Discussion Paper Series
*No. **109***

Department of Urban Engineering
University of Tokyo

Quantized spatial analysis: A new framework for analyzing the relations among the distributions of spatial objects

Yukio Sadahiro
Department of Urban Engineering, University of Tokyo

Abstract

This paper develops a new framework for analyzing the relations among distributions of spatial objects. Existing analytical methods usually focus on the distribution of one or two specific types of spatial objects. This paper, on the other hand, deals with the relations among more than two distributions of spatial objects. The framework is applicable independent of the type of spatial objects. It consists of four steps: 1) quantization, 2) analysis, 3) visualization, and 4) evaluation. Quantization introduces the concept of parts, bodies, and tags to describe spatial objects as abstract elements. Analysis explores spatial relations among the distributions of objects. Visualization graphically presents the spatial relations among objects. Evaluation measures the significance of the result of analysis. The paper applies the framework to the analysis of the distributions of commercial facilities in Chiba City, Japan. The result supports the technical soundness of the method as well as provides empirical findings.

1. Introduction

Geography deals with the distribution of a wide variety of geographical entities and phenomena. Geographical information science represents them as spatial objects such as points, lines, polygons, and surfaces and assigns them their attributes as nominal, categorical, and numerical variables. Geographers analyze their spatial distributions, the relationship between the distribution of objects and their attributes, and so forth.

The point is one of the most basic and fundamental spatial objects used in geographical information science. Statistical analysis of point distributions is often called point pattern analysis and numerous methods have been developed in the literature. They include quadrat method (Goodall, 1952; Greig-Smith, 1952; Pielou, 1969), nearest neighbor distance (Skellam, 1952; Clark and Evans, 1954; Diggle, 2003), K-function (Ripley, 1976, 1977, 1981), and kernel density estimation (Rosenblatt, 1956; Parzen, 1962; Silverman, 1986).

The line also plays a critical role in geographical information science. Lines represent water stream, gas pipelines, traffic, electric, and information networks. Analytical methods based on graph theory permit us to evaluate the properties of networks, say, size, density, connectivity, and centrality (Shimble, 1953; Haggett and Chorley, 1969). These methods are recently applied to the analysis of social and web networks (Wasserman and Faust, 1994; Carrington *et al.*, 2005; Knoke and Yang, 2008; Abraham *et al*., 2009).

The above methods treat the distribution of a single type of spatial objects. In the real world, however, a wide variety of spatial objects exists and affect with each other. It is clearly indispensable to analyze the relations among more than a single distribution of spatial objects to understand the whole picture of the real world.

In point pattern analysis, several methods are available to examine the relationship between two distributions of points. They include nearest neighbor contingency table (Pielou, 1961; Dixon, 1994), bivariate J-function (van Lieshout and Baddeley, 1999), cross K-function (Ripley, 1981), and nearest neighbor measures (Lee, 1979; Okabe and Miki, 1984).

Concerning the relationship between points and other types of spatial objects, a few methods have been proposed in the literature. Okabe and Fujii (1984) and Okabe *et al*. (1988) propose statistical methods for analyzing the relationship between points and a network and that between points and polygons, respectively. Sadahiro (1999) discusses the relationship between points and a surface.

Unfortunately, the type and the number of distributions dealt with in existing methods are quite limited. There are few methods that treat spatial objects other than

points and lines. Given more than two distributions of spatial objects, we have to evaluate every pair of two distributions separately. Since each method is tailored for a specific type of spatial objects, numerous methods have to be developed to treat a wide variety of spatial objects.

To resolve the problem, this paper proposes a new framework for analyzing the relations among the distributions of spatial objects. It is a generalized procedure based on those proposed in Sadahiro (2010, 2011, 2012b), Sadahiro and Kobayashi (2012), and Sadahiro *et al*. (2012), each of which focuses on a specific type of spatial objects: 1) point distributions on a discrete space, 2) spatial tessellations, 3) a set of single polygons, 4) trajectories on a network, and 5) spatially distributed time series data. On the basis of these papers, this paper proposes a general framework that is applicable independent of the type of spatial objects, say, points, lines, polygons, surfaces, tessellations, and so forth.

Section 2 describes a four-step procedure: 1) quantization, 2) analysis, 3) visualization, and 4) evaluation. The description proceeds with a concrete illustration of the analysis of point distributions on a continuous space. This aims to avoid a too abstract explanation of the method that is not easily accessible to readers. We choose point distributions on a continuous space because it is a more general case than those discussed in earlier papers. To evaluate the validity of the framework, Section 3 applies it to the analysis of the distributions of commercial facilities in Chiba City, Japan. Section 4 summarizes the conclusions with a discussion.

2. Framework

2.1 Quantization

We first define *parts*, the set of mutually exclusive elements that do not need further division into smaller pieces to compose all the distributions. Parts describe all the distributions as their subsets and serve as fundamental spatial objects in analysis.

Definition of parts depends on the topological relation among spatial objects. If spatial objects do not overlap with each other, parts are given by the union set of all the objects. In point distributions, for instance, each point forms a part. Every distribution can be represented as a set of points. Lines, polygons, and their mixture can be treated similarly if they do not overlap with each other. If spatial objects overlap, we make intersection of all the objects to obtain fragmented pieces each of which is defined as a part. Polygons with overlap are divided into smaller ones. Intersection of trajectories with overlap yields either lines or a mixture of points and lines. The former occurs when the trajectory data are obtained by map matching on a road network (Sadahiro *et al*.,

2012). The latter case applies to raw GPS data. Surfaces defined on a plane are treated as three-dimensional objects in making intersection. Sadahiro and Kobayashi (2012) takes a similar approach in the case of continuous functions defined on a one-dimensional space.

Having defined parts, we can describe any distribution as a set of parts. This paper calls a distribution of spatial objects a *body*. Let $\mathbf{B}=\{B_1, B_2, ..., B_M\}$ be the set of bodies. Body B_i is represented as a set of parts $\{P_{i1}, P_{i2}, ..., P_{imi}\}$. The number of elements and the ith element in set Q are denoted by $\#(Q)$ and $e(Q, i)$, respectively.

We then evaluate the spatial relations among parts. The method again depends on the topological relation among spatial objects. If spatial objects overlap with each other, spatial relations among parts can be defined by spatial adjacency. A pair of parts is either adjacent or nonadjacent. For spatial objects without overlap, we define *neighborhoods* of parts and evaluate their relations. A basic definition of neighborhoods is the buffer region of parts, where parts are points, lines, polygons, or their mixture. Parts whose neighborhoods overlap are regarded as proximal while others are separated. Spatial proximity is equivalent to spatial adjacency defined for objects with overlap.

Neighborhoods obtained by buffer operation depend on the buffer distance. It determines the scale of analysis as often seen in exploratory spatial analysis such as K-function and local statistics. A large value is appropriate for analysis at a global scale while local analysis requires a small value. This paper recommends starting from a small value and gradually increasing it for detecting spatial patterns at various scales that are useful for building research hypotheses.

Figure 1a shows four distributions of points and their neighborhoods. Point distributions are represented as bodies: $B_1=\{P_{11}, P_{12}, P_{13}, P_{14}, P_{15}, P_{16}\}$, $B_2=\{P_{21}, P_{22}, P_{23}, P_{24}\}$, $B_3=\{P_{31}, P_{32}, P_{33}, P_{34}\}$, $B_4=\{P_{41}, P_{42}\}$. In this figure, for instance, points P_{12}, P_{21}, and P_{32} are regarded as spatially proximal.

To describe explicitly the above spatial relations, we define spatial *tags*. In spatial objects with overlap, every part is assigned a single tag. Spatial adjacency of parts is represented as that of their assigned tags. When spatial objects do not overlap, every region generated by the intersection of all the neighborhoods is assigned a tag. Two parts are regarded as proximal if their neighborhoods share the same tag. A set of tags is denoted by $\mathbf{T}=\{T_1, T_2, ..., T_K\}$. Figure 1b shows the tags assigned to the regions generated by the intersection of neighborhoods.

[Figure 1]

Every tag has its own *size* denoted by $s(T_i)$. When tags are defined on a two-dimensional space, their size is the area of their assigned regions. On a one-dimensional space, tag size is the length of its assigned line segment.

Tags, parts, and bodies are related with each other. This paper calls this relationship *assignment*. Every body and part can be represented by a set of assigned tags. The set of bodies represented by the sets of assigned tags is denoted by $\mathbf{B_T}$. Every tag can also be represented as a set of assigned parts and bodies. In Figure 1, tag T_8 is assigned to points P_{12}, P_{21}, and P_{32}, bodies B_1, B_2, and B_3, while P_{12}, P_{21}, P_{32}, B_1, B_2, and B_3 are assigned to T_8. Part P_{12} and body B_2 can be represented as $\{T_5, T_6, T_8, T_9\}$ and $\{P_{21}, P_{22}, P_{23}, P_{24}\}=\{\{T_5, T_6, T_8, T_9\}, \{T_{14}, T_{15}\}, \{T_{16}, T_{17}\}, \{T_{21}, T_{22}, T_{23}, T_{24}, T_{25}, T_{26}, T_{27}\}\}$, respectively.

Tags relate parts in different bodies with each other. For instance, tag T_8 relates parts P_{12}, P_{21}, and P_{32}, and T_6 relates P_{12} and P_{21}. Tags even relate different bodies through parts. Tag T_8 relates bodies B_1, B_2, and B_3 through P_{12}, P_{21}, and P_{32}. Tag T_6 relates B_1 and B_2 through P_{12} and P_{21}.

We call the above process *quantization*. Quantization divides spatial objects into smaller elements such as points, lines, polygons, and their mixture. It also divides even continuous distributions such as surfaces defined on a two-dimensional space into discrete objects. Quantization then describes spatial objects and their relations by using abstract concepts, that is, parts, bodies, and tags. This abstraction permits us to analyze spatial objects independent of their type, say, points, lines, and polygons as seen later.

2.2 Analysis

The choice of analytical method depends on the spatial relations among the distributions of spatial objects that are focused in analysis. Spatial relations include topological relations, hierarchical relations, complementary relations, spatial proximity, and so forth. The objective of analysis is to detect such relations among distributions.

Topological relations are usually discussed in polygon distributions on a two-dimensional space (Sadahiro *et al*., 2012). However, they can also be defined for other types of spatial objects. Sadahiro (2010), for instance, considers the topological relations among point distributions on a discrete space. Sadahiro (2011) analyzes the topological relations among spatial tessellations.

Hierarchical relations are closely related to topological relations. Administrative units represented by spatial tessellations usually form a hierarchical structure. Points and lines can also exhibit hierarchical structure. Sadahiro (2011) and Sadahiro (2012b) discuss the hierarchical structure of tessellations and polygons,

respectively.

Complementary relations appear in the analysis of a distribution with respect to the union set of all the elements of distributions. Two distributions are complementary if their union covers all the elements. Sadahiro (2010) and Sadahiro (2011) discuss the complementary relations among points and tessellations, respectively.

Each of the above papers focuses on a specific type of spatial objects. However, once we interpret spatial objects in these papers as parts, tags, and bodies, we can use the methods independent of the type of spatial objects. For instance, we can analyze the hierarchical relations among trajectories on a network by using the method developed for polygons in Sadahiro (2012b). Quantization allows us to broadly discuss topological, hierarchical, and complementary relations among spatial distributions.

This paper, on the other hand, focuses on the spatial proximity among distributions. Though Sadahiro *et al*. (2012) and Sadahiro and Kobayashi (2012) discuss spatial proximity, they only treat spatial objects with overlap. Our focus is on spatial objects without overlap such as point distributions.

The objective of analysis is to detect local clusters of spatial objects that belong to different distributions. Suppose, for instance, the distributions of commercial facilities. Supermarkets, convenience stores, fast-food restaurants, and drug stores display different distributions at a global scale. However, we often find shopping malls that contain all the four types of commercial facilities. This implies that these distributions are very similar with each other at a local scale. The method proposed in this section detects such clusters and groups the distributions based on their local similarity, which cannot be extracted by existing methods of local spatial analysis such as scan statistics.

We call the areas such as above shopping malls *centers* denoted by $\mathbf{C}=\{C_1, C_2, ..., C_N\}$. They are formally defined as the sets of tags each of which are shared by many parts of different bodies. Each center consists of a set of tags assigned to the same set of parts. A body is said to be *assigned* to C_i if all the tags of C_i are assigned to parts in the body. Let Γ_i be a set of bodies assigned to center C_i. The set of body sets is denoted by $\mathbf{\Gamma}=\{\Gamma_1, \Gamma_2, ..., \Gamma_N\}$.

To detect centers, we propose Algorithm CB as shown below. It employs a set of tags Θ and that of bodies Ψ, both of which are initially empty. Only bodies assigned to all the tags in Θ are contained in Ψ throughout the detection process. Algorithm CB starts with choosing the tag shared by the most neighborhoods. The tag and its assigned bodies are substituted to Θ and Ψ, respectively. The algorithm gradually adds tags to Θ by removing bodies in Ψ until the tags in Θ become larger than a predetermined

threshold β.

Algorithm CB (Center detection and Body clustering)

Input:

Set of bodies represented by both **B** and $\mathbf{B_T}$

Set of tags **T**

Conditions $\{\vartheta_S, \vartheta_{A1}, \vartheta_{A2}\}$

Parameters $\{\alpha, \beta\}$

Output:

Set of tags representing centers **C** and their related tags **C'**

Sets of bodies assigned to centers **Γ** and related bodies **Γ'**

Algorithm:

1. $\mathbf{C}=\Gamma\in\mathbf{\Gamma}=C_i\in\mathbf{C}=\varnothing$. $j=0$.
2. Repeat the steps 3-26 while $\#(\Psi)\geq\alpha$ at step 6.
3. $\Theta=\Psi=\Psi'=\varnothing$. $k=1$.
4. Choose the set of tags in **T** satisfying ϑ_{A1}.
5. Add the tags to Θ.
6. Add the bodies in **B** assigned to the chosen tag to Ψ.
7. Repeat the steps 8-21 while $\#(\Psi)>1$.
8. If $s(\Theta)\geq\beta$ then
9. If $C_j\neq\varnothing$ then
10. Move the bodies from Ψ to Ψ' that do not contain all the elements of Θ.
11. If $C_j=\varnothing$ and $\#(\Psi)\geq\alpha$ then
12. $j=j+1$.
13. $C_j=\Theta$.
14. $\Gamma_j=\Psi$.
15. If $C_j=\varnothing$ and $\#(\Psi)<\alpha$ then
16. Move the bodies from Ψ to Ψ' that do not contain all the elements of Θ.
17. If $s(\Theta)<\beta$ then
18. Move the bodies from Ψ to Ψ' that do not contain all the elements of Θ.
19. Choose the set of tags in $\mathbf{T}\backslash\Theta$ satisfying ϑ_S and ϑ_{A2}.
20. Add the tags to Θ.
21. Move the bodies from Ψ to Ψ' that are not assigned to all the elements of Θ.

22. If $C_j \neq \varnothing$ then
23. $C_j' = \Theta$.
24. $\Gamma_j' = \Psi'$.
25. Remove the tags of parts of Γ_j containing C_j from $\mathbf{B_T}$.
26. If $C_j = \varnothing$ then $k = k+1$.
27. Return **C**, **C'**, **Γ**, and **Γ'**.

Algorithm CB detects centers represented by **C** and clusters bodies into similar groups based on **C** represented by **Γ**. Since it is designed as a general algorithm applicable independent of the type of spatial objects, conditions ϑ_S, ϑ_{A1}, and ϑ_{A2} and parameters α and β are not specified. They have to be defined in each application as well as tags, parts, and bodies.

Condition ϑ_S is the spatial requirement on tags in **C** while conditions ϑ_{A1} and ϑ_{A2} indicate requirements on their attributes. The former depends on whether or not each center has to be in a specific spatial form. Analysis of point distributions does not usually impose this condition. To obtain spatially proximal centers, we define ϑ_S as a tag adjacent to those in Θ.

Condition ϑ_{A1} is given to choose a tag that best represents many bodies. We usually define ϑ_{A1} as the largest tag among those assigned to the kth most bodies in $\mathbf{B_T}$. By slightly modifying condition ϑ_{A1}, we obtain condition ϑ_{A2}: the largest set of tags in $\mathbf{B_T}$ that are assigned to the most bodies in Ψ.

Parameters α and β are the minimum number of bodies assigned to a center and the minimum size of a center, respectively. Large α and β yield large centers assigned to many bodies. Though such centers are useful in analysis, large parameter values are often too restrictive so that Algorithm CB may detect only a few centers. In practice, we should start with small values, say, $\alpha = 0.001 \times M$ and $\beta = 0.0001 \times s(\mathbf{T})$, and gradually increase them until a reasonable number of centers are obtained.

Figure 2 shows the process of Algorithm CB applied to the point distributions shown in Figure 1. Let us first suppose the case where $\alpha = 3$ and $\beta = \beta_1$, the latter of which is indicated by the area of the dotted circle in Figure 2a. Algorithm CB chooses tag T_{27} at step 4 because it is the only tag assigned to four bodies. The sets Θ and Ψ become $\{T_{27}\}$ and $\{B_1, B_2, B_3, B_4\}$, respectively, the former of which is indicated by bold lines in Figure 2a. Since $s(T_{27}) < \beta_1$, Algorithm CB proceeds to step 17 and then 19 to choose $\{T_8, T_{25}\}$ because $s(T_8)+s(T_{25})$ is largest among the sets of tags assigned to three bodies in Ψ. The sets Θ and Ψ then become $\{\{T_{27}\}, \{T_8, T_{25}\}\}$ and $\{B_1, B_2, B_3\}$, respectively. The set $\{T_8, T_{25}\}$ is indicated by thin lines in Figure 2a. Algorithm CB returns to step 8. Since

$s(T_{27})+s(T_8)+s(T_{25})>\beta$, the tag set $\Theta=\{\{T_{27}\}, \{T_8, T_{25}\}\}$ is substituted to center C_1 at step 13. This center is indicated by dark gray shades in Figure 2b. Algorithm CB still continues until only a single element is contained in Ψ.

If $\alpha=3$ and $\beta=\beta_2$, Algorithm CB does not regard $\Theta=\{\{T_{27}\}, \{T_8, T_{25}\}\}$ as a center because $s(T_{27})+s(T_8)+s(T_{25})<\beta$. It proceeds to step 17 and adds tags $\{T_6, T_{14}, T_{17}, T_{24}\}$ to Θ. Though $s(\Theta)>\beta_2$, set Θ is not a center because $\#(\Psi)<\alpha$. Algorithm CB does not detect any center because the requirements for centers are too restrictive.

When $\alpha=2$ and $\beta=\beta_1$, Algorithm CB detects the center represented by $\{\{T_{27}\}, \{T_8, T_{25}\}, \{T_6, T_{14}, T_{17}, T_{22}\}\}$ indicated by light gray shades in Figure 2b. Two bodies $\{B_1, B_2\}$ are assigned to the center.

[Figure 2]

Algorithm CB is based on Algorithm TC proposed by Sadahiro and Kobayashi (2012) that was originally developed by Kharrat *et al*. (2008). While existing algorithms assume the type of spatial objects under limited circumstances, Algorithm CB is independent of the type of spatial objects.

In case of Figure 2, only a single center is detected. However, when a number of distributions are analyzed, many centers are often detected. Each body can be assigned to more than one center, and centers sharing the same tags spatially overlap with each other.

Algorithm CB generates two ordered sets of tags and bodies $\mathbf{C'}$ and $\mathbf{\Gamma'}$ as well as $\mathbf{C}$ and $\mathbf{\Gamma}$. In Figure 2, $\mathbf{C'}=\{\{T_{27}\}, \{T_8, T_{25}\}, \{T_6, T_{14}, T_{17}, T_{22}\}, \{T_1, T_3, T_5, T_9, T_{12}, T_{13}, T_{18}, T_{26}, T_{29}, T_{31}, T_{32}, T_{33}\}\}$ and $\mathbf{\Gamma'}=\{B_4, B_3, B_2, B_1\}$. The tag sets are arranged in the order of addition while the bodies are arranged in the order of removal. Since the orders reflect that of similarity of elements, $\mathbf{C'}$ and $\mathbf{\Gamma'}$ provide a means of classify tags and bodies. In $\mathbf{\Gamma'}=\{B_4, B_3, B_2, B_1\}$, for instance, bodies $\{B_3, B_2, B_1\}$ are more similar with each other than B_4. We can classify the bodies as $\{\{B_4\}, \{B_3, B_2, B_1\}\}$, $\{\{B_4\}, \{B_3\}, \{B_2, B_1\}\}$, or $\{\{B_4\}, \{B_3\}, \{B_2\}, \{B_1\}\}$. Classifications such as $\{\{B_4, B_2\}, \{B_3, B_1\}\}$ and $\{\{B_4, B_1\}, \{B_3, B_2\}\}$ are not permissible because they are inconsistent with the order of bodies in $\mathbf{\Gamma'}$.

This classification scheme permits each body or tag contained in more than one group when multiple centers are detected. To avoid this, we only have to remove all the tags of Γ_j in $\mathbf{B_T}$ at step 25. This enables the classification of all the bodies and tags without overlap.

Algorithm CB adds tags to Θ while removes bodies from Ψ under given

conditions. Tags and bodies are interchangeable because they can be both represented as the sets of the others as mentioned earlier. Exchanging tags and bodies, we obtain a dual algorithm of Algorithm CB (for details, see Sadahiro, 2012a).

2.3 Visualization

To visualize the result of analysis, this paper proposes a graph-based representation called *topology diagram*. Topology diagram has been originally developed in Sadahiro (2010, 2012b), Sadahiro and Kobayashi (2012), and Sadahiro *et al*. (2012). The power set of tags **T** and Boolean operations $\{\cap, \cup\}$ form a lattice (Anderson, 2002; Pemmaraju and Skiena, 2003), where the least and greatest elements are $\varnothing$ and the union set of all the tags. A lattice is a poset (partially ordered set), and consequently, visualized by Hasse diagram (Birkhoff, 1979; Davey and Priestley, 2002). Topology diagram is a modified subset of Hasse diagram that represents the topological relations among spatial objects. Nodes indicate tags, bodies, and centers, while edges indicate their topological relations. Parts are represented either explicitly by tags or implicitly by their composing tags. The vertical axis indicates the size of spatial objects.

Choice of topology diagram depends on the spatial relations focused in analysis. If topological relations are of importance, topology diagram proposed in Sadahiro (2011) or Sadahiro (2012b) would be appropriate. For hierarchical relations, the diagram shown in Sadahiro (2010) is useful.

To visualize the spatial proximity among distributions, this paper generates a topology diagram by tracing the entire process of detecting a single center in Algorithm CB. Addition of tags to Θ and removal of bodies from Ψ can be represented as a graph whose nodes indicate a center and its related bodies and tags.

Figure 3 shows the topology diagram representing the process of detecting the center in Figure 1 when α=2. Bold lines represent the growth of Θ. Algorithm CB adds tag sets $\{T_8, T_{25}\}$, $\{T_6, T_{14}, T_{17}, T_{22}\}$, $\{T_1, T_3, T_5, T_9, T_{12}, T_{13}, T_{18}, T_{26}, T_{29}, T_{31}, T_{32}, T_{33}\}$ to Θ while removes bodies B_4, B_3, and B_2 from Ψ. The former process is represented as a tree indicated by bold and thin solid lines while the latter is a tree consisting of bold solid and dotted lines in Figure 3.

[Figure 3]

Topology diagram permits us to grasp visually the entire process of Algorithm CB. In addition, it is convenient for classifying bodies and tags without inconsistency mentioned earlier. Cutting some edges connecting intermediate sets of tags in Θ, we

obtain groups of bodies and tags. In Figure 3, for instance, we cut E_1 to obtain two sets of bodies $\{B_1, B_2, B_3\}$, $\{B_4\}$ and two sets of tags $\{T_{27}\}$, $\{\{T_8, T_{25}\}, \{T_6, T_{14}, T_{17}, T_{22}\}, \{T_1, T_3, T_5, T_9, T_{12}, T_{13}, T_{18}, T_{26}, T_{29}, T_{31}, T_{32}, T_{33}\}\}$. Cutting E_1 and E_2, we obtain three sets of bodies. Topology diagram inherently prohibits classifications inconsistent with the order of bodies in $\mathbf{\Gamma'}$ such as $\{\{B_4, B_2\}, \{B_3, B_1\}\}$ and $\{\{B_4, B_1\}, \{B_3, B_2\}\}$.

Topology diagram can also visualize the relations among multiple centers. To generate a topology diagram for multiple centers, we employ Algorithm CB by regarding centers as bodies. Tracing the entire process of Algorithm CB, we obtain a topology diagram that indicates the topological relations among centers.

2.4 Evaluation

The result of analysis is evaluated by numerical measures. In the above analysis, the result is informative and useful when many centers are detected and many bodies are clustered into groups. Therefore, a basic measure is N, the number of centers detected by Algorithm CB. In addition, the ratios of the bodies and tags assigned to centers are also useful:

$$R_{NB} = \frac{\#\left(\bigcap_{\Gamma_i \in \mathbf{\Gamma}} \Gamma_i \right)}{M} \tag{1}$$

and

$$R_{NT} = \frac{\#\left(\bigcap_{C_i \in \mathbf{C}} C_i \right)}{K} . \tag{2}$$

The size of centers is also a critical measure of evaluating the effectiveness of analysis. Its standardized form is defined by

$$R_S = \frac{\sum_i \#(\Gamma_i) s(C_i)}{s(\mathbf{B_T})} . \tag{3}$$

The above measures reflect the similarity among the distributions of spatial objects. A high similarity permits Algorithm CB to detect large and many centers, and consequently, measures become large. Measures are small when a wide variety exists

among the distribution of spatial objects.

Measures can also be defined for individual centers. A center represents its assigned bodies and it is more representative and informative if it is large and assigned many bodies. The representativeness of a center can be evaluated by its size and the ratios of assigned bodies and tags:

$$r_{SN}\left(C_i\right)=\frac{\#\left(\Gamma_i\right)s\left(C_i\right)}{\sum_{B_j\in\Gamma_i}s\left(B_j\right)}, \tag{4}$$

$$r_{NB}\left(C_i\right)=\frac{\#\left(\Gamma_i\right)}{M}, \tag{5}$$

and

$$r_{NT}\left(C_i\right)=\frac{\#\left(C_i\right)}{K}. \tag{6}$$

The above measures are all ratio variables ranging from 0 to 1. Large values indicate the high representativeness of a single or multiples centers.

2.5 Rasterization

Quantization involves intersection of spatial objects or their neighborhoods. It often generates numerous parts and tags because they rapidly increase with an increase of spatial objects. Since the computing time heavily depends on the number of parts and tags, calculation may not terminate within a reasonable time.

A practical solution to this problem is to discretize the space on which spatial objects are distributed into small units. When the space is two-dimensional, a square lattice is a reasonable solution where cells serve as basic units. Parts, neighborhoods, and bodies are approximated by sets of cells and tags are assigned to the cells. If the space is one-dimensional such as a network space, we can use edges as basic units in discretization (Sadahiro *et al*., 2012).

Computational complexity of Algorithm CB before discretization in the worst case is $\mathrm{O}\left(2^{\sum_i m_i}\right)$. It reduces to O(*MU*) by discretization, where *M* and *U* are the

numbers of bodies and basic units in discretization, respectively. Since it is a linear function of M and U, computation time is practically acceptable.

A fine discretization sounds desirable because it provides a good approximation of spatial objects. However, a high resolution is not always effective because it often generates many small centers that are not significant in analysis. Considering the initial value of β mentioned earlier, this paper recommends to set m from ten thousands to a million. Sadahiro and Kobayashi (2012) reports that lattices of resolution 200×200 to 1000×1000 yielded almost similar results.

3. Empirical application

To validate the framework proposed in the previous section, this section applies it to the analysis of the distributions of commercial facilities in Chiba City, Japan. Chiba is located 30 kilometers away from the east of Tokyo. List of commercial facilities in the NTT telephone directory was converted into spatial data by geocoding and handled with in ArcGIS. Chiba has 16,311 commercial facilities of 235 categories.

Figure 4 shows the distribution of commercial facilities in Chiba. Downtown of Chiba is located in the southeast of Chiba station indicated by the largest cluster of darker cells. The northern and western parts of Chiba serve as residential areas for people working in Tokyo. The population density is higher in these areas. The southern and eastern parts are residential areas for people working in the downtown of Chiba. The population density in these areas is relatively low. Commercial facilities are concentrated mainly around railway stations. Chiba station has the largest cluster and stations in subcenters such as Inage and Tsuga also draw many facilities.

[Figure 4]

The objective of analysis is to detect local clusters of commercial facilities, each of which consists of the same combination of facilities such as supermarkets, convenience stores, fast-food restaurants, and drug stores as those mentioned in Section 2.2.

Concerning parameter values, we have tried various values to evaluate the relationship between parameters and result. Parameters α and β range from 1 to 10 and 10 to 500, respectively. Neighborhood of a part is defined as the circle of radius r ranging from 100 to 1000 meters.

Discretization employed lattices of resolutions ranging from 100×100 to 1000×1000. Since the obtained centers are almost consistent, we only discuss the result of 100×100 lattice in the following.

Figure 5 shows the relationship between parameter values and the result of analysis when α=5. In general, Algorithm CB detects more and larger centers with an increase in r and a decrease in β. Only the number of centers N decreases for large r as shown in Figure 5a because an increase in the size of centers is often accompanied by a decrease in the number of centers.

[Figure 5]

We then examine the details of the result. Table 1 shows the commercial facilities assigned to centers detected when r=300 and β=50. Nine centers C_1-C_9 are detected whose size is all close to 50 given by β.

A wide variation exists in the location of centers as seen in Figure 6. Tags assigned to C_1 and C_2 are found only around railway stations such as Chiba, Nishi-Chiba, Inage, and Kaihin-Makuhari. Center C_1 is characterized by bodies B_1-B_4: Japanese fast-food restaurants, banks, cosmetic stores, and coffee shops. These facilities are generally located around railway stations in suburban areas of Japan. The kinds and location of these facilities implies that center C_1 represents large shopping malls and districts around railway stations. Center C_2 is characterized by kimono and flower shops, whose distribution is quite similar to that of C_1. Since kimono shops are usually located in traditional shopping malls and department stores, center C_2 is considered to represent the traditional shopping districts around Chiba, Inage, and Tsuga. Kaihin-Makuhari has only C_1 tags because the station opened in 1986 and the shopping district around the station was developed in late 1980s.

Tags assigned to centers C_3-C_9 are dispersed all over Chiba City. They represent small shopping districts in and around residential areas. Table 1 suggests a classification of these centers into three groups: {C_3, C_4, C_5, C_6}, {C_7, C_8}, and {C_9}. Centers {C_3, C_4, C_5, C_6} are characterized by convenience stores, beauty shops, and laundry shops while {C_7, C_8} are characterized by sushi restaurants and barber shops. The former represents relatively new while the latter is traditional local shopping malls and districts in residential areas. The former can be further classified into {C_3, C_4} and {C_5, C_6} in Table 1, because Japanese and American pubs are assigned to only {C_3, C_4}. This reflects that local shopping districts sometimes contain local pubs for residents in Japan. Center C_9 is characterized by supermarkets, noodle restaurants, and fast-food

restaurants. It is a typical combination of commercial facilities in shopping centers in suburban and rural areas of Japan. It is confirmed in Figure 6c, though it shows some exceptions around railway stations.

[Table 1]

[Figure 6]

Figure 7 shows the upper half of topology diagrams of centers C_1 and C_2. Bodies B_{13}, B_{14}, B_{15}, B_{16}, and B_{19} are closely located in both diagrams. This implies that many shopping districts represented by C_1 and C_2 contain fast-food restaurants, Japanese and American pubs, convenience stores, and sushi restaurants. Bodies B_{10} and B_{20}, on the other hand, are located relatively far from the above bodies in both diagrams. Only some shopping districts represented by C_1 and C_2 have fruit and vegetable shops and pharmacies.

Bodies located on the left hand side of topology diagram have more similar spatial distributions. Bodies B_{13}, B_{14}, and B_{15} are similar in C_1 while B_{17}, B_{21}, and B_{18} are similar in C_2. The former represent fast-food restaurants, Japanese and American pubs while the latter are beauty shops, barber shops, and laundry shops. It is consistent with earlier discussion on centers C_1-C_6.

[Figure 7]

4. Conclusion

This paper proposes a new general framework for analyzing the relations among distributions of spatial objects. It has at least two advantages over existing methods. First, the framework can deal with more than two distributions of spatial objects. Second, it is applicable independent of the type of spatial objects.

Contribution of the paper compared with previous works (Sadahiro, 2010, 2011, 2012b, Sadahiro and Kobayashi, 2012, and Sadahiro *et al.*, 2012) is summarized as follows. First, the paper generalizes the methods proposed in these papers to be applicable independent of the type of spatial objects. Introduction of the concepts of parts, bodies, and tags permits us to apply the methods to spatial objects not considered in each paper. Second, the paper proposes a method for analyzing the proximal relations among the distributions of spatial objects without overlap such as point distributions. Though spatial proximity is discussed in Sadahiro et al. (2012) and Sadahiro and

Kobayashi (2012), the methods proposed in these papers cannot treat spatial objects without overlap. Spatial objects other than points can also be analyzed in a similar manner as described in the paper by defining parts, bodies, and tags appropriately.

Quantization allows us to represent spatial objects and their spatial relations as a graph, where nodes and links indicate tags and their spatial relations, respectively. This provides a potential for borrowing analytical methods developed in graph theory, network analysis, formal concept analysis, and discrete mathematics. For instance, we can use data mining techniques in formal concept analysis to detect useful and interesting spatial patterns.

Quantization clarifies the role of space in analysis. Space plays a primary role before analysis, that is, definition of parts, bodies, and tags, and their spatial relations. Analysis is performed based on quantized objects where the space is not explicitly considered. This treatment, however, sounds reasonable because space does not directly determine the status and properties of spatial objects in the real world. Space is a medium through which spatial objects affect with each other. Analysis should focus on the relations among spatial objects rather than the relationship between the space and spatial objects.

We finally discuss some limitations and extensions of the paper for future research.

First, the method should be extended to treat spatiotemporal distributions. Though Sadahiro and Kobayashi (2012) discusses spatially distributed time-series data, the method cannot be applicable directly to general spatiotemporal distributions. The method proposed in this paper needs further extension and improvement to deal with a wide variety of spatiotemporal distributions.

Second, spatial relations other than those discussed in Section 2.2 should be further explored. Spatial relations mentioned in the paper are fundamental and simple ones. More complicated relations would be necessary to describe and analyze the complicated real world.

Third, the framework of the method should be extended to confirmatory spatial analysis. This paper proposes the method for the use in exploratory spatial analysis, where we find research hypotheses that are tested in confirmatory spatial analysis. Though a gap usually exists between exploratory and confirmatory spatial analyses, it is desirable to perform both analyses within the same framework. Spatial modeling based on spatial relations should be further discussed.

References

Abraham, A., Hassanien, A. E., and Snasel, V. (2009). *Computational Social Network Analysis: Trends, Tools and Research Advances.* New York: Springer.

Anderson, I. (2002). *Combinatorics of Finite Sets.* New York: Dover Publication.

Birkhoff, G. (1979). *Lattice Theory (3rd Ed.).* Providence, RI: American Mathematical Society

Carrington, P. J., Scott, J., and Wasserman, S. (2005). *Models and Methods in Social Network Analysis.* Cambridge: Cambridge University Press

Clark, P. and Evans, F. (1954). "Distance to nearest neighbor as a measure of spatial relationships in populations." *Ecology*, 35, 445-453.

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

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns.* New York: Oxford University Press Inc.

Dixon, P. (1994). "Testing spatial segregation using a nearest-neighbor contingency table." *Ecology*, 75, 1940-1948.

Goodall, D. W. (1952). "Quantitative aspects of plant distribution." *Biological Reviews*, 27, 194-245.

Greig-Smith, P. (1952). "The use of random and contiguous quadrats in the study of the structure of plant communities." *Annals of Botany*, 16, 293-316.

Haggett, P. and Chorley, R. J. (1969). *Network Analysis in Geography*. London: Edward Arnold.

Kharrat, A., Popa, I. S., Zeitouni, K., and Faiz, S. (2008). "Clustering algorithm for network constraint trajectories." Ruas, A. and Gold, C. M. (eds.) *Headway in Spatial Data Handling 13th International Symposium on Spatial Data Handling, Lecture Notes in Geoinformation and Cartography.* Berlin: Springer, 631-647

Knoke, D. and Yang, S. (2008). *Social Network Analysis.* London: Sage.

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

van Lieshout, M. N. M. and Baddeley, A. J. (1999). Indices of dependence between types in multivariate point patterns *Scandinavian Journal of Statistics* 26 511-532

Okabe, A. and Fujii, A. (1984). "The statistical analysis through a computational method of a distribution of points in relation to its surrounding network." *Environment and Planning A*, 16, 163-171.

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

and Planning A 16 163-171

Okabe, A., Yoshikawa, T., Fujii, A., and Oikawa, K. (1988). "The statistical analysis of a distribution of activity points in relation to surface-like elements." *Environment and Planning A*, 20, 609-620.

Parzen, E. (1962). "On estimation of a probability density function and mode." *Annals of Mathematical Statistics*, 33, 1065-1076.

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

Pielou, E. C. (1961). "Segregation and symmetry in two-species populations as studied by nearest-neighbor relationships." *Journal of Ecology*, 49, 255-269.

Pielou, E. C. (1969). *An Introduction to Mathematical Ecology.* New York: Wiley.

Ripley, B. D. (1976). "The second-order analysis of stationary point process." *Journal of Applied Probability*, 13, 255-266.

Ripley, B. D. (1977). "Modelling spatial patterns." *Journal of the Royal Statistical Society Series B*, 41, 172-192.

Ripley, B. D. (1981). *Spatial Statistics* New York: Wiley.

Rosenblatt, M. (1956). "Remarks on some nonparametric estimates of a density function." *Annals of Mathematical Statistics*, 27, 832-837.

Sadahiro, Y. (1999). "Statistical methods for analyzing the distribution of spatial objects in relation to a surface." *Journal of Geographical Systems*, 1, 107-136.

Sadahiro, Y. (2010). "Analysis of the spatial relations among point distributions on a discrete space." *International Journal of Geographical Information Science*, 24, 997-1014.

Sadahiro, Y. (2011). "Analysis of the relations among spatial tessellations." *Journal of Geographical Systems*, 13, 373-391.

Sadahiro, Y. (2012a). "Object-oriented spatial analysis: Set-based exploratory analysis of the relations among the distributions of spatial objects." *Discussion Paper Series No. 107, Department of Urban Engineering, University of Tokyo* (available from http://ua.t.u-tokyo.ac.jp/pub/due-dp/107.pdf).

Sadahiro, Y. (2012b). "Exploratory analysis of polygons distributed with overlap." *Geographical Analysis*, to appear, (draft version is available from http://ua.t.u-tokyo.ac.jp/pub/due-dp/102.pdf).

Sadahiro, Y., Lay, R., and Kobayashi, T. (2012). "Trajectories of moving objects on a network: detection of similarities, visualization of relations, and classification of

trajectories." *Transactions in GIS* to, appear (draft version is available from http://ua.t.u-tokyo.ac.jp/pub/due-dp/105.pdf).

Sadahiro, Y. and Kobayashi, T. (2012). "Exploratory analysis of spatially distributed time series data: Detection of similarities, clustering and visualization of mutual relations." *Discussion Paper Series No. 108, Department of Urban Engineering, University of Tokyo* (available from http://ua.t.u-tokyo.ac.jp/pub/due-dp/108.pdf).

Shimble, A. (1953). "Structural parameters of communication networks." *Bulletin of Mathematical Biophysics,* 15, 501-507.

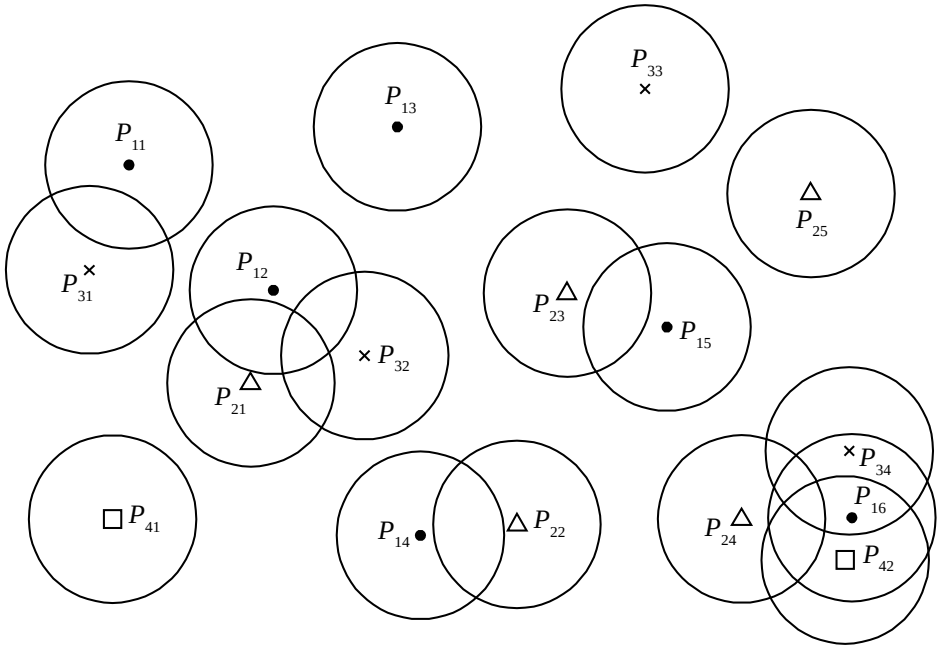
Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

Skellam, J. G. (1952). "Studies in statistical ecology. I. Spatial pattern." *Biometrica,* 39, 346-362.

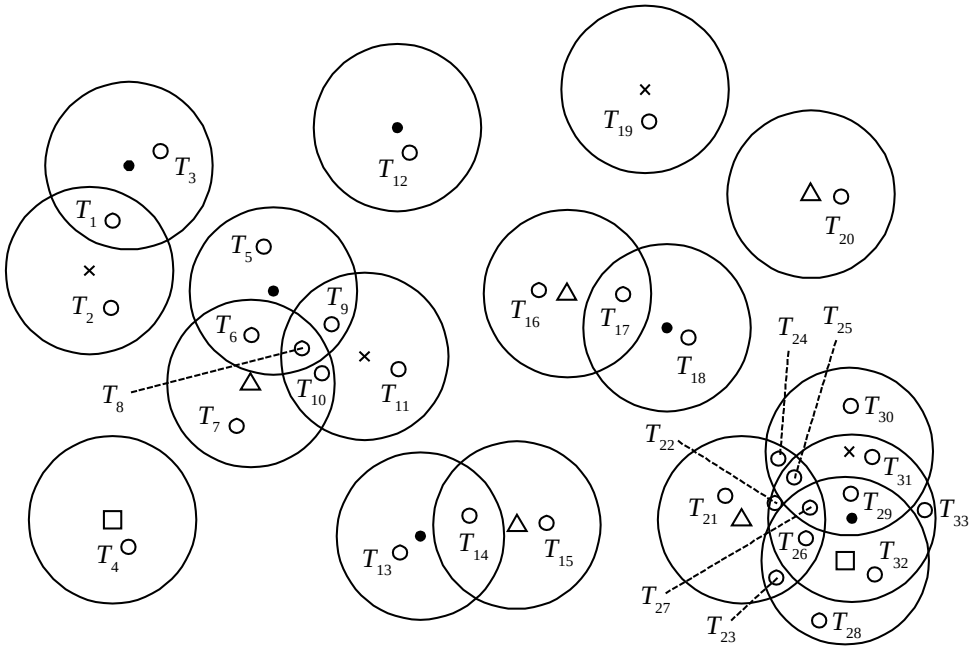
Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications.* Cambridge: Cambridge University Press.

Table 1 Commercial facilities assigned to centers detected when r=300 and β=50.

		C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
B_1	Japanese fast-food restaurants	○								
B_2	Banks	○								
B_3	Cosmetic stores	○								
B_4	Coffee shops	○								
B_5	Liquor shops	○				○			○	
B_6	Book stores	○					○			
B_7	Chinese restaurants	○						○		
B_8	Kimono shops		○							
B_9	Flower shops		○							
B_{10}	Fruits and vegetable shops	○	○							
B_{11}	Japanese noodle restaurants		○							○
B_{12}	Chinese noodle restaurants									○
B_{13}	Fast-food restaurants	○	○	○						○
B_{14}	Japanese pubs	○	○	○	○					
B_{15}	American pubs	○	○	○	○					
B_{16}	Convenience stores	○	○		○	○	○			
B_{17}	Beauty shops		○	○	○	○	○			
B_{18}	Laundry shops		○	○	○	○	○			
B_{19}	Sushi restaurants	○	○					○	○	
B_{20}	Pharmacies	○	○					○	○	
B_{21}	Barber shops		○			○		○	○	
B_{22}	Gas stations							○		
B_{23}	Supermarkets						○			○
B_{24}	Cram schools								○	○
B_{25}	Japanese pub-style restaurants				○					

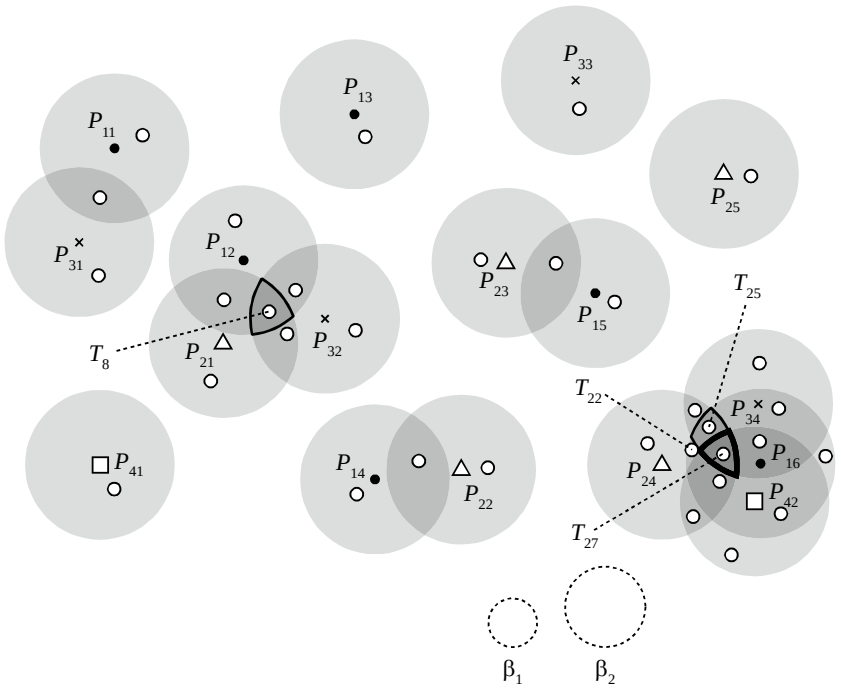


(a)

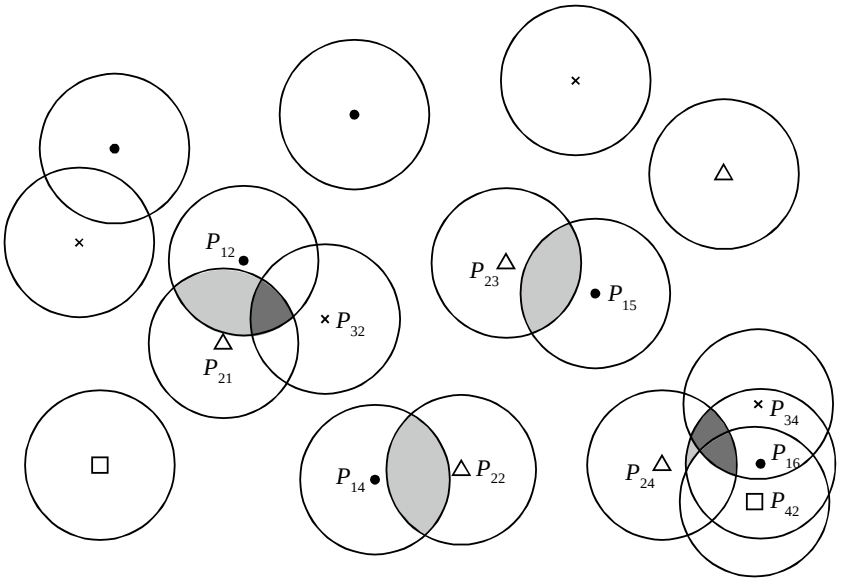


(b)

Figure 1. Point distributions, neighborhoods, and tags. Different symbols indicate different distributions. (a) Point distributions and their neighborhoods. (b) Tags assigned to individual regions.



(a)



(b)

Figure 2. Center detection and body clustering in point distributions. Parameter β is indicated by the area of dotted circles. (a) Tags chosen by Algorithm CB when α=3. Bold and thin lines indicate the first and second set of tags, respectively. (b) Centers detected by Algorithm CB. Dark and light gray shades indicate the centers when α=2 and α=3, respectively.

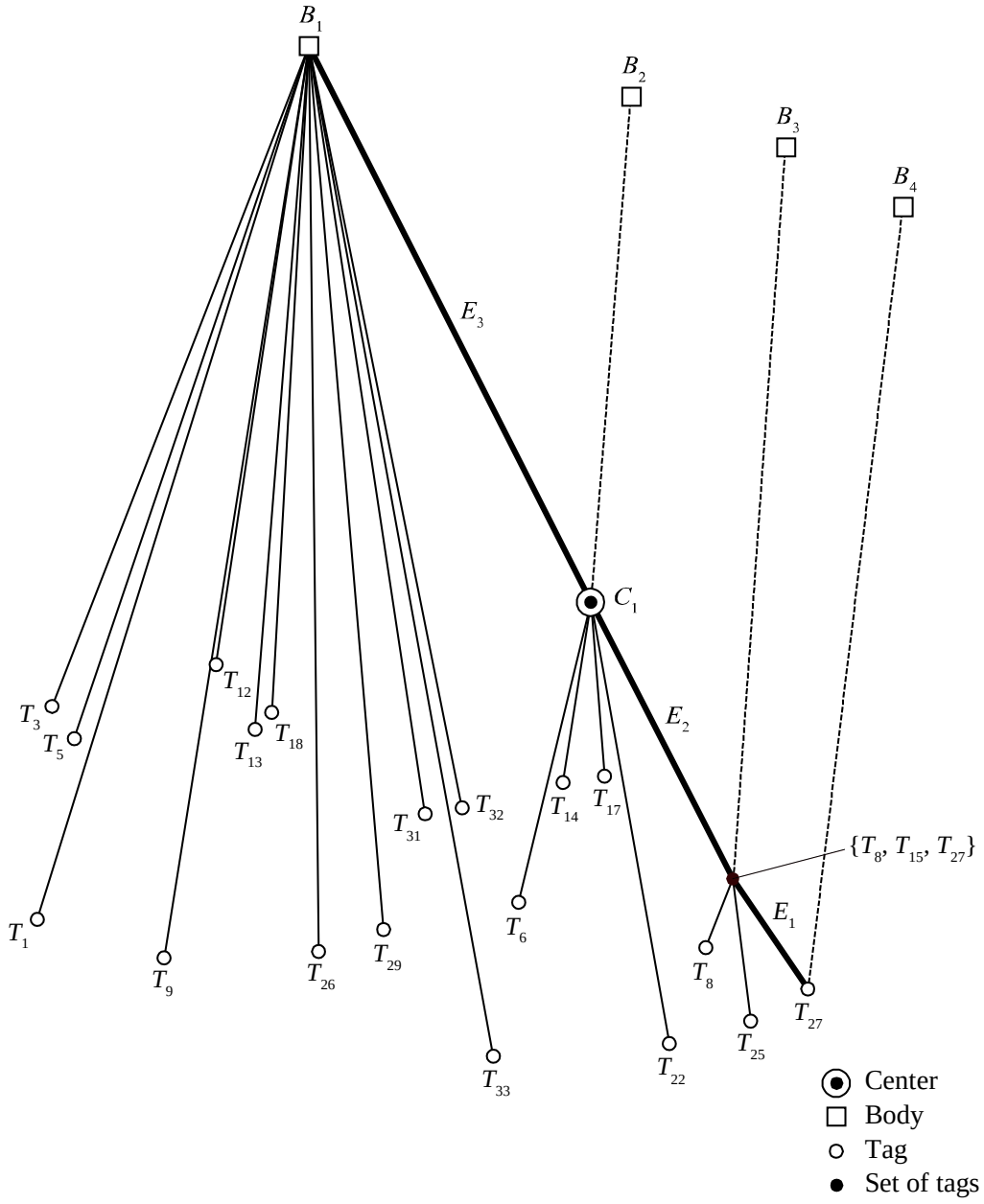


Figure 3. Topology diagram indicating the process of detecting a center in Algorithm CB. Bold and thin solid lines indicate the addition of tags to Θ. Bold solid lines and dotted lines indicate the removal of bodies from Ψ. Edges E_1-E_3 represent the growth of Θ.

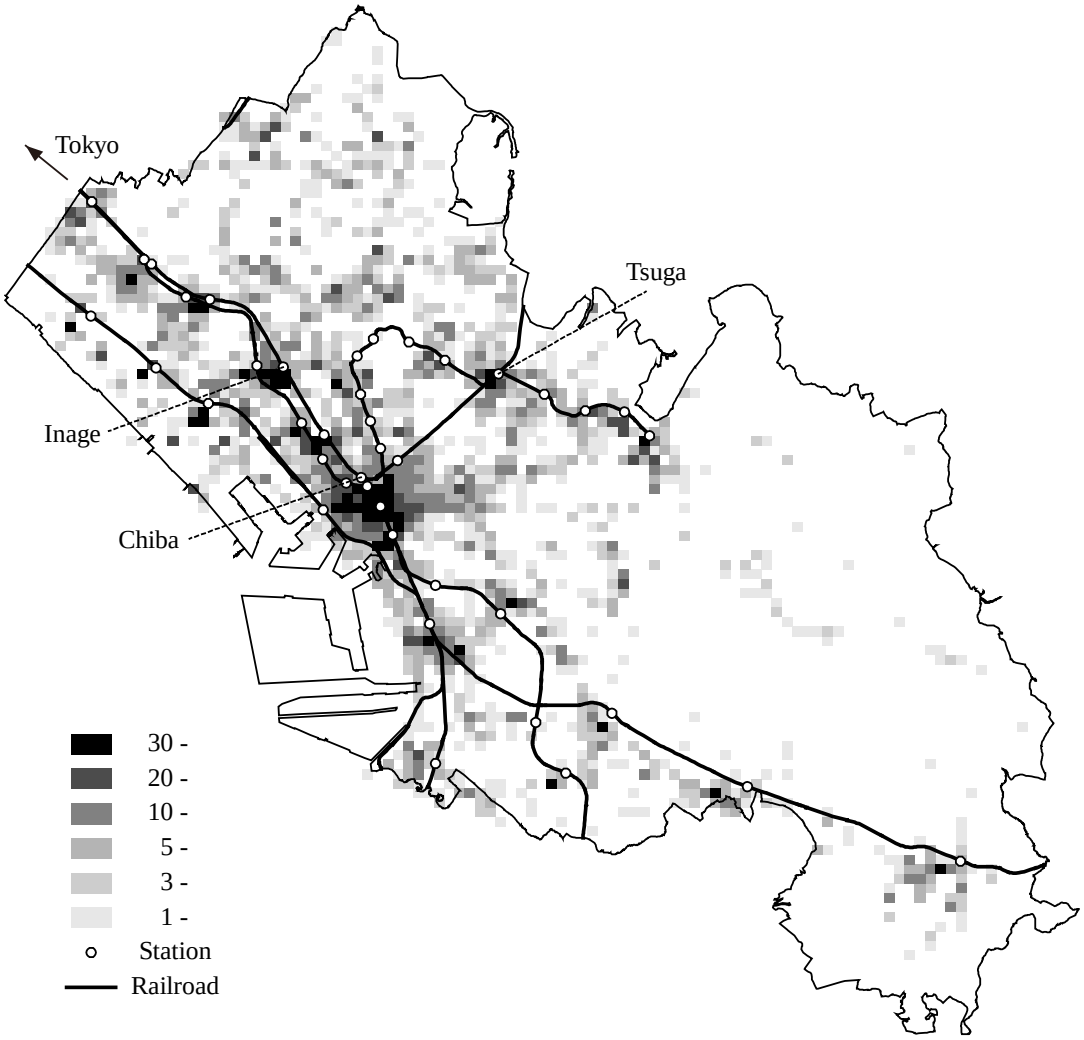


Figure 4. Distribution of commercial facilities in Chiba City, Japan.

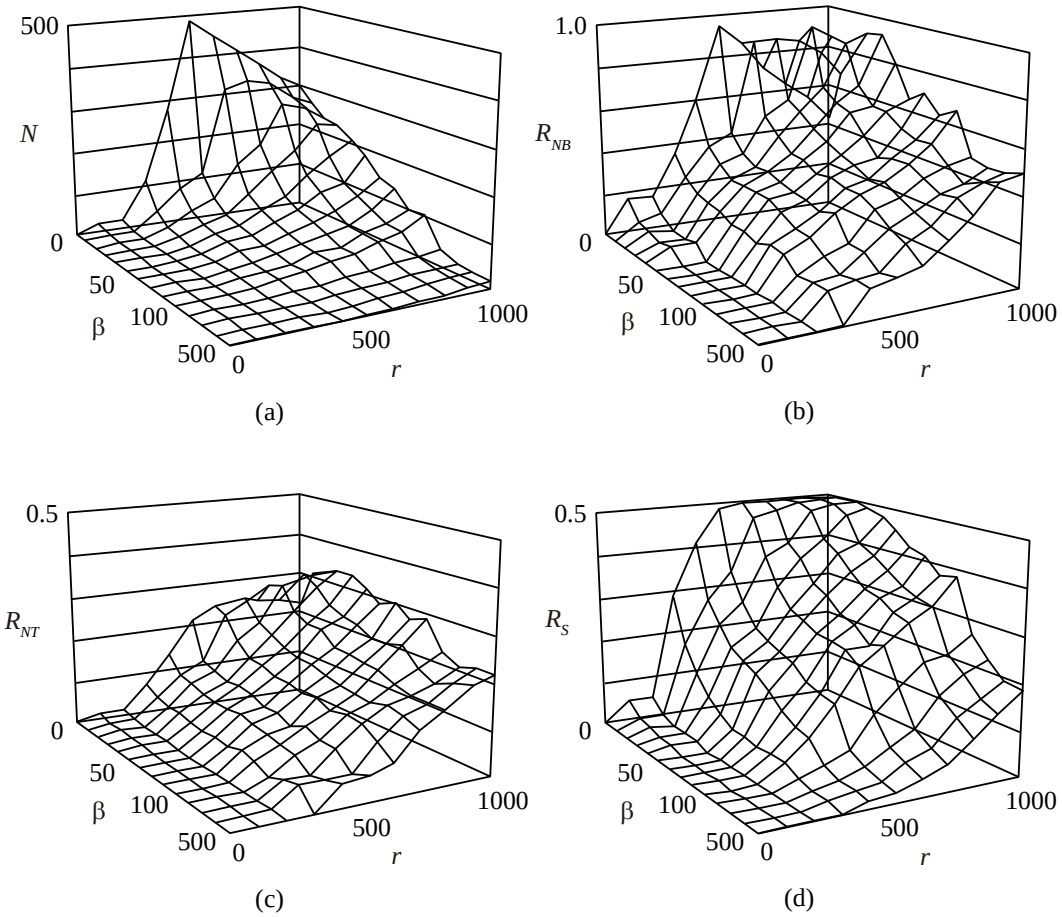
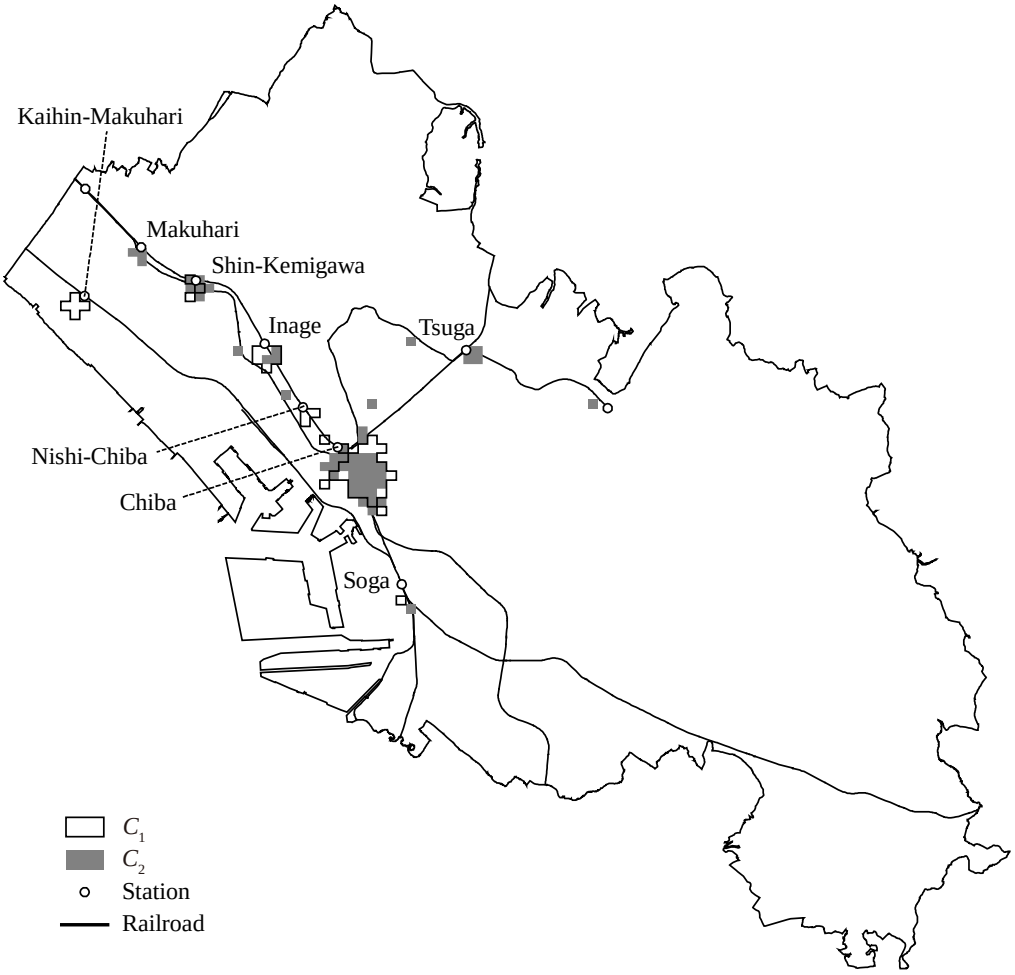
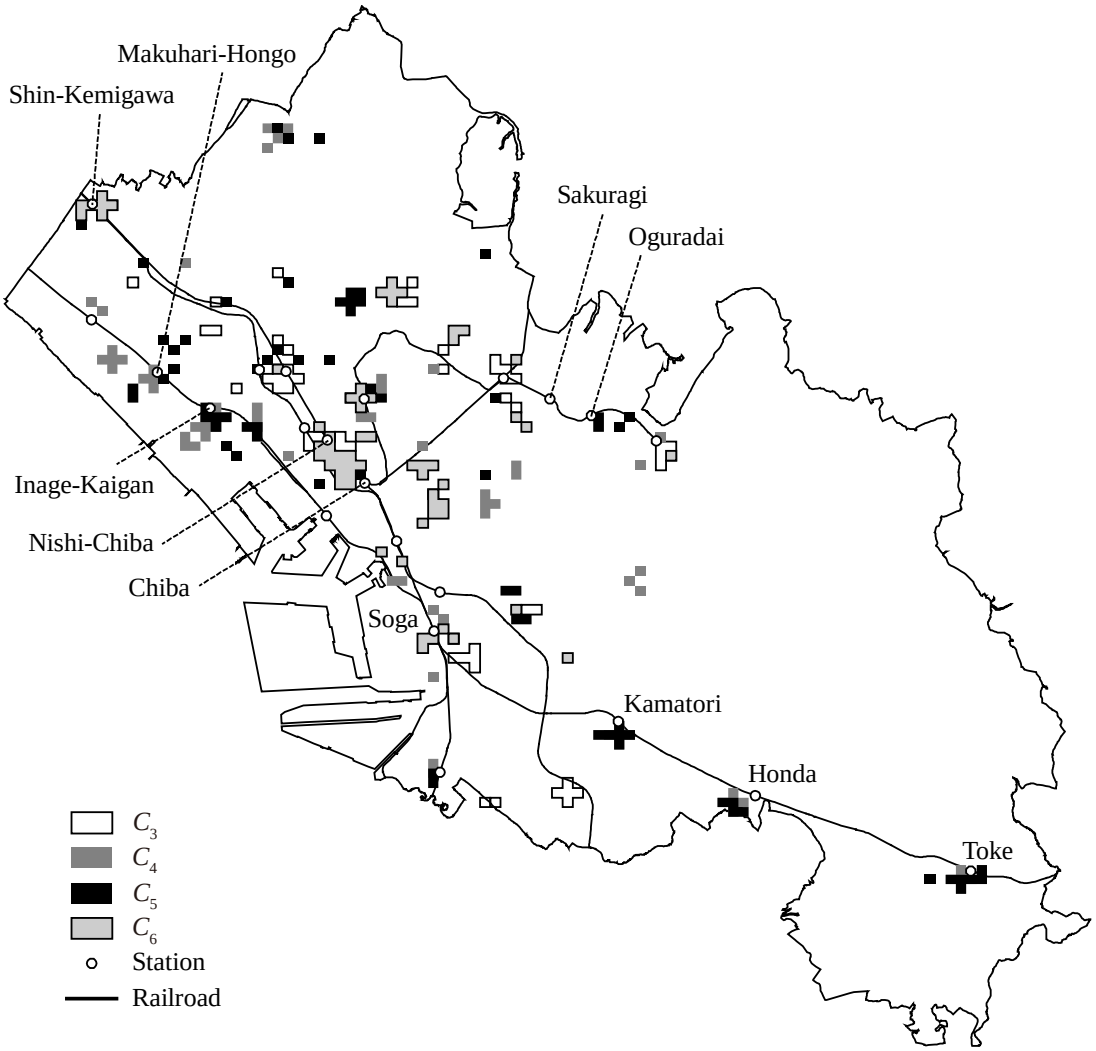


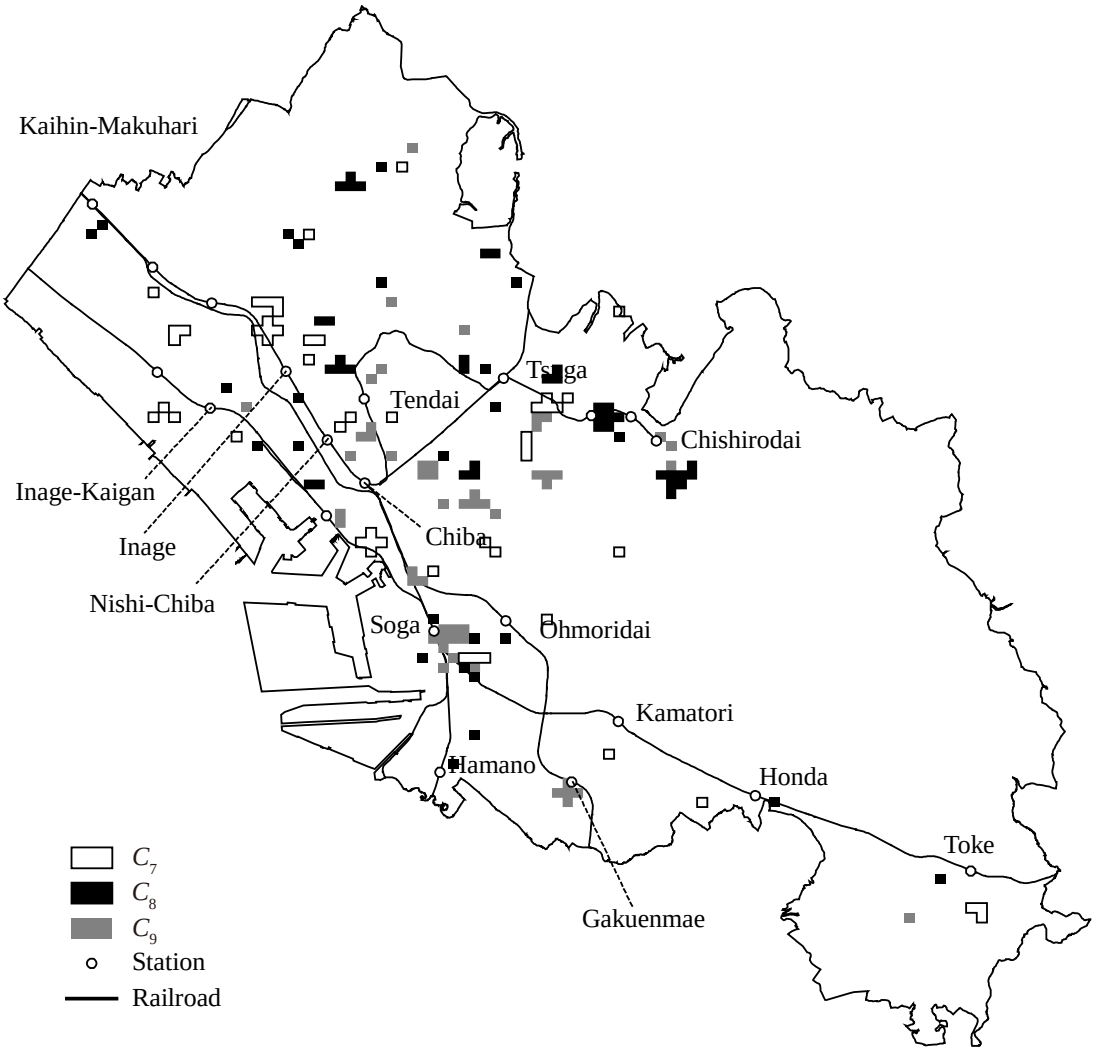
Figure 5. The relationship between parameter values and the result of analysis. (a) The number of centers, (b) the ratio of bodies assigned to centers, (c) the ratio of tags assigned to centers, (d) the standardized size of centers.


Kaihin-Makuhari
Makuhari
Shin-Kemigawa
Inage
Tsuga
Nishi-Chiba
Chiba
Soga
C_1
C_2
Station
Railroad

(a)

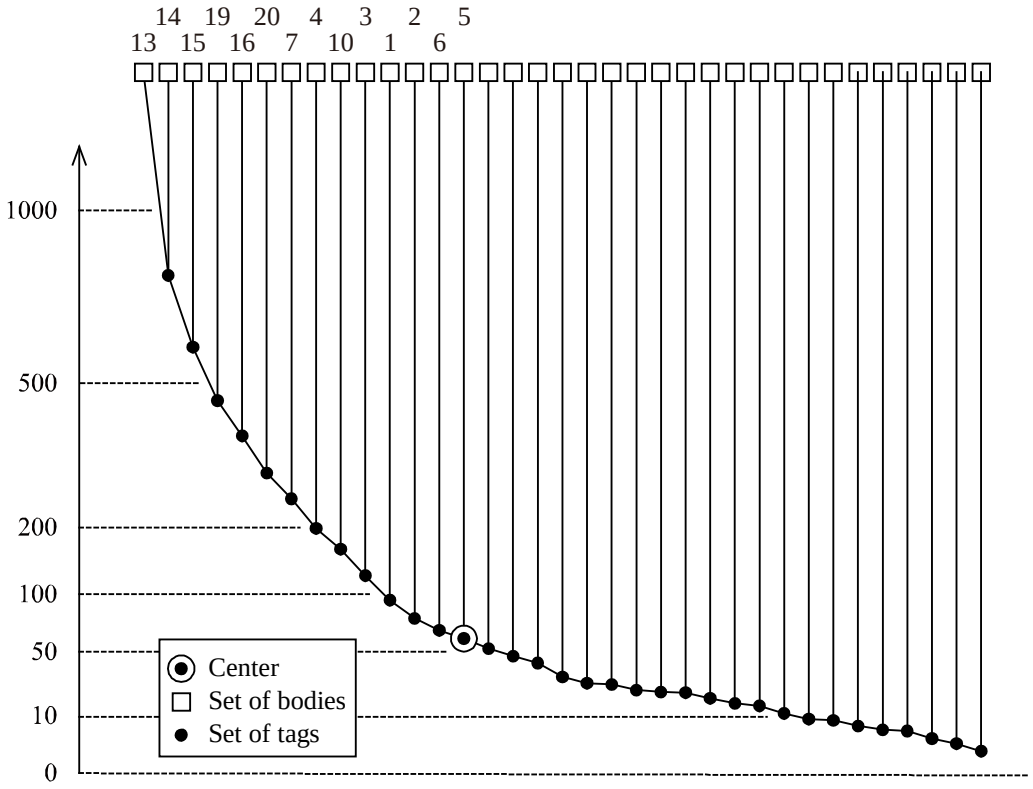

Makuhari-Hongo
Shin-Kemigawa
Sakuragi
Oguradai
Inage-Kaigan
Nishi-Chiba
Chiba
Soga
Kamatori
Honda
Toke
C_3
C_4
C_5
C_6
Station
Railroad

(b)

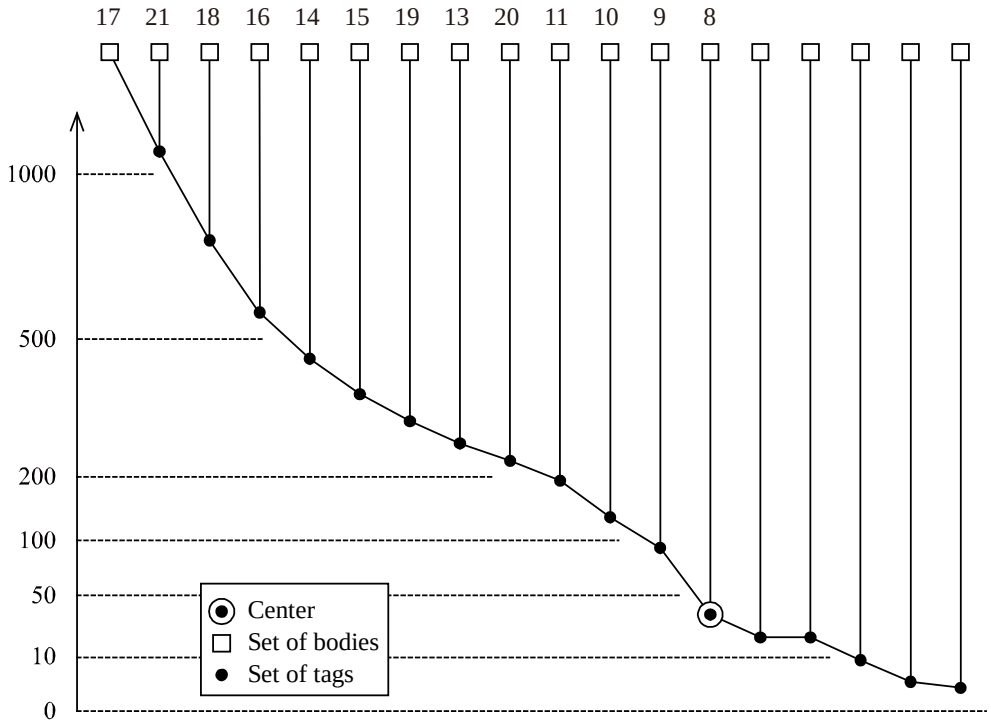


(c)

Figure 6. The location of tags assigned to centers detected when r=300 and β=50. (a) $\{C_1, C_2\}$, (b) $\{C_3, C_4, C_5, C_6\}$, (c) $\{C_7, C_8, C_9\}$.


14 19 20 4 3 2 5
13 15 16 7 10 1 6
1000
500
200
100
50
10
0
Center
Set of bodies
Set of tags

(a)



(b)

Figure 7. Topology diagrams of centers (a) C_1 and (b) C_2. The numbers over white squares indicate the suffix of bodies. The vertical axis indicates the size of spatial objects represented by the number of tags.