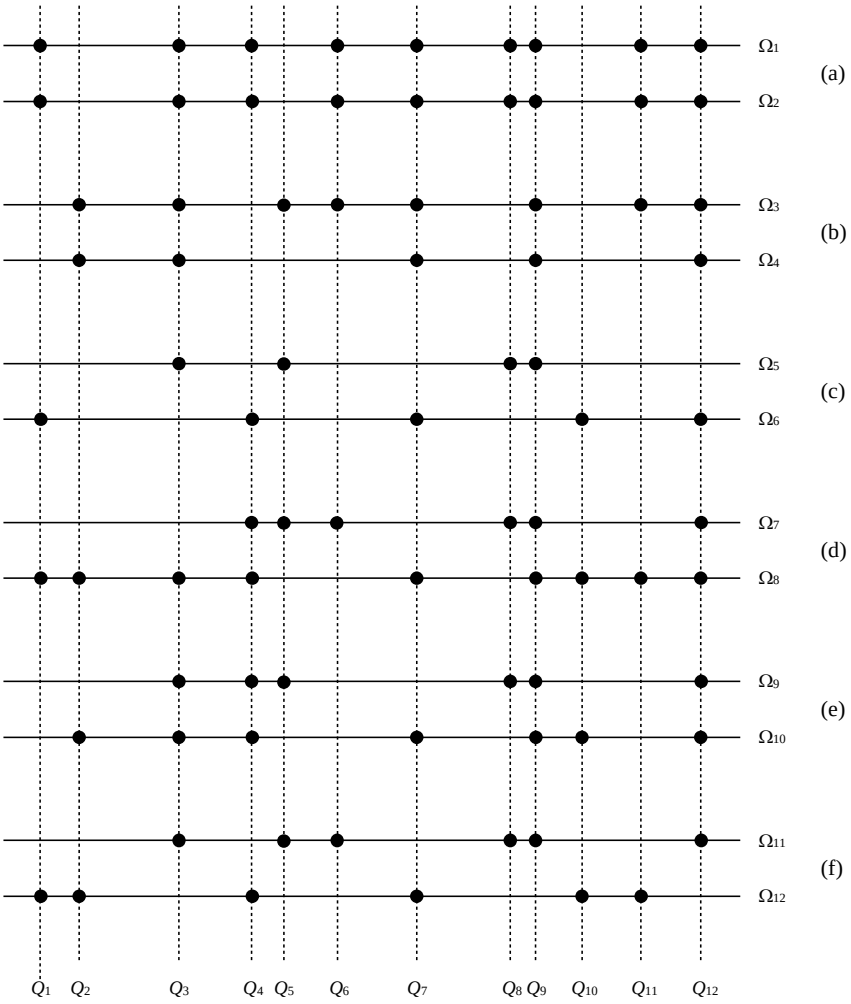

Figure 1 Relations between a pair of point distributions. (a) Equal, (b) inclusive, (c) exclusive, and (d) overlapping distributions. These four relations are mutually exclusive. Distributions in (d) are also complete, while those in (e) are incomplete. Distributions in (f) are called complementary because they are both complete and exclusive

Another set of relations is defined by considering Λ, the universal set of locations. Two sets Ω_i and Ω_j are *complete* if $\Omega_i \cup \Omega_j = \Lambda$ (Figure 1d), while they are *incomplete* if $\Omega_i \cup \Omega_j \neq \Lambda$ (Figure 1e). Since the two relations are independent of the previous four relations, eight relations are defined between Ω_i and Ω_j. For instance, Ω_i and Ω_j are complete and overlapping if $\Omega_i \cup \Omega_j = \Lambda$ and $\Omega_i \cap \Omega_j \neq \varnothing$ (Figure 1e). Two distributions are called *complementary* if they are complete and exclusive (Figure 1f).

3.2 Relation between a pair of points

The above discussion also applies to the relations between a pair of points.

Points Q_i and Q_j are called *equal* (Figure 2a) if any distribution in Ω contains either both or none of them. If every distribution containing Q_j also contains Q_i or vice versa, Q_i and Q_j are called *inclusive* (Figure 2b). In the former case Q_i is a higher level point of Q_j, while Q_i is a lower level point of Q_j in the latter case. Points Q_i and Q_j are *exclusive* (Figure 2c) if they are not contained in the same distribution simultaneously. If Q_i and Q_j are neither equal, inclusive, nor exclusive, they are called *overlapping* (Figure 2d).

Consideration of the universal set of distributions Ω gives another independent set of relations. Points Q_i and Q_j are *complete* (Figure 2d) if at least Q_i or Q_j is contained in any distribution. On the other hand, points Q_i and Q_j are *incomplete* (Figure 2e) if a distribution exists containing neither Q_i nor Q_j. As well as point distributions, points can be described by a pair of independent relations such as incomplete and overlapping (Figure 2e). If points Q_i and Q_j are complete and exclusive, they are called *complementary* (Figure 2f).

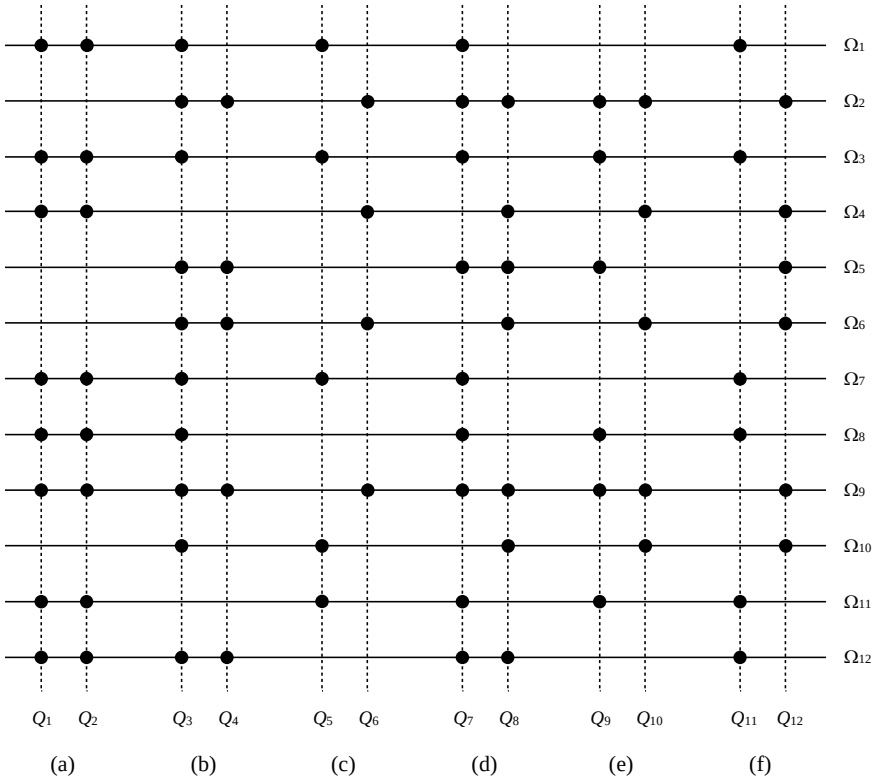

Figure 2 Relation between a pair of points. (a) Equal, (b) inclusive, (c) exclusive, and (d) overlapping points. Points in (d) are also complete, while points in (e) are incomplete. Points in (f) are called complementary because they are both complete and exclusive.

3.3 Relations among more than two sets of elements

Both the relations between point distributions and those between points can be further extended to the case of more than two sets of elements. If every pair of point distributions in Ω is consistently equal, inclusive, exclusive or overlapping, the set Ω is called by the name of the relation. For instance, Ω is called *inclusive* if every pair of point distributions is inclusive. It is also often called *hierarchical*. Set Ω is called *composite* if more than one relation are found in Ω.

Extension of complete and incomplete relations is rather different. Set Ω is called *complete* if the union of point distributions is the universal set of locations Λ; if not, they are *incomplete*. A complete and exclusive set is called *complementary*.

Extension is similarly possible for a set of points. A few examples will be shown in the next section.

4. Geographical representation of the relation among point distributions

Using the relations defined above, this section proposes a geographical representation of the relation among point distributions.

A *K-core* is a complete set of *K* points any smaller subset of which is not complete. For instance, Q_i is a 1-core if it is shared by all the distributions. Points Q_i and Q_j form a 2-core if they are complete and neither is 1-core (Figure 3). Point Q_i, Q_j and Q_k are a 3-core if any two of the three is not complete.

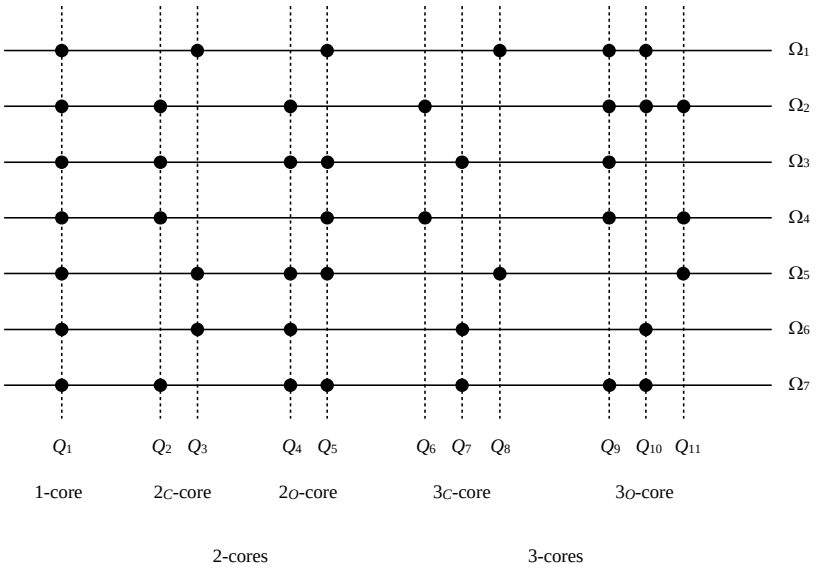

Figure 3 1-, 2- and 3-cores in Λ.

A complementary *K*-core is called a K_C*-core*. Ordinary *K*-cores are called

K_O-cores to distinguish them from K_C-cores. In a 3_C-core every distribution has exactly one point while a 3_O-core contains at least one point shared by more than one distribution. Since a *K*-core does not contain smaller *K*-cores, a *K*-core contains at least one point shared by only one distribution.

K-cores indicate the similarity in location among point distributions. Therefore, as shown later, mapping *K*-cores is helpful for understanding the outline and structure of Ω. However, maps often become complicated due to their overlapping nature. Since *K*-cores are permitted to share the same points, they may be visualized with too much overlapping.

To resolve this problem, three principles are proposed in choosing *K*-cores to visualize. Each principle evaluates *K*-cores in a different aspect: 1) simplicity of representation, 2) closeness of relation, and 3) amount of information.

The first principle claims that *K*-cores of fewer points should be chosen earlier. Simpler *K*-cores reveal the structure of point distributions more clearly and effectively. 1-cores should be chosen first, followed by 2- and 3-cores.

The second principle is that points in a closer relation should be chosen before those in a weaker relation. The first law of geography “near things are more related than distant things (Tobler, 1970)” naturally leads us to choose clustered points before dispersed ones. Proximity of points is measured by, for instance, the radius of their circumcircle.

The third principle claims that *K*-cores should be chosen in such a way that they convey as much information about point distributions as possible. The amount of information is measured by a decrease in entropy per point. Suppose two points Q_i and Q_j in Ω, a set of *N* point distributions. If no information is available about whether the points are contained in each distribution in Ω, 4^N cases can equally occur. Entropy of this situation is

$$-\log_2 \frac{1}{4^N} = 2N \,. \tag{4}$$

If Q_i and Q_j are known to be a 2-core, the number of possible cases reduces to 3^N-(2^{N+1}-1) (subtraction of 2^{N+1}-1 represents that a 2-core does not contain 1-cores). For a large *N*, entropy reduces to

$$-\log_2 \frac{1}{3^N - \left(2^N - 1\right)} \approx \log_2 3N \,. \approx 1.585N \tag{5}$$

Consequently, average amount of information per point is

$$J\left(2\text{-core}\right)=\frac{2-1.585}{2}N\ ,$$
$$\approx 0.208N$$

(6)

where J(K-core) is the average amount of information per point given by indicating a K-core.

If Q_i and Q_j are a 2_C-core, the entropy reduces accordingly to 2^N-2. Consequently,

$$J\left(2_C\text{-core}\right)=\frac{1}{2}\left(2N-\log_2\frac{1}{2^N-2}\right).$$
$$\approx 0.500N$$

(7)

The amount of information given by a 1-core is

$$J\left(1\text{-core}\right)=-\log_2\frac{1}{2^N}-\left(\log_2\frac{1}{1}\right)$$
$$=N\ .$$

(8)

For 3_C- and 3_O-cores, the average information is approximately $0.472N$ and $0.066N$, respectively (for details, see Sadahiro (2009)). These measures give us a means of ordering K-cores to visualize. If 2_C- and 2-cores are overlapping, the former should be visualized because it conveys more information. If 3_C-, 2_C- and 2_O-cores share a single point, the 2_C-core should be chosen while the others are omitted. The priority order of 1-, 2- and 3-cores is

$$1\text{-core}\succ 2_C\text{-core}\succ 3_C\text{-core}\succ 2_O\text{-core}\succ 3_O\text{-core}.$$

(9)

Note that the three principles are not strict rules to obey. Since they are local rather than global principles, they only help us to choose one K-core from a small set of K-cores. In addition, the principles are not comparable with each other. Consequently, a general principle is indispensable that should be determined by analysts in light of the objective of analysis. One option is to choose K-cores that convey the largest amount of information without overlapping. Since it is a combinatorial problem, heuristic methods would be effective to obtain a reasonable result. Section 7 will illustrate another example of implementation.

5. Structural representation of the relation among point distributions

The previous section has proposed a method of visualizing the relation among

point distributions in a geographical dimension. This section shifts the focus to the structural aspect of the relation among point distributions.

5.1 Graph representation of the relation among point distributions

Suppose the power set of Λ, denoted by $\Psi=\{\Psi_1, \Psi_2, ..., \Psi_P\}$. This contains all the point distributions in Ω. Since they are defined as sets of locations, Boolean operations such as intersection, union, and complement can be applied. Consequently, the algebraic structure $(\Psi, \cap, \cup)$ is a lattice, where the least and greatest elements are $\varnothing$ and Λ, respectively.

A lattice as a partially ordered set can be visualized as a Hasse diagram (Davey and Priestley, 2002; Pemmaraju and Skiena, 2003). Hasse diagram is a simple representation of a partially ordered set that permits us to evaluate the relation among elements in Ψ. In the Hasse diagram, nodes and links represent point distributions in Ψ and inclusion relation between them (Figure 4), respectively. Binary operations applied to point distributions are given by tracing links either upward or downward from the point distributions. The intersection and union of Ψ_i and Ψ_j are the point distributions at the lowest and highest levels connected either directly or indirectly to both Ψ_i and Ψ_j , respectively. In Figure 4, for instance, union of $\{Q_1, Q_2\}$ and $\{Q_2, Q_4\}$ is $\{Q_1, Q_2, Q_4\}$, the point distribution of the lowest level connected to both $\{Q_1, Q_2\}$ and $\{Q_2, Q_4\}$.

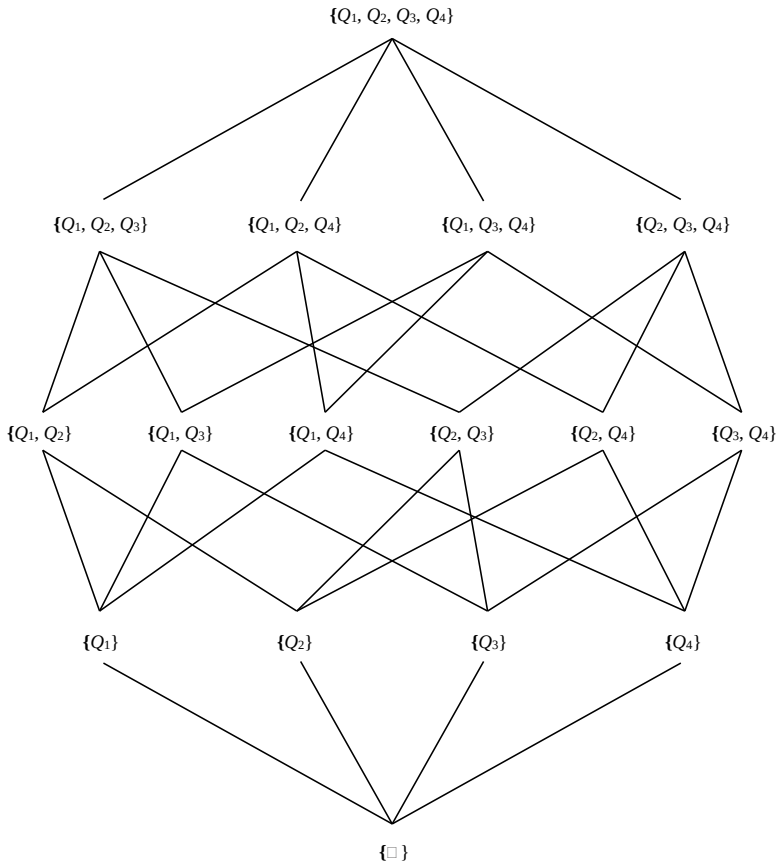

Figure 4 Hasse diagram of the power set of $\{Q_1, Q_2, Q_3, Q_4\}$.

Hasse diagram can also be used to visualize the closeness of relation between point distributions. To this end, two distances are defined between Ψ_i and Ψ_j.

Size distance is the difference in the number of elements in Ψ_i and Ψ_j:

$$d_S\left(\Psi_i, \Psi_j\right) = \left| n\left(\Psi_i\right) - n\left(\Psi_j\right) \right|. \tag{10}$$

It is zero if two distributions consist of the same number of points.

Inclusion distance evaluates the separation from inclusive relation. Suppose four distributions $\Psi_1=\{Q_1, Q_2, Q_3, Q_4\}$, $\Psi_2=\{Q_1, Q_2, Q_3\}$, $\Psi_3=\{Q_1, Q_2, Q_5\}$, and $\Psi_4=\{Q_1, Q_5, Q_6\}$. Distribution Ψ_1 contains all the three elements in Ψ_2, two of three elements in Ψ_3, and only one element in Ψ_4. The relation between Ψ_1 and Ψ_2 is inclusive while Ψ_1 and Ψ_3 are overlapping. The relation between Ψ_1 and Ψ_4 is also overlapping but the relation is weaker than that between Ψ_1 and Ψ_3. *Inclusion distance* evaluates such closeness of relation between distributions:

$$d_I\left(\Psi_i,\Psi_j\right)=n\left(\Psi_j\cup\Psi_j\right)-\max\left(n\left(\Psi_i\right),n\left(\Psi_j\right)\right)$$
$$=\min\left(n\left(\Psi_i\right),n\left(\Psi_j\right)\right)-n\left(\Psi_j\cap\Psi_j\right) \quad (11)$$

It is zero if Ψ_i and Ψ_j are in inclusion relation, and increases as their intersection becomes smaller.

Taking $n(\Psi_i)$ as the vertical axis, Hasse diagram tells us the closeness of relation between two distributions (Figure 5). Inclusion distance $d_I(\Psi_i, \Psi_j)$ is represented as the length of the shorter link connected to $\Psi_i\cup\Psi_j$. If Ψ_i and Ψ_j are inclusive, $d_I(\Psi_i, \Psi_j)=0$ and thus Ψ_i and Ψ_j are connected directly by a single link. If the relation is mostly inclusive, the shorter link becomes so short that Ψ_i and Ψ_j look like connected directly by a single link. Size distance is the difference in length of two links connected to $\Psi_i\cup\Psi_j$. If Ψ_i and Ψ_j consist of the same number of elements, they are connected indirectly by two links of the same length. Four links connecting Ψ_i, Ψ_j, $\Psi_i\cup\Psi_j$, and $\Psi_i\cap\Psi_j$ form a rhombus.

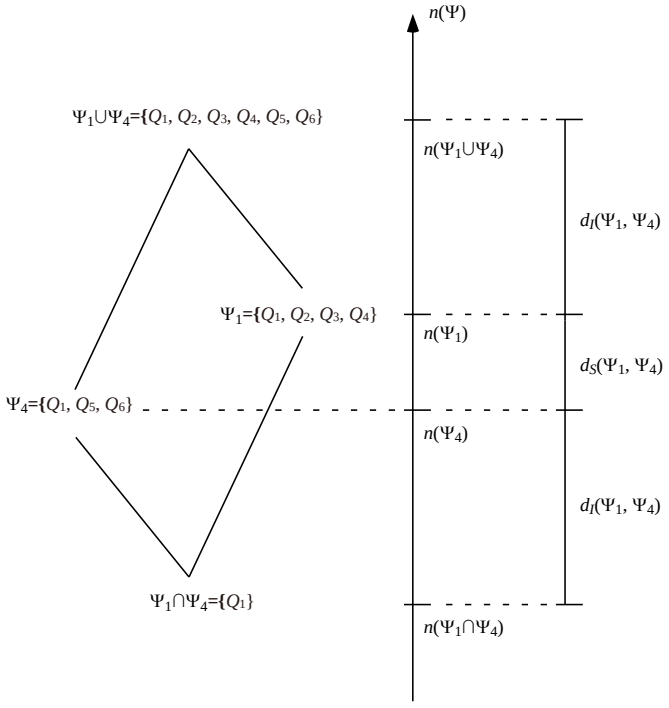

Figure 5 Relationship among the number of elements in sets, distance measures and binary operations.

5.2 Tree representation of the relation among point distributions

Though Hasse diagram is a useful tool for visualizing the relation among point distributions, it is not effective for a large set of distributions because it becomes much complicated. Hasse diagram is complemented by *intersection tree*, a visual representation of a stepwise process that groups point distributions into one according to the similarity between distributions (see also Sadahiro and Sasaya (2008)). It is constructed in a way similar to hierarchical methods in cluster analysis. Similarity between a pair of point distributions is evaluated by a measure based on the size and inclusion distances such as $d(\Omega_i, \Omega_j)=d_S(\Omega_i, \Omega_j)+d_I(\Omega_i, \Omega_j)$ and $(d_S(\Omega_i, \Omega_j)+d_I(\Omega_i, \Omega_j))/n(\Omega_i \cap \Omega_j)$. The measure is calculated for all the pairs of point distributions in Ω, from which the most similar pair is chosen. They are replaced by their intersection so that the elements in Ω decrease by one. This process continues until Ω consists of only one element. Figure 6 illustrates an intersection tree constructed from three distributions of points. The tree grows downward from original distributions.

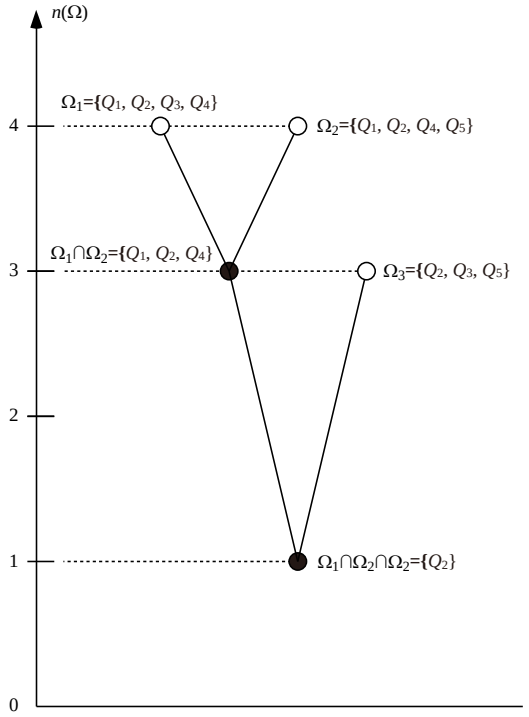

Figure 6 An intersection tree constructed from Ω_1, Ω_2 and Ω_3. White and black circles indicate original distributions and intersections replacing the most similar pair of distributions, respectively. Links represent intersection operation applied to point distributions. The similarity between distributions is measured by $d_S(\Omega_i, \Omega_j)+d_I(\Omega_i, \Omega_j)$. Distributions Ω_1 and Ω_2 are chosen first to generate $\Omega_1 \cap \Omega_2$, and then Ω_3 and $\Omega_1 \cap \Omega_2$ yield their intersection $\Omega_1 \cap \Omega_2 \cap \Omega_3$.

Using union instead of intersection yields another tree called a *union tree*. It is generated by replacing the most similar pair of point distributions with their union. A union tree extends upward from original distributions.

Both intersection and union trees consist of at most N-1 links. While they inherit the properties of Hasse diagram, they are not so complicated as to allow us intuitive understanding of the relation among point distributions.

Intersection and union trees can also be used to classify point distributions. Their partial trees naturally define groups of similar point distributions. Note, however, that intersection and union trees are different from dendrograms used in cluster analysis. The vertical axis of the trees indicates $d_S(\Omega_i, \Omega_j)$, which is not necessarily used as a distance measure in clustering process. Since the trees do not directly indicate the order of clustering, the process of classification should be visualized by using dendrograms.

6. Extension of the method: smoothing of point distributions

The theory of the method proposed heavily relies on the concept of intersection and union of point distributions. This may cause a practical problem when a large number of distributions are analyzed because their intersection often becomes empty. To resolve this problem, smoothing operation is applied to point distributions that transforms points into a surface. Surface function of point distribution Ω_i at Q_j is defined as

$$s\left(\Omega_i, Q_j\right) = \frac{\sum_k p\left(Q_j, Q_k\right) n\left(I\left(\Omega_i, Q_k\right)\right)}{\sum_k p\left(Q_j, Q_k\right)}, \tag{12}$$

where $p(Q_j, Q_k)$ is spatial proximity between Q_j and Q_k, typically given by a negative exponential function of the distance between two points. The surface function is continuous ranging from 0 to 1 defined on discrete locations Λ. It shows a large value if points are densely distributed around Q_j in Ω_i.

Transformation of points into a surface involves redefinition of spatial objects, operations and variables. Distribution Ω_i is represented by a set of continuous values:

$$S\left(\Omega_i\right) = \left\{ s\left(\Omega_i, Q_j\right), j \in \mathfrak{M} \right\}. \tag{13}$$

The number of elements in Ω_i is given by

$$n_S\left(\Omega_i\right)=\sum_j s\left(\Omega_i,Q_j\right).$$

(14)

Intersection of sets is also a set of continuous values:

$$I_S\left(S\left(\Omega\right)\right)=\left\{\min_{i\in\mathfrak{N}}\left(s\left(\Omega_i,Q_j\right)\right),\ j\in\mathfrak{M}\right\}.$$

(15)

Intersection is usually positive so that the problem mentioned earlier can be avoided. The mean and variance of the number of points remain the same because the surface function is standardized. Diversity is given by

$$\gamma_S\left(\Omega\right)=1-\frac{N\sum_j \min_{i\in\mathfrak{N}}\left(s\left(\Omega_i,Q_j\right)\right)}{\sum_i m_i}.$$

(16)

Intersection and union trees can still be constructed in the same way. On the other hand, *K*-cores need careful consideration because they can be redefined in a wide variety of ways. Due to the limitation of space, this paper just suggests two options: one is to visualize the value of $s(\Omega_i, Q_j)$ at Λ, and the other is to show the *K*-cores extracted from original point distributions.

7. Application

This section applies the method proposed to school location planning in Inage and Wakaba wards in Chiba City, Japan (Figure 7).

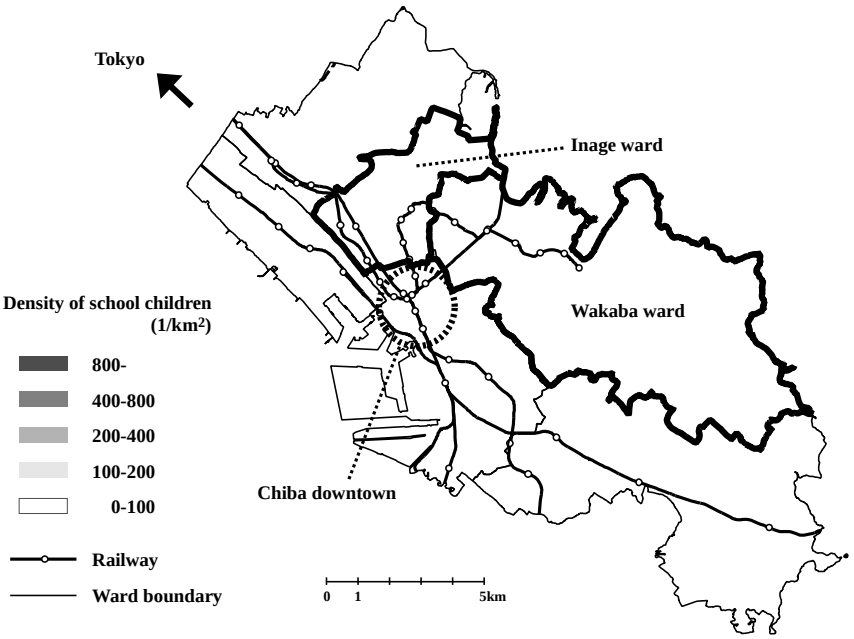

Figure 7 Inage and Wakaba wards in Chiba City. They are located 40-60 minutes from the central area of Tokyo. Residents are working in Tokyo and Chiba downtown.

There were 16 and 20 public elementary schools in 2008 in these wards, respectively (Figure 8). With a rapid decrease in birth rate, however, children of elementary schools have been decreasing since 1981. School reduction has been being discussed to improve educational environment of schools and economic efficiency of school operation.

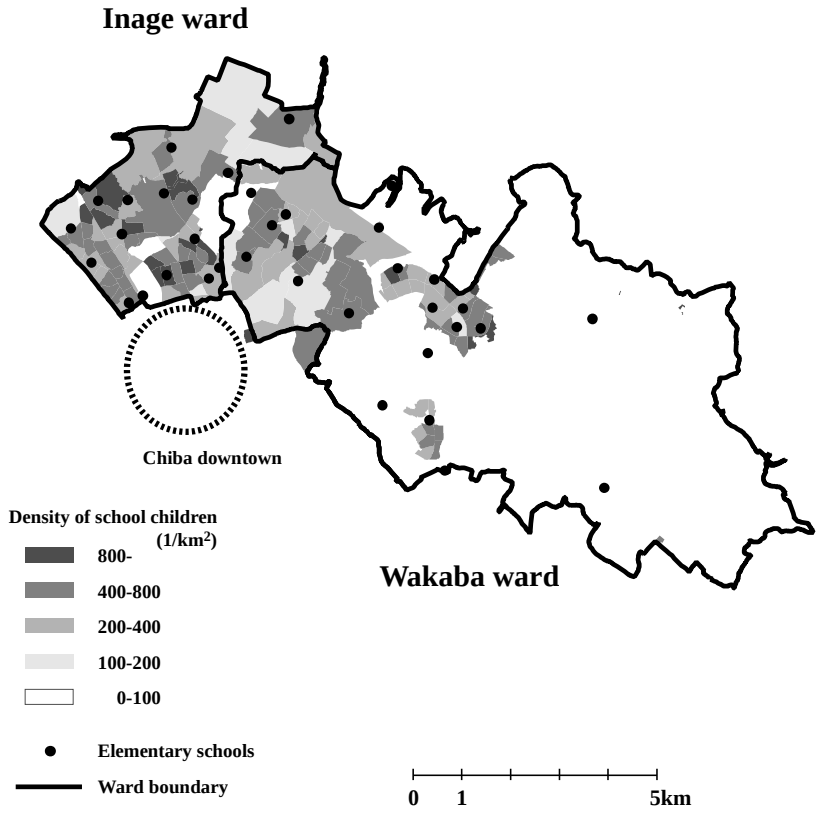

Figure 8 Public elementary schools and density distribution of children aged 6-12. Children concentrate around Chiba downtown and railway stations.

Future plans of school location were derived as the solutions of spatial optimization problem. The minimum numbers of schools were calculated that satisfied the given constraints including the maximum distance from home to school, the minimum and maximum capacity of schools, and so forth.

The solutions were 11 and 15 schools in Inage and Wakaba wards, respectively. Since the constraints were not so restrictive, the solutions could be achieved by numerous combinations of 11 and 15 schools chosen from existing ones. Fifty and a hundred alternatives were extracted in Inage and Wakaba wards, respectively, in the ascending order of average distance from home to school.

The method proposed was utilized to narrow the alternatives to a few desirable ones. They were analyzed in both their original form and surfaces generated by smoothing operation. The spatial proximity was defined by

$$p\left(Q_j, Q_k\right) = \begin{cases} 1 & \text{if } D\left(Q_j, Q_k\right) \le D_{\max} \\ 0 & \text{otherwise} \end{cases},$$

(17)

where $D(Q_i, Q_j)$ was the graph distance between Q_i and Q_j on Delaunay triangulation generated from Λ. The maximum distance D_{max} was set to 1 or 2.

Table 1 shows the summary statistics of the result. Diversity of original distributions is one in both wards indicating that there is no school shared by all the plans. Weighted diversity shows similar values to those of ordinary measures, while spatial smoothing greatly reduces diversity among point distributions. Though weighting and smoothing both extend the ordinary diversity measure, they are different in that the former evaluates points at different locations while the latter considers points in other distributions. In this empirical study, it is shown that the latter is more effective to obtain a significant result.

Table 1 Summary statistics of school plans in Inage and Wakaba wards.

Ward	Number of plans	Number of schools	Point distributions	Diversity	
				Ordinary	Weighted
Inage	50	11	Original	1.000	0.972
			Smoothed (D_{max}=1)	0.593	0.584
			Smoothed (D_{max}=2)	0.173	0.170
Wakaba	100	15	Original	1.000	0.985
			Smoothed (D_{max}=1)	0.521	0.517
			Smoothed (D_{max}=2)	0.214	0.210

School plans were then classified into several groups by using intersection tree. Similarity between plans was evaluated by

$$d\left(\Psi_i, \Psi_j\right) = \frac{d_S\left(\Psi_i, \Psi_i \cap \Psi_j\right) + d_S\left(\Psi_j, \Psi_i \cap \Psi_j\right)}{n\left(\Psi_i \cap \Psi_j\right)}.$$

(18)

Table 2 shows the summary of classification result where four and nine groups were obtained. As seen in this table, smoothing operation greatly affects the result of classification. The number of schools in each group considerably varies among three cases in both wards. The result also suggests that the desirable number of groups depends on the number of school plans. In Inage ward classification into four groups seems better because nine groups include many small ones. In Wakaba ward, on the other hand, school plans are classified more uniformly into nine groups.

Table 2 Classification of school plans in Inage and Wakaba wards.

Ward	Distribution	Number of school plans in each group	
		Classification into 4 groups	Classification into 9 groups
Inage	Original	18, 13, 11, 8	9, 8, 7, 7, 5, 5, 4, 3, 2
	Smoothed (D_{max}=1)	39, 5, 3, 3	29, 7, 4, 3, 2, 2, 1, 1, 1
	Smoothed (D_{max}=2)	21, 16, 12, 1	14, 11, 9, 6, 5, 2, 1, 1, 1
Wakaba	Original	54, 19, 16, 11	24, 13, 13, 11, 11, 9, 8, 6, 5
	Smoothed (D_{max}=1)	56, 31, 12, 1	29, 18, 13, 12, 11, 10, 5, 1, 1
	Smoothed (D_{max}=2)	77, 15, 6, 2	52, 15, 13, 6, 5, 5, 2, 1, 1

Figure 9 shows the dendrograms of school plans in Wakaba ward. In Figure 9(a), groups A, B, C, and D are connected by dashed lines because their intersection is empty. Among the three cases, classification based on the original distributions seems most effective because the variation in size is small and the classification is hierarchical. A hundred plans are classified into a few groups of similar size and then each group is further classified uniformly into smaller ones.

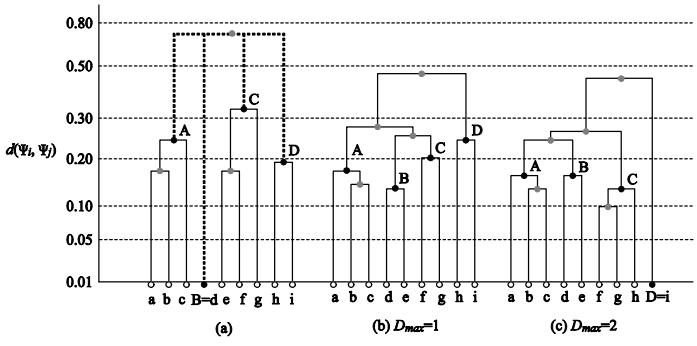

Figure 9 Dendrograms of school plans in Wakaba ward. School plans were classified based on (a) original distribution, (b) smoothed distribution (D_{max}=1), and (c) smoothed distribution (D_{max}=2). Black and while circles indicate groups of school plans where a hundred plans are classified into four and nine groups, respectively. The vertical axis

$d(\Psi_i, \Psi_j)$ is the distance between sets of distributions when clustered into one group. In dendrogram (a), groups A, B, C, and D are connected by dashed lines because there is no point shared by all the distributions.

Each group of school plans was geographically visualized by using 2- and 3-cores as follows. Delaunay triangulation was generated from school locations to extract the pairs and triplets of schools located within 800m connected directly by links. From all the pairs and triplets, the 2- or 3-core having the largest amount of information was chosen. The next *K*-core was chosen from those not overlapping with the first one. This process continued until no *K*-core remains.

Figures 10, 11, and 12 show the *K*-cores of each group in Wakaba ward. Plans were classified into nine groups, among which those consisting of more than nine plans are presented.

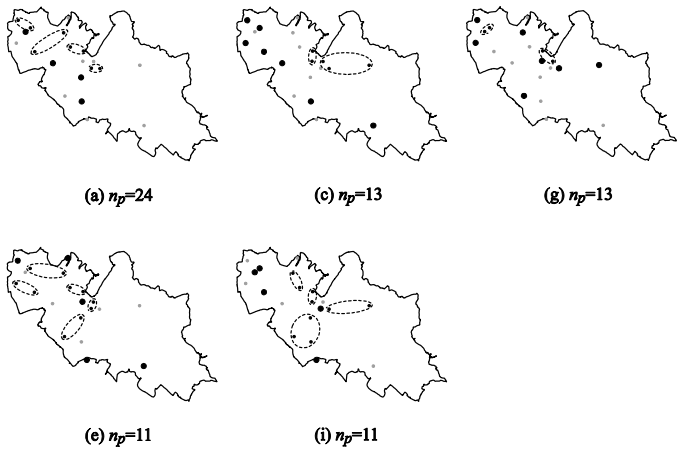

Figure 10 *K*-cores in school plans in Wakaba ward. Classification is based on original distributions. Groups consisting of more than nine plans are presented (n_p is the number of plans in each group). Large black dots indicate 1-cores. Small black dots encircled by solid lines are 2_C- and 3_C-cores. Small black dots encircled by dashed lines are 2_O- and 3_O-cores.

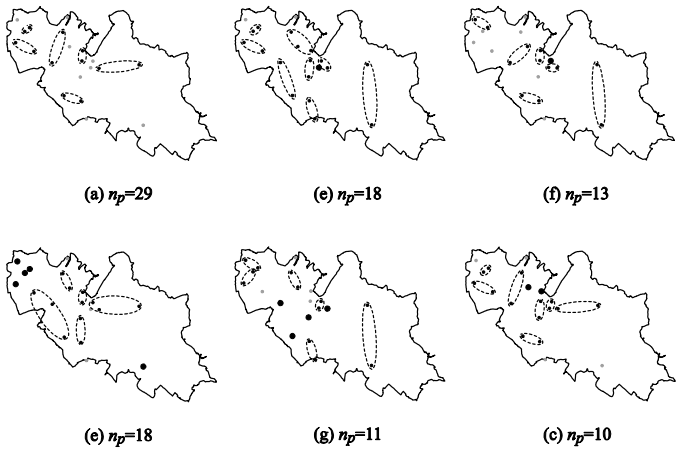

Figure 11 *K*-cores of school plans in Wakaba ward. Classification is based on smoothed distributions (D_{max}=1). The same symbols are used as in Figure 10.

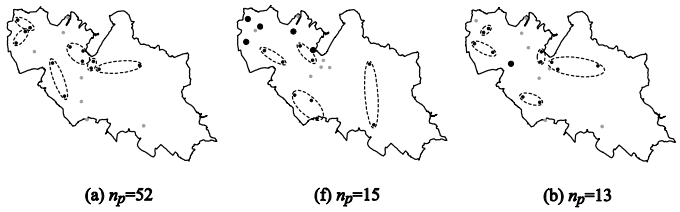

Figure 12 *K*-cores in school plans in Wakaba ward. Classification is based on smoothed distributions (D_{max}=2). The same symbols are used as in Figure 10.

These figures show that small groups have many 1-cores while large groups have 2- and 3-cores. It is reasonable because larger groups have greater variation in school plans than smaller ones. Figure 10 contains more 1-cores than Figures 11 and 12. Since 1-cores are defined by points exactly at the same location, they hardly exist in groups obtained by the classification where a location difference of points is permitted.
In school location planning, the result can be utilized as follows. In Wakaba ward, public elementary schools have to be reduced from 20 to 16. A hundred school plans are classified into several groups, and larger groups are discussed in detail. In each group, schools represented by 1-cores are determined to remain. Then, one school is chosen from every 2_C- and 3_C-core. Finally, schools are chosen from 2_O- and 3_O-cores and others up to 16 in such a way that the given constraints are satisfied. The results are compared among different groups in other aspects such as the racial composition, minority distribution, and school quality to make a final decision.

8. Conclusion

This paper has proposed a new method of analyzing the relation among point distributions on a discrete space. Properties of distributions are described by numerical measures and geographical representation. Structural relation among point distributions is described by Hasse diagram and intersection tree, the latter of which also serves as a means of classifying distributions into similar groups. The method was applied to school location planning in Chiba City, Japan. The result revealed the properties of the method as well as illustrated its utilization in school location planning.

We finally discuss some limitations and extensions of the paper for future research. First, this paper adopts a hierarchical method to construct intersection trees. It is primarily because implementation is straightforward and result is easy to interpret. One drawback is that such a hierarchical method does not assure the global optimality of result. Other non-hierarchical and heuristic methods such as k-means should also be considered. Second, this paper focuses on points distributed on a discrete space. This is a reasonable setting in urban environment where land blocks are predetermined by land ownership system. In ecology and epidemiology, however, it is more realistic to assume a continuous space in point distributions. Extension in this direction is indispensable to treat a wider variety of point distributions. Third, points and point distributions are dual in the sense that point distributions are defined by points that compose the distributions while points are characterized by point distributions that they are included in. This permits us to transfer the relations defined on point distributions to points as seen in Section 3. This implies that numerical measures, *K*-cores, and intersection trees can also be considered on points. Intersection trees, for instance, provide a means of classifying the points with respect to their inclusion in point distributions. Though this paper focuses on point distributions, it would be useful to extend the method for analyzing the relation among points.

References

Bailey, T. C. and Gatrell, A. C., 1995. *Interactive Spatial Data Analysis*. London: Taylor & Francis.

Besag, J. and Newell, J., 1991. The Detection of Clusters in Rare Diseases. *Journal of the Royal Statistical Society, Series A,* **154**, 143–155.

Cressie, N. (1993) *Statistics for Spatial Data.* New York: John Wiley.

Davey, B. A. and Priestley, H. A., 2002. *Introduction to Lattice and Order*. Cambridge: Cambridge University Press.

Department of Epidemiology, UCLA, 2009. John Snow – A Historical Giant in Epidemiology. Available from: URL:http://www.ph.ucla.edu/epi/snow.html [04/2009].

Fortin, M.-J. and Dale, M. R. T., 2005. *Spatial Analysis: A Guide for Ecologists*. Cambridge: Cambridge University Press.

Glaz, J., Naus, J., and Wallenstein, S., 2001. *Scan Statistics*. Berlin: Springer.

Jacquez, G. M., 1996. A K-nearest Test for Space-Time Interaction. *Statistics in Medicine*, **15**, 1935-1949.

Knox, E. G., 1964. The Detection of Space-Time Interaction. *Applied Statistics*, **13**, 25-29.

Kulldorff, M. and Nagawalla, N., 1994. Spatial Disease Clusters: Detection and Inference. *Statistics in Medicine*, **13**, 1–12.

Kulldorff, A., 1997. A Spatial Scan Statistics. *Communications in Statistics: Theory and Methods*, **26**, 1481-1496.

Lai, P.-C., So., F.-M., and Wing, C.-K., 2008. *Spatial Epidemiological Approaches in Disease Mapping and Analysis*. Boca Raton: CRC Press.

Mantel, N., 1967. The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, **27**, 209-220.

Okabe, A. and Yamada, I., 2001. The K-function Method on a Network and Its Computational Implementation. *Geographical Analysis*, **33**, 271-290.

Okabe, A., Okunuki, K., and Shiode, S., 2006. The SANET Toolbox: New Methods for Network Spatial Analysis. *Transactions in GIS*, **10**, 535-550.

Openshaw, S., Charlton, M., Wymer, C., and Craft, A., 1987. A Mark 1 Geographical Analysis Machine for the Automated Analysis of Point Data Set. *International Journal of Geographical Information Systems*, **1**, 335–358.

Openshaw, S., 1994. Two Exploratory Space-Time-Attribute Pattern Analyzers Relevant to GIS. In *Spatial Analysis and GIS*, edited by S. Fotheringham and P. Rogerson (London: Taylor & Francis), 83-104.

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

Pemmaraju, S. and Skiena, S., 2003. *Computational Discrete Mathematics: Combinatorics and Graph Theory with Mathematica* . Cambridge University Press, Cambridge.

Ripley, R. D., 1981. *Spatial Statistics*. New York: John Wiley.

Sadahiro, Y., and Sasaya, T., 2008. Analysis of the relationship among spatial tessellations. *Discussion Paper*, **96**, Department of Urban Engineering, University of Tokyo.

Sadahiro, Y., 2009. Analysis of the relation among point distributions on a discrete space. *Discussion Paper*, **98r1**, Department of Urban Engineering, University of Tokyo.

Tobler, W., 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, **46**, 234-240.

Upton, G. J. G. and Fingleton, B., 1985. *Spatial Data Analysis by Example: Point Pattern and Quantitative Data*. New York: John Wiley.

Yamada, I. and Thill, J.-C., 2007. Local Indicators of Network-Constrained Clusters in Spatial Point Patterns. *Geographical Analysis*, **39**, 268-292.