CSIS Discussion Paper No. **122**

A method for analyzing the segregation between point distributions

Yukio Sadahiro

September 2013

Center for Spatial Information Science, 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

This paper proposes a new method for analyzing the segregation between point distributions. Though there have been proposed numerous methods and measures in segregation analysis, they have at least three deficiencies: 1) statistical significance of segregation is not evaluated, 2) aspatial properties of points are not considered, and 3) the relationship between different dimensions of segregation is not fully discussed. To resolve these problems, this paper proposes a new method for analyzing the segregation between point distributions. We introduce a general procedure of evaluating the individual components of segregation. This procedure helps us find independent components of segregation, and provides a means of assessing their statistical significance. To test the validity of the proposed method, we apply it to three datasets of different sizes. The result supports the technical soundness of the method, and provides empirical findings.

Keywords: segregation; point distributions; statistical analysis; aspatial properties

1. Introduction

Segregation is a fundamental concept in geographical information science. In geographical context it refers to the spatial isolation of different types of objects. Ethnic segregation has drawn much attention in geography, demography, and sociology (Duncan and Duncan 1955; Morgan 1983; White 1983, 1986; James and Taeuber 1985; Massey 1985; Morrill 1995). Segregation of plant and animal species has been widely discussed in biology and ecology (O'Neill *et al.* 1988; Li and Reynolds 1993; Riitters *et al.* 1996; Mucientes *et al.* 2009).

There have been proposed numerous measures to evaluate the degree of segregation. Though earlier measures are easy to understand and calculate, they have been often criticized for the modifiable areal unit problem (Wong 1993, 2001; Reardon and O’Sullivan 2004). Recent methods are more sophisticated (Johnston *et al.* 2007, 2011; Poulsen *et al.* 2007; Rey and Folch 2011; Páez *et al.* 2012), and some measures are defined based on the location of individual objects to avoid the modifiable areal unit problem (Reardon and O’Sullivan 2004; Reardon *et al*. 2009; Reardon and Bischoff 2011; Sadahiro and Hong 2013). However, there still remain at least three problems to be solved: 1) statistical evaluation of segregation, 2) consideration of aspatial properties, and 3) independent dimensions of segregation. We discuss these problems successively in the following.

Statistical tests permit us to judge whether or not segregation can occur by chance. Some papers evaluate the statistical significance of segregation by using the measures of spatial autocorrelation. Johnston *et al.* (2011) and Poulsen *et al.* (2011) use global Moran's *I* to evaluate the segregation between ethnic groups. Logan *et al.* (2002) and Brown and Chung (2006) adopt local Moran's *I* to identify local clusters of ethnic and racial groups. However, these measures are not appropriate for the segregation between multiple groups since they were originally developed to evaluate the spatial autocorrelation of a single group of objects. Though these measures can evaluate the segregation of one group from the others, they cannot assess the statistical significance of the entire segregation between multiple groups. Other measures of spatial correlation such as those proposed by Wartenberg (1985) and Stephane *et al.* (2008) evaluate the segregation between every pair of groups, not the entire segregation between multiple groups.

Another aspect overlooked in existing studies is the aspatial properties of points. It often happens in the real world that one group is more similar to another than the others. Austronesian peoples are more similar to Papuan peoples in some aspects than Indo-European peoples. Segregation between Austronesian and Indo-European peoples may be more crucial than that between Austronesian and Papuan peoples. However, such aspatial properties of points have not been considered explicitly in segregation analysis.

The third problem lies in the discussion of segregation dimensions. Massey and Denton (1988) defines dimensions as the primary axes for measuring the degree of segregation and advocated five dimensions called evenness, exposure, concentration, centralization, and clustering.

Reardon and O’Sullivan (2004) claims that these dimensions are not fully distinct and proposes two independent dimensions called spatial exposure and spatial evenness. Johnston *et al*. (2007) uses principal component analysis in case studies to find two independent dimensions called separateness and location. However, the independence between these dimensions has not yet been fully examined. Reardon and O’Sullivan (2004) claims the independence of their dimensions without showing a theoretical proof, and empirical approach taken by Johnson *et al*. (2007) does not assure that the two dimensions exist in any circumstances.

To resolve the above problems, this paper proposes a new method for analyzing the segregation between point distributions. Though our method is developed based on the information about the location of individual points, it is also effective for aggregated spatial data that have been frequently used in the literature. Section 2 introduces a general procedure of evaluating the individual components of segregation. This procedure helps us find independent components of segregation, and provides a means of assessing their statistical significance. To test the validity of the proposed method, Section 3 applies it to the analysis of two synthetic datasets and one real dataset. Section 4 summarizes the conclusions with discussion.

2. Methodology

This paper introduces two terms *factors* and *components* to refer to the causes and results of segregation, respectively. The latter almost correspond to dimensions discussed in existing studies such as evenness, exposure, and clustering. *Independent components* refer to the components that are mutually exclusive and collectively exhaustive, i.e., components that form the entire segregation without overlapping. This definition is equivalent to those used in Reardon and O’Sullivan (2004) and Johnson *et al*. (2007).

Sadahiro and Hong (2013) proposes a method for evaluating segregation based on the location of individual points. We extend their method to resolve the three problems mentioned in Section 1. This section starts with the measurement of segregation without considering the aspatial properties of points. We then extend the method by incorporating the aspatial properties of points. This permits us to evaluate individual components separately, and helps us find independent components of segregation. We finally propose a method for testing the statistical significance of segregation.

2.1 Measurement of segregation without considering the aspatial properties of points

Suppose a set of K types of points in region R of area T. Let P_{ij} and $\mathbf{z}_{ij}$ be jth point of type i and its location, respectively. The number of type i points and its summation are denoted by N_i and N=N_1+ N_2+... +N_K, respectively. The set of type i points is denoted by $\boldsymbol{P}_i$={P_{i1}, P_{i2}, ..., P_{iNi}}.

Reardon and O’Sullivan (2004) and Sadahiro and Hong (2013) evaluate the segregation between points by using their density distributions. Following this line, we define $D_i(\mathbf{x})$ as the density function of $\boldsymbol{P}_i$ at location $\mathbf{x}$. The set of the density functions is denoted by $\boldsymbol{D}=\{D_1(\mathbf{x}), D_2(\mathbf{x}), ..., D_K(\mathbf{x})\}$. Let $\phi(\mathbf{x}, \mathbf{z}_{ij})$ be a proximity function that indicates the spatial proximity between location $\mathbf{x}$ and point P_{ij}. One definition of $\phi(\mathbf{x}, \mathbf{z}_{ij})$ is a distance-decay function such as

$$\phi\left(\mathbf{x},\mathbf{z}_{ij}\right)=\exp\left(-\alpha\left|\mathbf{x}-\mathbf{z}_{ij}\right|\right).\tag{1}$$

The density of type i points at $\mathbf{x}$ is defined as the summation of proximity functions of all the type i points:

$$D_i\left(\mathbf{x}\right)=\frac{N_i\sum_j\phi\left(\mathbf{x},\mathbf{z}_{ij}\right)}{\int_{\mathbf{y}\in R}\sum_j\phi\left(\mathbf{y},\mathbf{z}_{ij}\right)d\mathbf{y}}.\tag{2}$$

This definition standardizes the summation of proximity functions in such a way that the integration of a density function is equal to the number of points, i.e.,

$$\int_{\mathbf{x}\in R}D_i\left(\mathbf{x}\right)d\mathbf{x}=N_i.\tag{3}$$

The above definition is based on the location of individual points. If the locational data of points are aggregated by spatial units, we substitute the density of points in individual units for $D_i(\mathbf{x})$. The method described in the following is applicable to both aggregated and disaggregated spatial data.

Using the density functions of points, we define the local segregation at $\mathbf{x}$ as

$$\begin{aligned}s\left(\mathbf{x};\boldsymbol{D}\right)&=\sum_i\left\{\frac{D_i\left(\mathbf{x}\right)}{\sum_j D_j\left(\mathbf{x}\right)}\right\}^2\\&=\frac{\sum_i\left\{D_i\left(\mathbf{x}\right)\right\}^2}{\left\{\sum_j D_j\left(\mathbf{x}\right)\right\}^2}.\end{aligned}\tag{4}$$

We adopt this definition for its simplicity and flexibility (Sadahiro and Hong, 2013). Integrating Equation (4), we obtain the global measure of segregation:

$$S\left(\boldsymbol{D}\right)=\frac{\int_{\mathbf{x}\in R}D\left(\mathbf{x}\right)s\left(\mathbf{x};\boldsymbol{D}\right)\mathrm{d}\mathbf{x}}{\int_{\mathbf{x}\in R}D\left(\mathbf{x}\right)\mathrm{d}\mathbf{x}},\tag{5}$$

where

$$D\left(\mathbf{x}\right)=\sum_i D_i\left(\mathbf{x}\right).\tag{6}$$

The measure ranges from zero to one. A large value appears when different groups are separated while a small value indicates the integration of different groups.

2.2 Measurement of segregation with considering the aspatial properties of points

This subsection extends the proposed measures to incorporate the aspatial properties of points. Let us suppose the segregation of black, dark-grey, light-grey, and white points. Dark-grey and light-grey points are the composites of black and white points at 70/30 and 30/70 ratios, respectively. The properties of points are represented by the proportion of black and white, i.e., {1.0, 0.0}, {0.7, 0.3}, {0.3, 0.7}, and {0.0, 1.0}.

To formalize this example, we introduce a positive $K\times M$ matrix **A** that represents the aspatial properties of points:

$$\mathbf{A}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1M} \\ a_{21} & a_{22} & \cdots & a_{2M} \\ \vdots & \vdots & & \vdots \\ a_{K1} & a_{K2} & \cdots & a_{KM} \end{bmatrix}.\tag{7}$$

Element a_{im} indicates the mth attribute of type i points. We assume that every set of variables $\{a_{1m}, a_{2m}, ..., a_{Km}\}$ is given in a standardized form, i.e.,

$$\sum_i a_{i1}=\sum_i a_{i2}=\cdots\sum_i a_{iM}\ .\tag{8}$$

The attributes of four types of points mentioned above is represented by

$$\mathbf{A}=\begin{bmatrix} 1.0 & 0.0 \\ 0.7 & 0.3 \\ 0.3 & 0.7 \\ 0.0 & 1.0 \end{bmatrix}.\tag{9}$$

Here, we consider the spatial distribution of attributes instead of that of points themselves. The spatial distribution of *m*th attribute is given by

$$A_m(\mathbf{x}) = \sum_i a_{im} D_i(\mathbf{x}) .$$

(10)

We replace $\boldsymbol{D}$={$D_1(\mathbf{x})$, $D_2(\mathbf{x})$, ..., $D_K(\mathbf{x})$} by {$A_1(\mathbf{x})$, $A_2(\mathbf{x})$, ..., $A_M(\mathbf{x})$} to evaluate the segregation of point distributions. The local and global measures of segregation become

$$\begin{aligned} s(\mathbf{x};\boldsymbol{D},\mathbf{A}) &= \sum_i \left\{ \frac{A_i(\mathbf{x})}{\sum_j A_j(\mathbf{x})} \right\}^2 \\ &= \sum_i \left\{ \frac{\sum_j a_{ji} D_j(\mathbf{x})}{\sum_j \sum_k a_{kj} D_k(\mathbf{x})} \right\}^2 \end{aligned}$$

(11)

and

$$S(\boldsymbol{D},\mathbf{A}) = \frac{\int_{\mathbf{x}\in R} D(\mathbf{x}) s(\mathbf{x};\boldsymbol{D},\mathbf{A}) \mathrm{d}\mathbf{x}}{\int_{\mathbf{x}\in R} D(\mathbf{x}) \mathrm{d}\mathbf{x}} ,$$

(12)

respectively. We call the latter the *overall segregation*.

The measure $S(\boldsymbol{D})$ is a special case of $S(\boldsymbol{D}, \mathbf{A})$. The former is obtained by defining **A** as the $K{\times}K$ identity matrix in Equation (12):

$$\mathbf{A_I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} .$$

(13)

So far we have assumed attributes to be represented by the proportion of elements. However, the proposed measures are effective for a wider range of properties measured on a ratio scale. The only requirement for attributes is that they can be standardized as shown in Equation (8).

2.3 Evaluation of individual components of segregation

This subsection proposes a method for evaluating individual components of segregation. To consider a certain component, we compare two situations where its underlying factors are present and absent. The difference in the overall segregation $S(\boldsymbol{D}, \mathbf{A})$ between the two situations reflects the effect of the factors, and consequently, provides us a means of evaluating the component.

Suppose point distributions in Figure 1a, where black and white points are uniformly distributed. In Figure 1b, white points are uniformly distributed while black points are distributed only on the left side. Segregation is caused by the non-uniform distribution of black points. Consequently, the difference in $S(\boldsymbol{D}, \mathbf{A})$ between the two patterns indicates the effect of non-uniformity in black points on segregation.

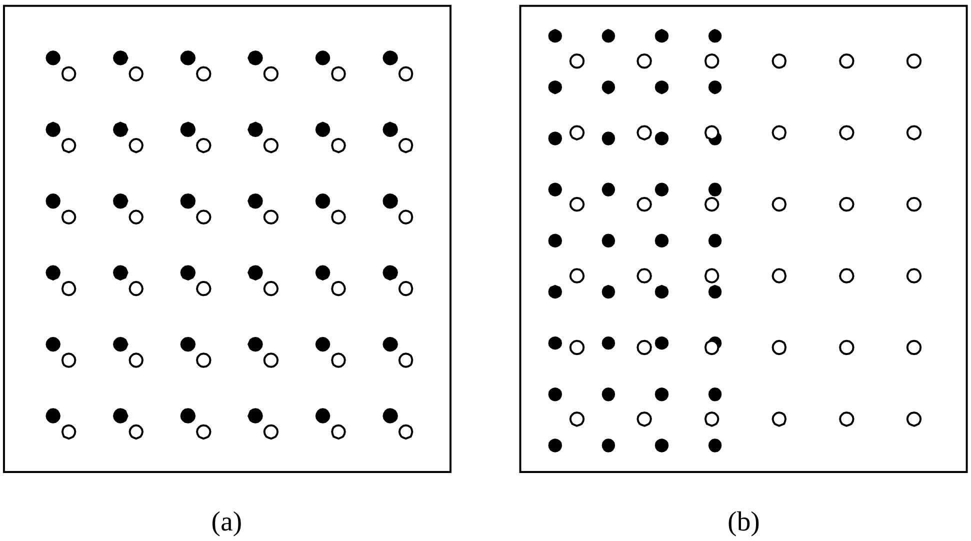


Figure 1 The distributions of black and white points. (a) Both black and white points are uniformly distributed. (b) Black points are distributed only on the left side, which causes segregation between points.

We evaluate the effect of non-uniformity in type i points by considering the situation where type i points are uniformly distributed. The set of density function $\boldsymbol{D'}$ is given by

$$\boldsymbol{D'} = \left\{ D_1(\mathbf{x}), D_2(\mathbf{x}), \ldots, \frac{N_i}{T}, \ldots, D_M(\mathbf{x}) \right\}. \tag{14}$$

The effect of the non-uniformity in type i points is measured by

$$\Delta S(\boldsymbol{D}, \mathbf{A}) = S(\boldsymbol{D}, \mathbf{A}) - S(\boldsymbol{D'}, \mathbf{A}). \tag{15}$$

We then suppose Figure 2 where the different number of black and white points are uniformly distributed. Though Figure 2a exhibits no obvious segregation, segregation occurs at a local level in Figure 2b. Points are integrated only in the neighborhood of white points, and we can find many places where white points are absent. Segregation occurs and increases with the difference in the number of points.

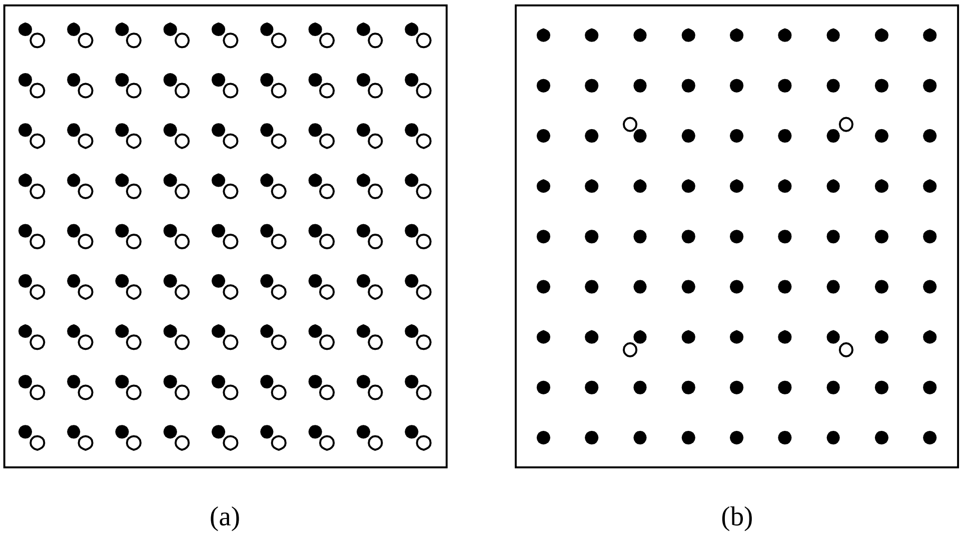


Figure 2 The uniform distributions of black and white points. (a) The same number of black and white points are distributed. (b) White points are fewer than black points. Segregation occurs at a local level where white points are absent.

We evaluate the effect of the difference in the number of type *i* points by replacing N_i with the average number of other types of points:

$$\frac{\sum_{j \neq i} N_j}{K-1} = \frac{N - N_i}{K-1} . \tag{16}$$

The density functions become

$$\boldsymbol{D}'' = \left\{ D_1(\mathbf{x}), D_2(\mathbf{x}), ..., \frac{N - N_i}{(K-1) N_i} D_i(\mathbf{x}), ..., D_K(\mathbf{x}) \right\} . \tag{17}$$

The effect is measures by

$$\Delta S(\boldsymbol{D}, \mathbf{A}) = S(\boldsymbol{D}, \mathbf{A}) - S(\boldsymbol{D}'', \mathbf{A}) . \tag{18}$$

As seen above, evaluation of a certain component requires us to compare two situations where its factors are present and absent. However, it is sometimes difficult to remove only relevant factors without affecting the others. To resolve the problem, we propose an indirect method for evaluating a component of segregation.

Suppose two components C_1 and C_2 each of which is caused by different factors. If we can remove only the factors of C_1, we can evaluate C_2 even if we cannot remove the factors of C_2 separately. Let S and S' be the overall segregation of two situations where the factors of both C_1 and C_2 are present and absent, respectively. The difference S-S' represents C_1+C_2. Let S'' be the overall segregation where the factors of C_1 are absent. The difference S-S'' represents C_1. Consequently, we can evaluate C_2 by substituting S-S'' from C_1+C_2:

$$\begin{aligned} C_2 &= \left(S - S'\right) - \left(S - S''\right) \\ &= S'' - S' \end{aligned} . \tag{19}$$

2.4 Primary factors and components of segregation

The proposed procedure permits us to evaluate the individual components of segregation. This subsection discusses four primary components of segregation using this procedure: spatial segregation, compositional segregation, qualitative segregation, and attribute segregation. They are caused by four factors what we call spatial unevenness, compositional unevenness, qualitative uniformity, and attribute similarity, respectively. We will discuss the above factors and components successively in the following.

Spatial unevenness refers to the non-uniformity in point distributions, which causes spatial segregation. To evaluate spatial segregation, we can take either absolute or relative approach. Absolute approach evaluates absolute spatial unevenness, which refers to the fluctuation of each distribution. Relative approach considers relative spatial unevenness, which refers to the fluctuation of the proportion of each distribution to all the distributions. Though both approaches yield the same segregation measures, they provide different bases for statistical tests as discussed later.

Absolute approach evaluates spatial segregation by considering the situation where absolute spatial unevenness does not exist. It is represented by the set of uniform density functions:

$$\boldsymbol{D}_{\boldsymbol{AS}} = \left\{ \frac{N_1}{T}, \frac{N_2}{T}, \ldots, \frac{N_K}{T} \right\} . \tag{20}$$

Suppose the density distributions of points defined on a one-dimensional space shown in Figure 3a. Figure 3b shows the situation where absolute spatial unevenness is absent. Overall segregation reduces from $S(\boldsymbol{D}, \mathbf{A})$ to

$$S\left(\boldsymbol{D}_{\boldsymbol{AS}}, \mathbf{A}\right) = \sum_i \left(\frac{N_i}{N}\right)^2 .$$

(21)

The reduction represents *absolute spatial segregation*:

$$\begin{aligned} S_{AS}\left(\boldsymbol{D},\mathbf{A}\right) &= S\left(\boldsymbol{D},\mathbf{A}\right) - S\left(\boldsymbol{D_{AS}},\mathbf{A}\right) \\ &= S\left(\boldsymbol{D},\mathbf{A}\right) - \sum_i \left(\frac{N_i}{N}\right)^2 \end{aligned} .$$

(22)

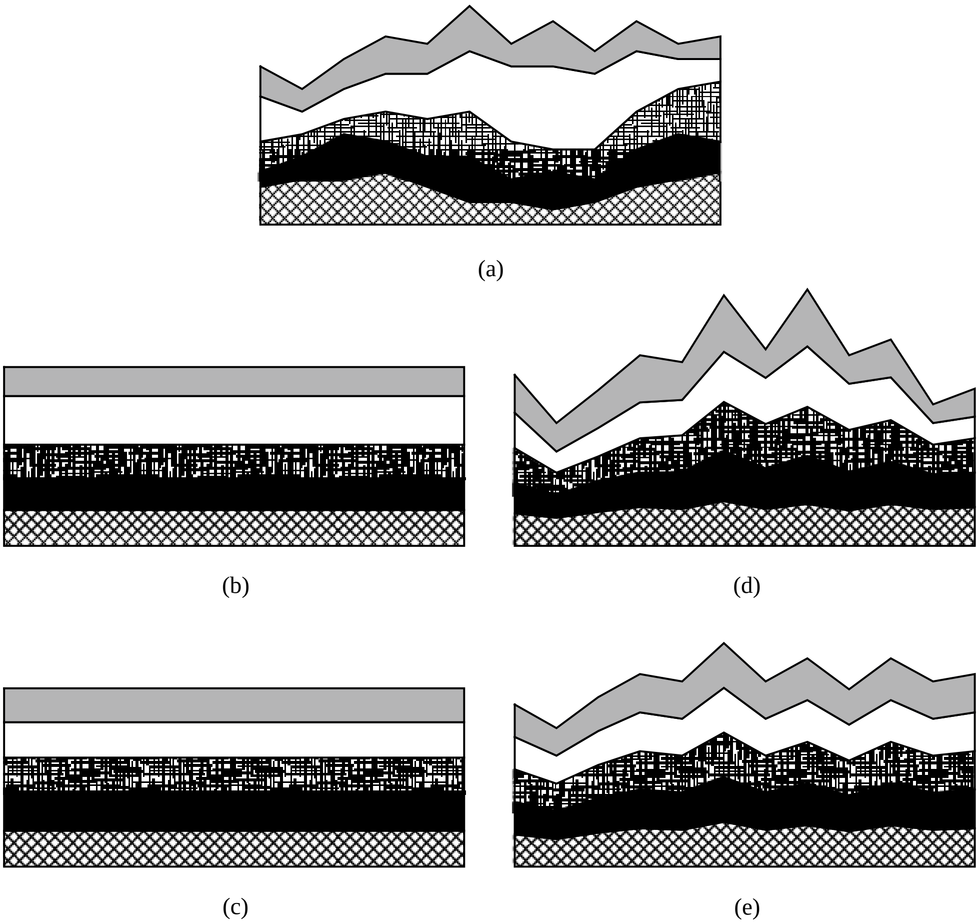


Figure 3 Evaluation of spatial and compositional segregations on a one-dimensional space. Figure 3a shows the density distributions of points to be evaluated while the others indicate density distributions where some factors are absent. Absent factors are (b) absolute spatial unevenness, (c) absolute spatial unevenness and compositional unevenness, (d) relative spatial unevenness, (e) relative spatial unevenness and compositional unevenness.

Relative approach introduces another representation of overall segregation. Let ***r*** be a set of variables indicating the ratio of $D_i(\mathbf{x})$ to $D(\mathbf{x})$:

$$\begin{aligned}\boldsymbol{r} &= \left\{ r_1(\mathbf{x}), r_2(\mathbf{x}), \dots, r_K(\mathbf{x}) \right\} \\ &= \left\{ \frac{D_1(\mathbf{x})}{D(\mathbf{x})}, \frac{D_2(\mathbf{x})}{D(\mathbf{x})}, \dots, \frac{D_K(\mathbf{x})}{D(\mathbf{x})} \right\} .\end{aligned}$$

(23)

The set of density distributions and their overall segregation are represented by

$$\boldsymbol{D} = \left\{ r_1(\mathbf{x}) D(\mathbf{x}), r_2(\mathbf{x}) D(\mathbf{x}), \dots, r_K(\mathbf{x}) D(\mathbf{x}) \right\}$$

(24)

and

$$S(\boldsymbol{D}, \mathbf{A}) = S(\boldsymbol{r}, D(\mathbf{x}), \mathbf{A}),$$

(25)

respectively.

Relative approach evaluates spatial segregation by considering the situation where the proportion of every type of points is uniform in R. It is represented by

$$\boldsymbol{r_{RS}} = \left\{ \frac{N_1}{N}, \frac{N_2}{N}, \dots, \frac{N_K}{N} \right\} .$$

(26)

Figure 3d shows the absence of relative spatial unevenness. As seen in this figure, relative approach keeps the distribution of the total density of points. *Relative spatial segregation* is the reduction of overall segregation from $S(\boldsymbol{r}, D(\mathbf{x}), \mathbf{A})$ to $S(\boldsymbol{r_{RS}}, D(\mathbf{x}), \mathbf{A})$:

$$\begin{aligned}S_{RS}(\boldsymbol{D}, \mathbf{A}) &= S(\boldsymbol{r}, D(\mathbf{x}), \mathbf{A}) - S(\boldsymbol{r_{RS}}, D(\mathbf{x}), \mathbf{A}) \\ &= S(\boldsymbol{D}, \mathbf{A}) - \sum_i \left(\frac{N_i}{N} \right)^2 .\end{aligned}$$

(27)

Since $S_{AS}(\boldsymbol{D}, \mathbf{A})=S_{RS}(\boldsymbol{D}, \mathbf{A})$, we define *spatial segregation* as

$$S_S(\boldsymbol{D}, \mathbf{A}) = S_{AS}(\boldsymbol{D}, \mathbf{A}) = S_{RS}(\boldsymbol{D}, \mathbf{A}) = S(\boldsymbol{D}, \mathbf{A}) - \sum_i \left(\frac{N_i}{N} \right)^2 .$$

(28)

Compositional unevenness refers to the variation in the number of points between different groups, which leads to compositional segregation. Compositional unevenness is another critical factor of segregation as seen in Figure 2. We take the indirect method to evaluate compositional segregation because we cannot remove compositional unevenness without affecting the other factors.

Suppose the situation where both spatial and compositional unevenness are absent. In absolute approach, it is represented by Figure 3c where all the density distributions are defined by the same uniform function:

$$\boldsymbol{D}_{\boldsymbol{ASC}} = \left\{ \frac{N}{KT}, \frac{N}{KT}, \dots, \frac{N}{KT} \right\}. \tag{29}$$

The reduction of overall segregation is

$$S\left(\boldsymbol{D}, \mathbf{A}\right) - S\left(\boldsymbol{D}_{\boldsymbol{ASC}}, \mathbf{A}\right) = S\left(\boldsymbol{D}, \mathbf{A}\right) - \frac{1}{K}. \tag{30}$$

Relative approach changes Figure 3a into Figure 3e. The latter are represented by

$$\boldsymbol{r}_{\boldsymbol{RSC}} = \left\{ \frac{1}{K}, \frac{1}{K}, \dots, \frac{1}{K} \right\}. \tag{31}$$

The reduction of overall segregation is

$$S\left(\boldsymbol{D}, \mathbf{A}\right) - S\left(\boldsymbol{r}_{\boldsymbol{RSC}}, D\left(\mathbf{x}\right), \mathbf{A}\right) = S\left(\boldsymbol{D}, \mathbf{A}\right) - \frac{1}{K}. \tag{32}$$

Equations (30) and (32) represent the effect of spatial and compositional unevenness, i.e., the summation of spatial and compositional segregations. We thus subtract spatial segregation from these equations to measure *compositional segregation*:

$$\begin{aligned} S_C\left(\boldsymbol{D}, \mathbf{A}\right) &= S\left(\boldsymbol{D}, \mathbf{A}\right) - \frac{1}{K} - S_{AS}\left(\boldsymbol{D}, \mathbf{A}\right) \\ &= \sum_i \left(\frac{N_i}{N} \right)^2 - \frac{1}{K} \end{aligned}. \tag{33}$$

This equation indicates that compositional segregation is defined as a function of only aspatial variables. Compositional segregation evaluates the aspatial aspect of segregation.

We then discuss the third factor of segregation what we call qualitative uniformity. It refers to the lack of the diversity of points advocated by Sadahiro and Hong (2013). When only a few types of points exist, points of the same group tend to cluster, and consequently, segregation of points increases. To evaluate the effect of qualitative uniformity, we take the indirect method that assumes the absence of spatial unevenness, compositional unevenness and qualitative uniformity. Absence of qualitative uniformity is represented by an infinite increase in the variety of points, which is mathematically represented as infinite operation $K \to \infty$. Absence of spatial and compositional unevenness reduces overall segregation from $S(\boldsymbol{D}, \mathbf{A})$ to $1/K$, and absence of qualitative uniformity further reduces it infinitely to zero. Consequently, the effect of qualitative uniformity is evaluated by

$$S_Q\left(\boldsymbol{D},\mathbf{A}\right)=\frac{1}{K}. \tag{34}$$

We call this measure *qualitative segregation*.

The relationship between spatial, compositional, and qualitative segregations is represented as

$$S\left(\boldsymbol{D},\mathbf{A}\right)=S_S\left(\boldsymbol{D},\mathbf{A}\right)+S_C\left(\boldsymbol{D},\mathbf{A}\right)+S_Q\left(\boldsymbol{D},\mathbf{A}\right). \tag{35}$$

This equation indicates that overall segregation is decomposed into three components. Spatial, compositional, and qualitative segregations are independent components of segregation since they comprise the entire segregation without overlapping.

The final factor of segregation is attribute similarity, which refers to the similarity in the aspatial properties of points. To measure the effect of attribute similarity, we consider the situation where different groups have completely different properties. Since it is mathematically represented as Equation (13), the effect of attribute similarity is measured by

$$S_A\left(\boldsymbol{D},\mathbf{A}\right)=S\left(\boldsymbol{D},\mathbf{A}\right)-S\left(\boldsymbol{D},\mathbf{A}_{\mathbf{I}}\right). \tag{36}$$

We call this measure *attribute segregation*. Attribute segregation is usually negative or zero since segregation decreases with the similarity between different groups.

We have introduced four primary components of segregation: spatial segregation, compositional segregation, qualitative segregation, and attribute segregation. Spatial segregation represents the spatial aspect of segregation while the others evaluate the aspatial aspect of segregation. Spatial, compositional, and qualitative segregations are independent components as mentioned earlier, while attribute segregation is not independent of the others.

2.5 Statistical analysis of segregation

This subsection proposes a method for evaluating the statistical significance of individual components of segregation. The null hypothesis assumes a stochastic process where the factor of a component to be examined is absent. We derive the probability distribution of the segregation measure by using Monte Carlo simulation and test the statistical significance of the component.

The significance of spatial segregation can be evaluated by either absolute or relative approach. Absolute approach assumes a situation where every type of points are distributed independently according to the uniform distribution. We calculate the probability distribution of S_S(***D***,

A) in this circumstance and evaluate the statistical significance of spatial segregation. Relative approach considers the random permutation of points. We randomize the type of points with keeping their location and calculate $S_S(\boldsymbol{D}, \mathbf{A})$ to obtain its probability distribution under the null hypothesis.

Test of compositional segregation employs $S_C(\boldsymbol{D}, \mathbf{A})$ as the statistic. Since $S_C(\boldsymbol{D}, \mathbf{A})$ is a purely aspatial variable, we perform its statistical test within an aspatial framework. The null hypothesis considers a situation where every point is randomly assigned to one of M types. We calculate the probability distribution of $S_C(\boldsymbol{D}, \mathbf{A})$ in this circumstance and evaluate the statistical significance of compositional segregation.

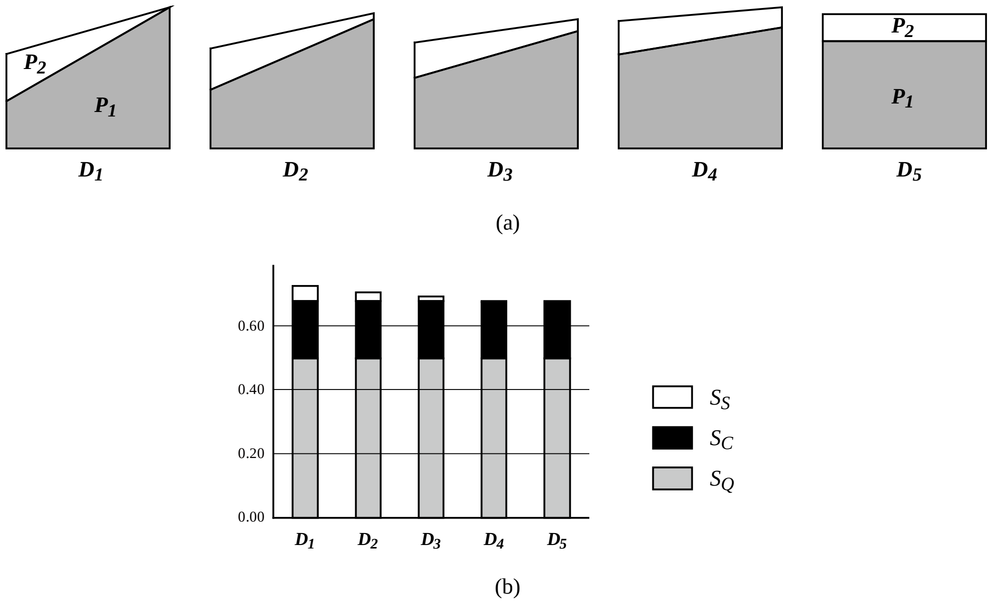
3. Applications

To test the validity of the proposed method, this section applies it to the analysis of three different datasets. We omit the indicator ($\boldsymbol{D}$, $\mathbf{A}$) used in segregation measures for simplicity hereafter.

3.1 Properties of segregation measures

This subsection investigates the properties of segregation measures using a small synthetic dataset. We consider two types of points $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ distributed on a one-dimensional space of length one.

We first focus on the spatial and compositional segregations by assuming matrix $\mathbf{A}$ as $\mathbf{A_I}$. Figure 4a shows the density distributions of $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ that gradually become uniform with keeping their total volume from $\boldsymbol{D_1}$ to $\boldsymbol{D_5}$. Spatial segregation S_S decreases with spatial unevenness as shown in Figure 4b. Measures S_C and S_Q remain unchanged because the proportion of $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ does not change.

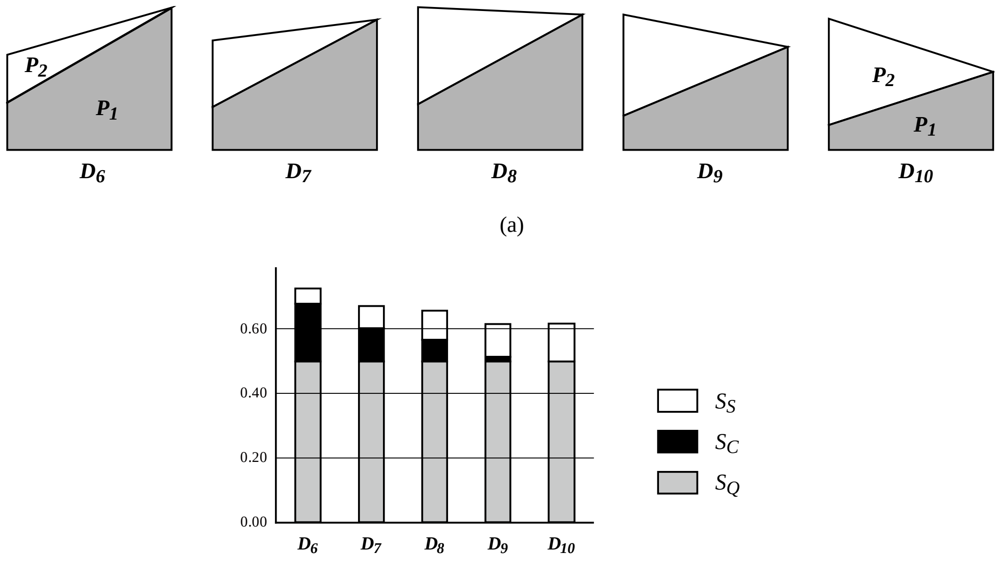


(a)

(b)

Figure 4 The relationship between spatial unevenness and segregation measures. Spatial segregation decreases with spatial unevenness from $\boldsymbol{D_1}$ to $\boldsymbol{D_5}$.

Figure 5a shows a decrease in compositional unevenness where the total volume of $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ gradually approaches with each other. Figure 5b indicates that compositional segregation decreases monotonically with compositional unevenness from $\boldsymbol{D_6}$ to $\boldsymbol{D_{10}}$. Though spatial segregation increases, overall segregation decreases due to the decrease in compositional segregation.



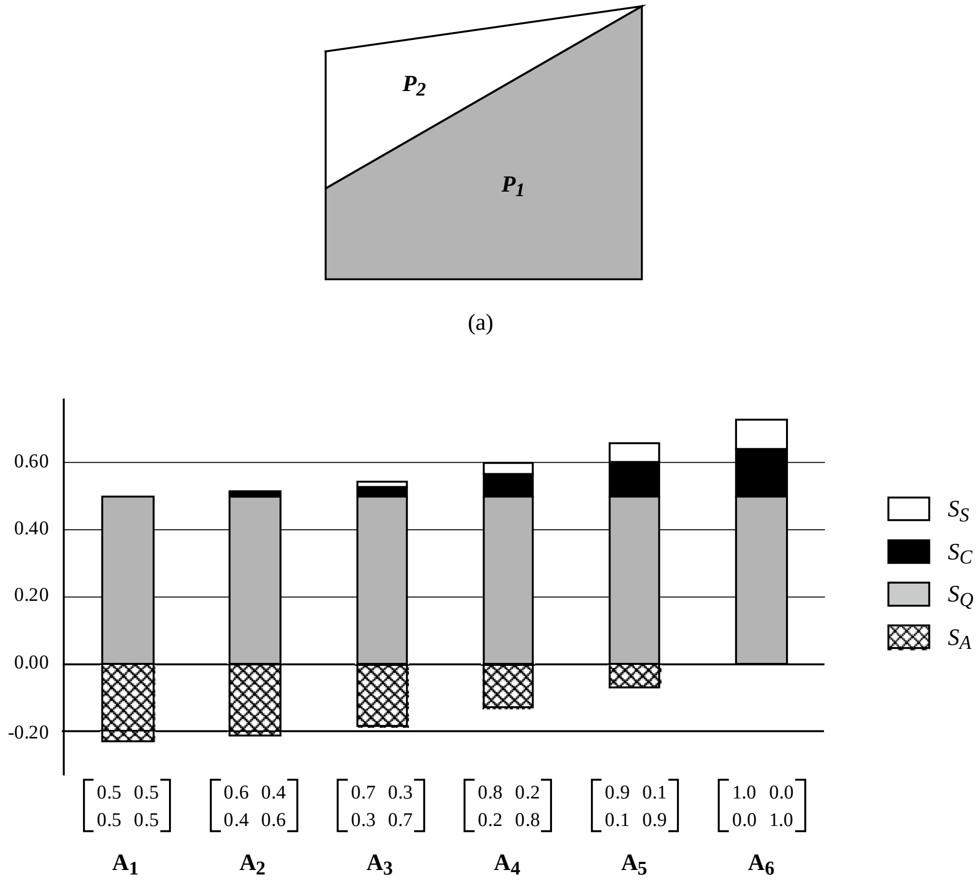
(a)

(b)

Figure 5 The relationship between compositional unevenness and segregation measures. Compositional segregation decreases with compositional unevenness from $\boldsymbol{D_6}$ to $\boldsymbol{D_{10}}$.

We then discuss attribute segregation by using the point distributions shown in Figure 6a. Their attributes are represented by 2×2 matrix **A** shown in Figure 6b, where attribute similarity decreases from $\mathbf{A_1}$ to $\mathbf{A_6}$. We can confirm in Figure 6b that attribute segregation S_A increases with a decrease in attribute similarity. We also notice that S_S+S_C+S_Q-S_A remains unchanged from $\mathbf{A_1}$ to $\mathbf{A_6}$. This can be confirmed by substituting Equation (35) into (36):

$$\begin{aligned} S_S\left(\boldsymbol{D},\mathbf{A}\right)+S_C\left(\boldsymbol{D},\mathbf{A}\right)+S_Q\left(\boldsymbol{D},\mathbf{A}\right)-S_A\left(\boldsymbol{D},\mathbf{A}\right)&=S\left(\boldsymbol{D},\mathbf{A}\right)-S_A\left(\boldsymbol{D},\mathbf{A}\right)\\ &=S\left(\boldsymbol{D},\mathbf{A_I}\right) \end{aligned}. \tag{37}$$



(a)

(b)

Figure 6 The relationship between attribute similarity and segregation measures. (a) Density distributions of points. (b) Attributes of points and segregation measures. Attribute similarity decreases from $\mathbf{A_1}$ to $\mathbf{A_6}$.

3.2 Statistical significance of segregation measures

This subsection discusses the statistical significance of spatial and compositional segregations. We again suppose two types of points $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ distributed on a one-dimensional space of length one.

We evaluate the significance of spatial segregation in point distributions shown in Figure 7. Spatial unevenness increases from left to right while compositional unevenness decreases from up to down. The points $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ follow the density distributions defined by

$$f_1(x) = 1 + b\left(x - \frac{1}{2}\right) \tag{38}$$

and

$$f_2(x) = -b\left(x - \frac{1}{2}\right), \tag{39}$$

respectively ($0 \le b \le 2$). We keep N_1 as 500 while N_2 varies from 100 to 500.

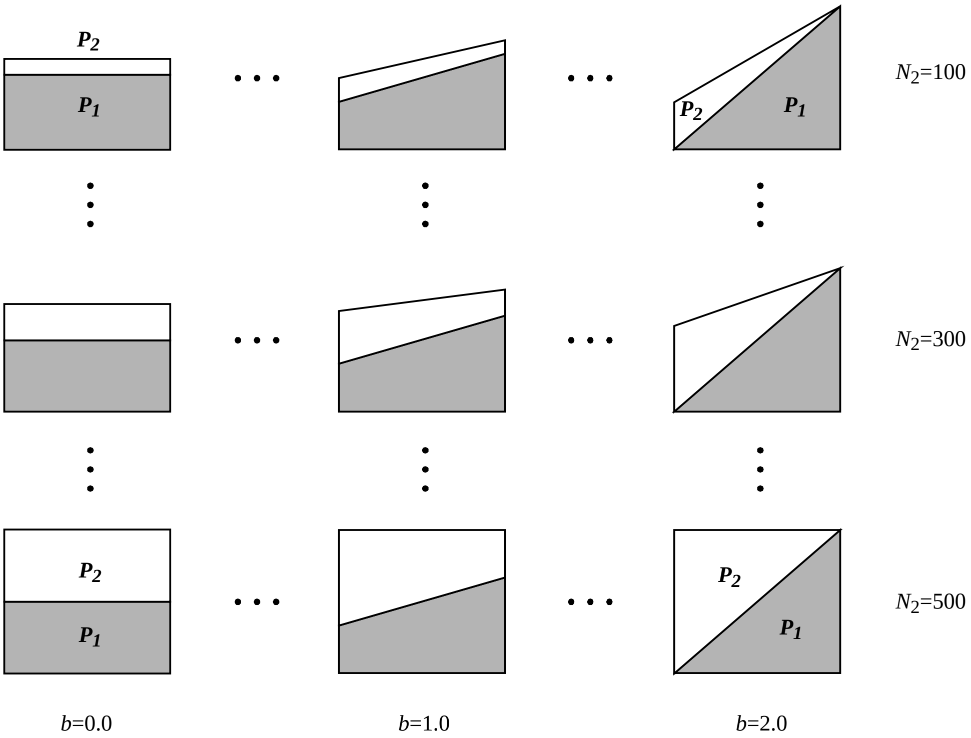


Figure 7 Density distributions of $\boldsymbol{P_1}$ and $\boldsymbol{P_2}$. Spatial unevenness increases from left to right while compositional unevenness decreases from up to down. The number of $\boldsymbol{P_1}$ is 500 while that of $\boldsymbol{P_2}$ varies from 100 to 500.

We employ both absolute and relative approaches in statistical tests to evaluate the effect of different null hypotheses. The significance level of spatial segregation under a null hypothesis is defined by

$$\gamma = \int_S^\infty g(s)\mathrm{d}s\,, \tag{40}$$

where $g(s)$ is the probability density distribution of spatial segregation under the null hypothesis. A small value of γ indicates a high significance of spatial segregation.

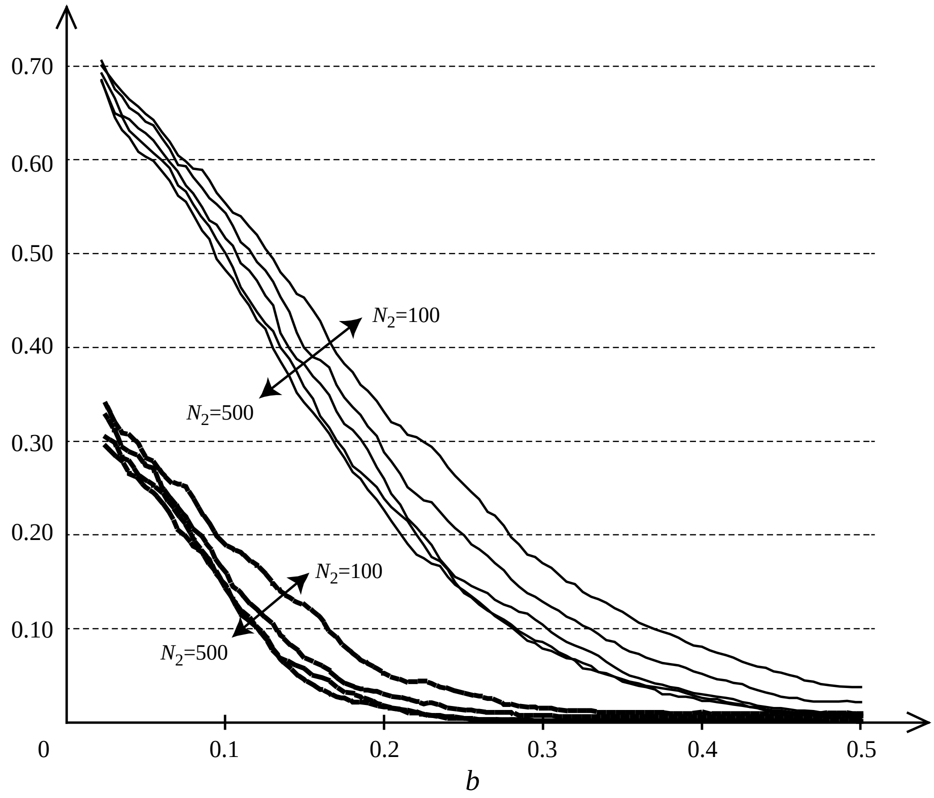


Figure 8 The significance level of spatial segregation under null hypothesis. The number of $\boldsymbol{P_2}$ varies from 100 to 500 as indicated by arrows. Solid and dotted lines indicate the significance level evaluated by absolute and relative approaches, respectively.

Figure 8 shows the significance level of spatial segregation in absolute and relative approaches. As seen in this figure, spatial segregation becomes more significant with an increase in spatial unevenness represented by *b*. The figure also shows that spatial segregation becomes more significant with a decrease in N_2. This implies that segregation is more probable to occur when fewer points are distributed. Comparing absolute and relative approaches, we notice that the former is less likely to evaluate spatial segregation as statistically significant. It is because absolute approach considers a wider variety of point distributions in the null hypothesis. Absolute approach considers the random distribution of points while relative approach only randomizes the type of points with keeping their location. A narrower range of variation in the null hypothesis lowers the requirement for statistical significance.

We then turn to the statistical significance of compositional segregation. Figure 9 shows the distributions of $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$, where N_1 is 500 while N_2 varies from 100 to 500. Compositional unevenness decreases from left to right.

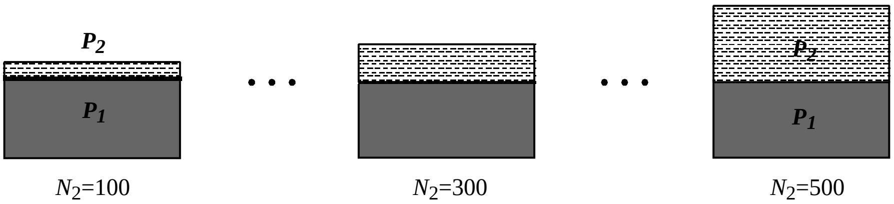


Figure 9 Density distributions of $\boldsymbol{P_1}$ and $\boldsymbol{P_2}$. The number of $\boldsymbol{P_1}$ points is 500 while that of $\boldsymbol{P_2}$ points varies from 100 to 500.

Figure 10 shows the significance level of compositional segregation under the null hypotheses. We can confirm that compositional segregation becomes less significant with a decrease in compositional unevenness. Statistical significance of compositional segregation at 5% level requires that the ratio of $\boldsymbol{P_2}$ to $\boldsymbol{P_1}$ is smaller than 0.75 (=375/500).

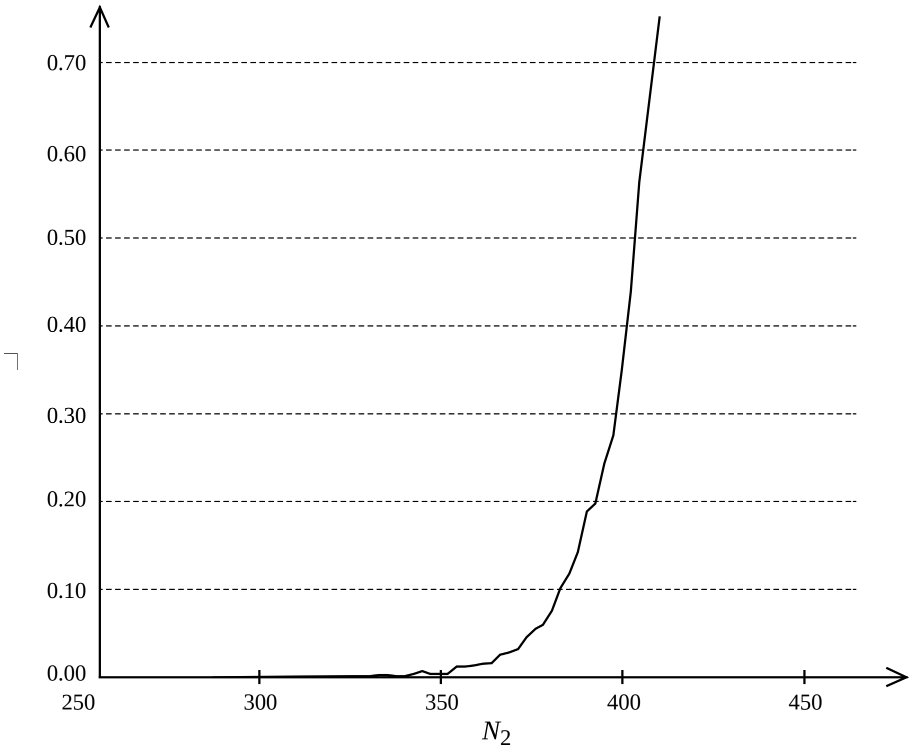


Figure 10 The significance level of compositional segregation under the null hypothesis.

3.3 Real dataset

We finally apply the proposed method to the analysis of a real dataset to test its practical feasibility. We examine the segregation of commercial facilities in Chiba, Japan. Chiba is located 30 kilometers away in a suburb of Tokyo. We converted the list of commercial facilities in the NTT telephone directory into spatial data by geocoding. Figure 11 shows the density distribution of commercial facilities in Chiba.

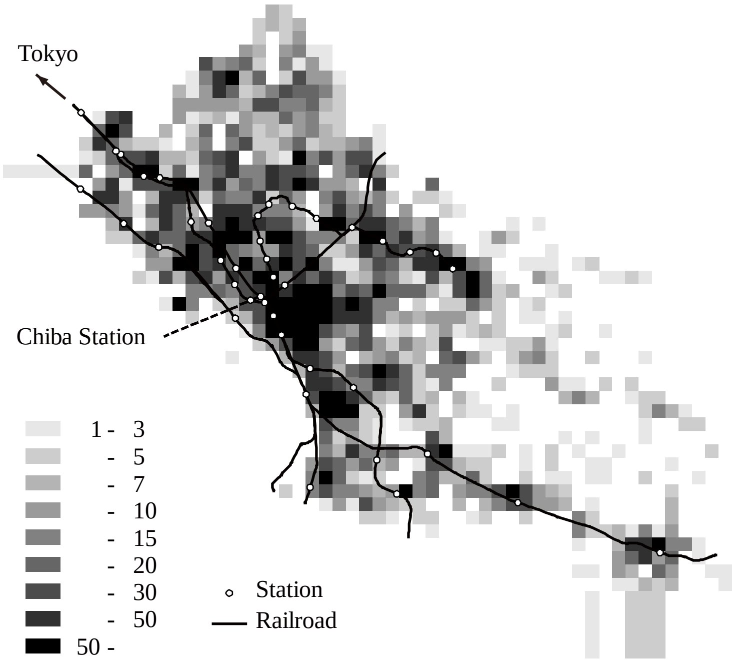
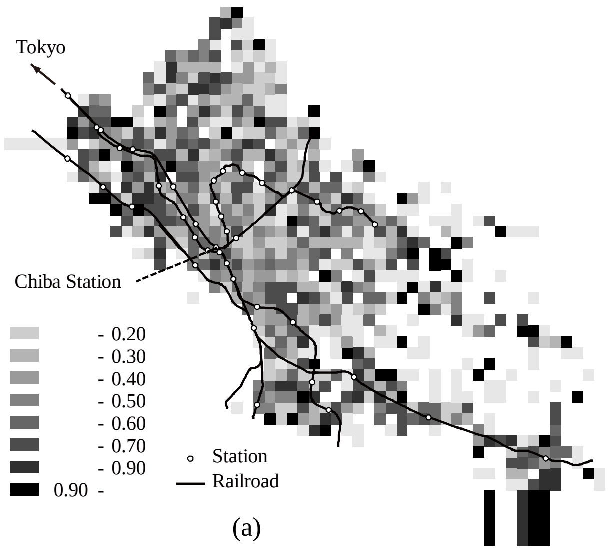


Figure 11 The density distribution of commercial facilities in Chiba [1/km^2].

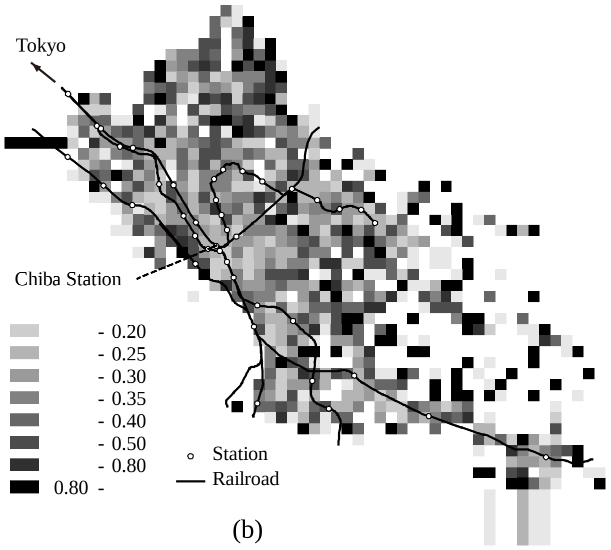
We classified the commercial facilities into different categories by three classification schemes and calculated their segregation measures. Scheme C-1 classified the commercial facilities into three categories:

Group 1: retails (clothing stores, accessory stores, grocery store, ...)
Group 2: services (beauty shops, laundries, hotels, ...)
Group 3: restaurants

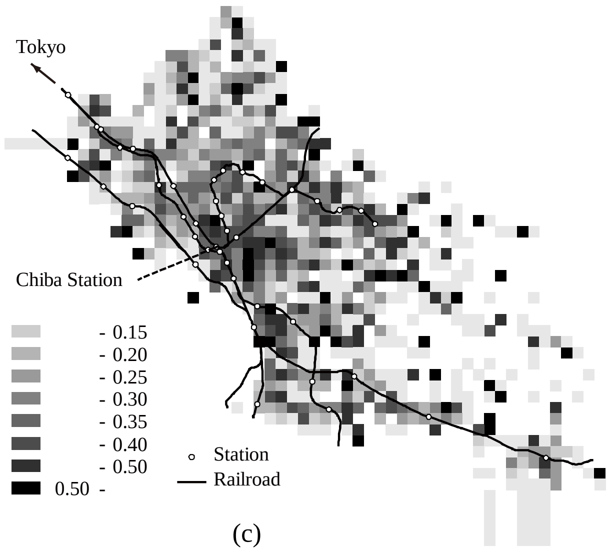
These groups consist of 6236, 4872, and 5197 facilities, respectively. Figure 12 shows the proportion of three types of commercial facilities.


Tokyo
Chiba Station
- 0.20
- 0.30
- 0.40
- 0.50
- 0.60
- 0.70
- 0.90
0.90 -
Station
Railroad

(a)


Tokyo
Chiba Station
- 0.20
- 0.25
- 0.30
- 0.35
- 0.40
- 0.50
- 0.80
0.80 -
Station
Railroad

(b)


Tokyo
Chiba Station
- 0.15
- 0.20
- 0.25
- 0.30
- 0.35
- 0.40
- 0.50
0.50 -
Station
Railroad

(c)

Figure 12 The proportion of three types of commercial facilities in Chiba. (a) Retails, (b) services, (c) restaurants.

Figure 13 shows the distribution of local segregation measure $s(\mathbf{x}; \mathbf{D})$ defined by Equation (4). Comparing Figures 11 and 13, we notice a negative correlation between the density and segregation of commercial facilities. Segregation tends to occur where only a few facilities exist.

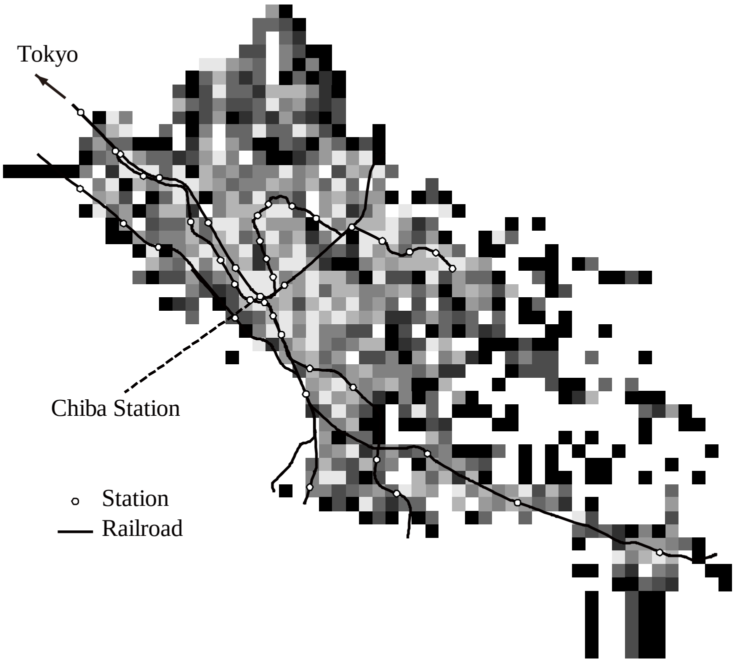


Figure 13 The distribution of local segregation measure $s(\mathbf{x}; \mathbf{D})$.

Table 1 shows the measures of individual components of segregation. The case study evaluates not only the primary components but also the spatial segregation of individual categories.

We first examine the result for classification scheme C-1. Table 1a shows that only spatial segregation is statistically significant at 5% level. Segregation of commercial facilities in Chiba is primarily caused by their spatial unevenness. Table 1b shows that S_{S1} and S_{S2} (retails and services) are significant while S_{S3} (restaurants) is not significant. This implies that restaurants are distributed more uniformly than retails and services. Restaurants are common places for people to visit in suburban areas in Japan such as Chiba.

Scheme C-2 classified the commercial facilities into eight categories each of which contains 1434, 1999, 420, 1370, 1013, 3461, 1411, and 5197 facilities. The result is almost consistent with that for scheme C-1 as shown in Table 1. Only spatial segregation is statistically significant in Table 1a and S_{S8} (restaurants) is not significant among eight categories in Table 1b.

Scheme C-3 classified restaurants further into three subcategories: 1) Japanese restaurants, 2) Western and Chinese restaurants, 3) pubs, bars, and taverns. In this scheme, all the components are not significant as shown in Table 1. This suggests that any type of restaurant is so common everywhere in Chiba.

Table 1 Segregation measures of commercial facilities in Chiba. (a) Segregation measures of primary components, (b) spatial segregation of individual categories. Asterisks indicate the statistical significance at 5% level in both absolute and relative approaches.

Classification scheme	S	S_S	S_C	S_Q
C-1	0.5644	0.2272*	0.0038	0.3333
C-2	0.4331	0.2446*	0.0635	0.1250
C-3	0.6313	0.2645	0.0038	0.3333

(a)

Classification scheme	S_{S1}	S_{S2}	S_{S3}	S_{S4}	S_{S5}	S_{S6}	S_{S7}	S_{S8}
C-1	0.0574*	0.0263*	0.0231	N/A	N/A	N/A	N/A	N/A
C-2	0.1153*	0.0091*	0.0964*	0.0061*	0.0881*	0.1110*	0.0423*	0.0013
C-3	0.0307	0.1108	0.0847	N/A	N/A	N/A	N/A	N/A

(b)

4. Concluding discussion

This paper has proposed a new method for analyzing the segregation between point distributions. Section 2 introduces a general procedure of evaluating individual components of segregation. This procedure helps us find independent components of segregation, and provides a means of assessing their statistical significance. To test the validity of the proposed method, we applied it to the analysis of two synthetic and one real datasets. The result supports the technical soundness of the method, and provides empirical findings.

This paper has several advantages over existing ones. First, this paper proposes a method for evaluating the statistical significance of segregation. Statistical test permits us to judge whether or not a segregation measure can be obtained by chance. Second, our method takes into account explicitly the aspatial properties of points. We can consider spatial and aspatial aspects of point distributions simultaneously in segregation analysis. Third, this paper proposes a general procedure of evaluating individual components of segregation. This procedure helps us consider a wider variety of components and find independent components as shown in Section 2.

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

First, we should extend our method into the spatiotemporal domain. The location and properties of points often change over time, and so does their segregation. One method to treat the temporal dimension is to calculate segregation measures at different times and discuss their change over time. This approach, however, is not sensitive to the change in the spatial arrangement of points. Segregation measures remain the same even if a drastic change occurs in point distributions. A new approach is necessary that considers both the spatial and temporal domains simultaneously in segregation analysis.

Second, we should develop a method for analyzing the segregation of continuous variables. Though our method partially fulfills this purpose, statistical evaluation of segregation is not straightforward due to the difficulty in the choice of null hypothesis for continuous variables. A new statistical method needs to be developed.

Third, segregation analysis should explicitly consider the uncertainty in spatial data. Accuracy of spatial data has been long discussed in geographical information science (Gopal and Goodchild 1989; Hunsaker *et al.* 2002; Shi *et al.* 2002). Though there exist spatial data models that explicitly represent locational uncertainty, they have not yet been fully incorporated into analytical methods. An extension of our method to this direction is an important topic for future research.

References

Brown LA and Chung S-Y (2006) Spatial segregation, segregation indices and the geographical perspective. *Population, Space and Place* 12: 125-143

Duncan OD and Duncan B (1955) A methodological analysis of segregation indexes. *American Sociological Review* 20: 210-217

Gopal U and Goodchild MF 1989. *The Accuracy of Spatial Databases.* CRC Press, London

Hunsaker CT, Goodchild MF, Friedl MA, Case TJ (2001) *Spatial Uncertainty in Ecology: Implications for Remote Sensing and GIS Applications.* Springer, Berlin

James DR, Taeuber KE, 1985. Measures of segregation. *Sociological Methodology* 14: 1-32

Johnston R, Poulsen M, Forrest J (2007) Ethnic and racial segregation in U.S. metropolitan areas, 1980-2000. *Urban Affairs Review* 42: 479-504

Johnston R, Poulsen M, Forrest J (2011) Evaluating changing residential segregation in Auckland, New Zealand, using spatial statistics. *Tijdschrift voor Economische en Sociale Geografie* 102: 1-23

Li H., Reynolds JF (1993) A new contagion index to quantify spatial patterns of landscapes. *Landscape Ecology* 8: 155-162

Logan JR, Zhang W, Alba RD (2002) Immigrant enclaves and ethnic communities in New York and Los Angeles. *American Sociological Review* 67: 299-322

Massey DS, Denton SA (1988) The dimensions of residential segregation. *Social Forces* 67: 281-315

Morgan BS (1983) An alternative approach to the development of a distance-based measure of racial segregation. *American Journal of Sociology* 88: 1237-1249

Morrill RL (1995) Racial segregation and class in a liberal metropolis. *Geographical Analysis* 27: 22-41

Mucientes GR, Queiroz N, Sousa LL, Tarroso P, Sims DW (2009) Sexual segregation of pelagic sharks and the potential threat from fisheries. *Biology Letters* 5: 156-159

O'Neill RV, Krummel JR, Gardner RH, Sugihara G, Jackson B, DeAngelis DL, Milne BT, Turner MG, Zygmunt B, Christensen SW, Dale VH, Graham RL (1988) Indices of landscape pattern. *Landscape Ecology* 1: 153-162

Páez A, Ruiz M, Lopez F, Logan J (2012) Measuring ethnic clustering and exposure with the Q statistic: an exploratory analysis of Irish, Germans, and Yankees in 1880 Newark. *Annals of the Association of American Geographers* 102: 84-102

Poulsen M, Johnston R, Forrest J (2011) Using local statistics and neighbourhood classifications to portray ethnic residential segregation: a London example. *Environment and Planning B* 38: 636-658

Reardon SF, O’Sullivan D (2004) Measures of spatial segregation. *Sociological Methodology* 34: 121-162

Reardon SF, Farrell CR, Matthews SA, O’Sullivan D, Bischoff K, Firebaugh G (2009) Race and space in the 1990s: Changes in the geographic scale of racial residential segregation, 1990-2000. *Social Science Research* 38: 55-70

Reardon SF, Bischoff, K (2011) Income inequality and income segregation. *American Journal of Sociology* 116: 1092-1153

Rey SJ, Folch DC (2011) Impact of spatial effects on income segregation indices. *Computers, Environment and Urban Systems* 35: 431-441

Riitters KH, O'Neill RV, Wickham JD, Jones KB (1996) A note on contagion indices for landscape analysis. *Landscape Ecology* 11: 197-202

Sadahiro Y, Hong S-Y (2013) Decomposition approach to the measurement of spatial segregation. *CSIS Discussion Paper Series No. 119* Center for Spatial Information Science, The University of Tokyo (http://www.csis.u-tokyo.ac.jp/dp/119.pdf)

Shi W, Fisher P, Goodchild MF (2002) *Spatial Data Quality.* CRC Press, London

Stephane D, Conia S, Francois D (2008) Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. *Journal of Vegetation Science* 19: 45-56

Wartenberg D (1985) Multivariate spatial correlation: a method for exploratory geographical analysis. *Geographical Analysis* 17: 263-283

White MJ (1983) The measurement of spatial segregation. *American Journal of Sociology* 88: 1008-1018

White MJ (1986) Segregation and diversity measures of spatial segregation. *American Population Index* 52: 198-221

Wong DWS (1993) Spatial indices of segregation. *Urban Studies* 30: 559-572

Wong DWS (2001) Location-specific cumulative distribution functions (LSCDF): An alternative to spatial correlation analysis. *Geographical Analysis* 33: 76-93