Journal of Geographical Systems
https://doi.org/10.1007/s10109-024-00447-y

ORIGINAL ARTICLE

Point cluster analysis using weighted random labeling

Yukio Sadahiro[1] · **Ikuho Yamada**[2]

Received: 12 January 2024 / Accepted: 7 August 2024

Abstract
This paper proposes a new method of point cluster analysis. There are at least three important points that we need to consider in the evaluation of point clusters. The first is spatial inhomogeneity, i.e., the inhomogeneity of locations where points can be located. The second is aspatial inhomogeneity, which indicates the inhomogeneity of point characteristics. The third is an explicit representation of the geographic scale of analysis. This paper proposes a method that considers these points in a statistical framework. We develop two measures of point clusters: local and global. The former permits us to discuss the spatial variation in point clusters, while the latter indicates the global tendency of point clusters. To test the method's validity, this paper applies it to the analysis of hypothetical and real datasets. The results supported the soundness of the proposed method.

Keywords Point clusters · Spatial inhomogeneity · Aspatial inhomogeneity · Weighted random labeling · Geographical scale of analysis

JEL Classification C65 · R10

1 Introduction

The concept of a cluster of points is one of the most important concepts in point pattern analysis. Point cluster analysis judges whether a point pattern is clustered, dispersed (regular), or random and detects local point clusters. An objective is to reveal the underlying structure of point patterns, i.e., how and why point clusters are generated. Geography considers the clusters of retail stores and restaurants (Scott 1970; Dawson 2012). Epidemiology discusses the clusters of disease cases (Elliot et al. 2000; Lawson 2013). Criminology analyzes the clusters of crime spots (Brantingham and Brantingham

✉ Yukio Sadahiro
sada@csis.u-tokyo.ac.jp

[1] Interfaculty Initiative in Information Studies, The University of Tokyo, 7-3-1, Hongo, Bunkyo-Ku, Tokyo 113-8656, Japan

[2] Center for Spatial Information Science, The University of Tokyo, Tokyo, Japan

Published online: 10 September 2024

1981; Wortley et al. 2008). Point cluster analysis has drawn much attention in various academic fields related to spatial phenomena.

There are at least three important points that we need to consider in the analysis of point clusters. The first is spatial inhomogeneity, which refers to the inhomogeneity of locations where points can be located. Suppose retail stores such as clothing and shoe stores. Zoning regulations restrict the locations of retail stores to commercial zones, and thus, the potential locations are inhomogeneous. Cuzick and Edwards (1990) considers the clusters of disease cases. Their locations are limited only to the residences of individuals, which is also usually inhomogeneous.

The second point is what we call aspatial inhomogeneity, which indicates the inhomogeneity of point characteristics. Pubs and bars prefer small buildings in commercial areas. Home decor and sporting goods shops tend to be located at larger places along highways. Older people are more likely to contract heart disease and diabetes (Brown et al. 2011; Kirkman et al. 2012). The height and diameter of trees affect the selection of hole-nesting birds (Van Balen et al. 1982; Peterson and Gauthier 1985).

We cannot neglect these two inhomogeneities in point cluster analysis since it may lead to erroneous conclusions. Suppose a statistical analysis concludes disease cases as clustered, suggesting an infectious disease. This, however, can happen by chance when the residences of individuals are clustered, even if the disease is not infectious. Birds' nests often form spatial clusters, but it may be caused by the characteristics of trees, such as their height and diameter, rather than their spatial locations.

The third point we need to consider is the geographic scale of analysis. Geographic scale refers to the spatial extent and resolution of analysis (Dabiri and Blaschke 2019; Oshan et al. 2022). Consideration of geographic scale is critical since the analysis results heavily depend on the geographic scale. Ripley's *K*-function, for instance, explicitly considers the analytical scale in point cluster analysis, which is represented by the radius of circles.

Point cluster analysis has been discussed in various academic fields, and numerous methods have been developed for this purpose. Existing methods, however, do not fully cover the above three points, as discussed in the following section, which motivated us to develop a new analytical method. We focus on the case where the locations of points are discrete and limited, such as individuals and buildings mentioned earlier. Our question is whether a certain type of points, such as disease cases and retail stores, are spatially clustered in this setting. We consider both the global and local point clusters, i.e., the global tendency and spatial variation in point clusters. Section 2 discusses the advantages and disadvantages of existing methods. Section 3 describes our method in detail. Section 4 tests the method's validity by applying it to hypothetical and real datasets. Section 5 summarizes the conclusion and discusses the topics of future research.

2 Related works

2.1 Methods based on the complete spatial randomness

The nearest neighbor method is a simple but effective tool for classifying point patterns (Clark and Evans 1954; Clark and Evans 1955; Diggle 1975). It measures the average distance between points and their nearest neighbor points and compares it with the average distance obtained under complete spatial randomness. A drawback is that the nearest neighbor method does not explicitly consider the geographic scale of analysis (Upton and Fingleton 1985; Boots and Getis 1988; Quattrochi and Goodchild 1997; Zhang et al. 2014). Different point patterns can have the same nearest neighbor distance, which implies that the nearest neighbor cannot distinguish many different patterns.

Ripley's *K*-function resolves this problem (Ripley 1976; Ripley 1979). It places circles around points and counts the number of other points inside the circles. The *K*-function then compares it with that obtained under the complete spatial randomness. While the *K*-function evaluates the global tendency of clustering, scan statistic (Kulldorff and Nagarwalla 1995; Kulldorff 1997) focuses on local clusters of points. Placing circles of various sizes at various locations, scan statistic compares the numbers of points inside the circles with that outside the circles. Unfortunately, *K*-function and scan statistic in their original forms do not consider the spatial inhomogeneity of points. The complete spatial randomness assumed as the null hypothesis is often too relaxed in the real world (Cuzick and Edwards 1990).

2.2 Methods considering the spatial inhomogeneity of points

A model-based approach is one option to control the spatial inhomogeneity of points. Spatial statistics have developed stochastic point processes that describe the spatial patterns of points (Cliff and Ord 1981; Diggle and Rowlingson 1994; Baddeley 2007). We can generate point patterns based on a spatial point process and compare them with an observed pattern. A difficulty lies in the choice of the point process. Appropriate choice requires us to have enough knowledge of point processes, which is not always satisfied, especially at an early stage of analysis.

An exploratory approach is another option, and many methods are available to treat spatial inhomogeneity (Kulldorff 2006 provides a comprehensive review). The *k* nearest neighbors (*k*-NN) test developed by Cuzick and Edwards (1990) is one of the most popular methods and is widely used, especially in epidemiology (Gatrell et al. 1996; Haining 2003; Diggle 2013). The test considers the location of disease cases and controls, and the null hypothesis randomizes individuals' labels (case/

control) without changing their locations to evaluate the degree of point clustering. Ripley's cross *K*-function is also applicable to evaluate point clusters under spatial inhomogeneity (Diggle 1983; Cressie 2015). Though it usually assumes complete spatial randomness as the null hypothesis, we can include spatial inhomogeneity by using random labeling (Lynch and Moorcroft 2008; Tao and Thill 2019). Cumulative and maximum χ^2 tests are also often used to control spatial inhomogeneity (Hirotsu 1986; Lagazio et al. 1996; Rogerson 2006; Boulesteix and Strobl 2007). Though these χ^2 tests were not originally developed for spatial analysis, they are applicable to treat spatial inhomogeneity.

A drawback of the above exploratory methods is that they do not consider the aspatial inhomogeneity, i.e., the inhomogeneity of point characteristics. These methods assume that all points have the same probability of being assigned a certain label, which is unrealistic in real-world situations and thus should be relaxed.

2.3 Methods considering the aspatial inhomogeneity of points

Matched case–control design is one solution to control the aspatial inhomogeneity, which is often used in experiment designs in medical and biological sciences (Chetwynd et al. 2001; Jacquez et al. 2005; Pearce 2016). The design considers characteristics of individuals, such as age or gender, and chooses the controls in such a way that the distribution of their characteristics is close to those of cases. Though this method does not aim for spatial analysis, we can extend it into the spatial domain. A disadvantage is that it requires many individuals to be chosen as controls, especially when characteristics vary considerably among individuals.

Weighted random sampling is a procedure of selecting elements from a set according to a weighted probability distribution (Ahrens and Dieter 1985; Devroye 2006; Hübschle-Schneider and Sanders 2022). Unlike matched case–control design, weighted random sampling does not require many points. It is a candidate for controlling aspatial inhomogeneity in point cluster analysis.

2.4 Method considering geographic scale of analysis

There are at least two approaches to representing the geographic scale of analysis. One is to use an absolute spatial measure, such as the distance between locations, as a scale parameter. The *K*-function, for instance, utilizes circles to count the number of points. The radius of circles works as a parameter of representing the analytical scale. Similarly, scan statistic uses circles to detect point clusters, where the circle radius is a scale parameter.

Another approach is to use a relative spatial measure. Cuzick and Edwards (1990) consider the *k*th nearest neighbor points, where *k* represents the analytical scale. Jacquez (1996) also considers the *k*th nearest neighbor point to analyze the space–time

interaction in point distributions. The colocation quotient is defined based on the type of the *k*th nearest neighbor points (Leslie and Kronenfeld 2011).

The two approaches have both advantages and disadvantages. An advantage of absolute measures is that we can easily understand the role and effect of analytical scale since they are represented by real values measured on a concrete space (Rogerson 2006). Relative measures are not easily interpretable since the distance to the *k*th nearest neighbor point varies among locations, which yields difficulty in choosing appropriate k (Chetwynd et al. 2001; Song and Kulldorff 2003; Tango 2007). An advantage of relative measures is that they explicitly consider the spatial inhomogeneity in analysis (Leslie and Kronenfeld 2011). Absolute measures implicitly assume homogeneous space; thus, they are not directly applicable to point cluster analysis under spatial inhomogeneity.

As seen above, existing methods do not fully satisfy all three points of our demand, i.e., simultaneously considering spatial inhomogeneity, aspatial inhomogeneity, and analytical scale. However, they provide us with effective tools for challenging our problems. The randomization test is effective to control the spatial inhomogeneity. Extending weighted random sampling, we can treat the aspatial inhomogeneity of points. Concerning the representation of the geographic scale of analysis, we choose an absolute measure complemented by the randomization test to treat the spatial inhomogeneity. We will describe our method in detail in the following section.

3 Method

Suppose a region Ξ contains N points, denoted as $Z_1, Z_2, \ldots Z_N$. Each point is labeled P or Q, which may represent cases of a disease or trees having birds' nests mentioned in Sect. 1. N_P and N_Q denote the numbers of P and Q points, respectively. Our question is whether P points are clustered in the whole distribution. We assume a single characteristic of points considered closely related to the label, such as the age of individuals and the size of trees. We call this characteristic *attribute* hereafter. The attribute plays a key role in controlling the aspatial inhomogeneity.

3.1 Relationship between the label and the attribute

This subsection discusses the relationship between the label and the attribute. There are two types of attributes: categorical variables and numerical variables. The following treats these cases successively.

We first assume that the attribute is a categorical variable. Suppose that N points represent buildings and that labels P and Q indicate buildings of fast food restaurants and other buildings, respectively. We classify these buildings into

three categories, i.e., those in urban, suburban, and rural areas. The area category is the attribute of buildings. Fast food restaurants tend to be located in urban rather than suburban or rural areas, implying that buildings in urban areas are more likely to be labeled *P*. We calculate the ratio of the buildings of fast food restaurants in each of the three area categories, which indicates the tendency for a building to be labeled *P*. We use the ratio as the *weight* in the null hypothesis of the statistical test described in the next subsection. Buildings with larger weights are more likely to be labeled *P*.

We then consider the case where the attribute is a numerical variable. Again, we consider the labels *P* and *Q*, which indicate the type of building mentioned earlier. We take the floor size of buildings as the attribute. Assume that fast food restaurants avoid very small and very large buildings and prefer middle-sized buildings. The floor size distribution of fast food restaurants has a bell shape. We then fit a Gaussian distribution to the size distribution and estimate the probability distribution. The estimated distribution indicates the relationship between the type of building and floor size, i.e., the tendency for a building to be labeled *P*. Using the estimated distribution, we calculate the weight of each point. Log normal and beta distributions are alternative options if the size distribution is skewed. A logistic distribution is useful when the tendency of being labeled *P* or *Q* monotonically increases or decreases. This applies to the relationship between diabetes and body weight since overweight monotonically increases the risk of diabetes (Colditz et al. 1990; Feldman et al. 2017).

As above, we first clarify the relationship between the label and the attribute. The weight quantitatively measures this relationship and works as a control variable of aspatial inhomogeneity.

3.2 Evaluation of point clustering

This subsection evaluates the clusters of points labeled *P*. We first discuss local analysis and then move to the global analysis. The former aims to capture the spatial variation in point clusters, while the latter aims to understand the global tendency of point clusters.

The local analysis starts by drawing a circle of radius r at a location X, denoted by $C(r, X)$. We count the points labeled *P* and *Q* in $C(r, X)$, denoted by n_P and n_Q, respectively. The ratio of *P* points in $C(r, X)$ is given by

$$\alpha(r, X) = \frac{n_P}{n_P + n_Q}. \tag{1}$$

We compare $\alpha(r, X)$ with the ratio of *P* points in Ξ, as done in scan statistics:

$$\alpha_0 = \frac{N_P}{N}. \tag{2}$$

If P points are clustered in $C(r, X)$, $\alpha(r, X)$ is larger than α_0. We perform a Monte Carlo simulation to evaluate the statistical significance of $\alpha(r, X)$. The null hypothesis assumes that $\alpha(r, X)=\alpha_0$, i.e., the probability that a point is labeled P, is the same inside and outside $C(r, X)$. The alternative hypothesis assumes that $\alpha(r, X) > \alpha_0$, i.e., the probability that a point is labeled P is greater in $C(r, X)$ than in its outside.

We extend the weighted random sampling as follows. We randomly label all the points without changing their locations in each simulation. A single simulation consists of N steps, which is equal to the total number of points. In each step, we choose a label, P or Q, and a point to be labeled following a statistical procedure. The probability that we choose a label is proportional to the number of points to be labeled. We denote the probabilities of choosing P and Q as s_P and s_Q, respectively. They are initially given by

$$s_P = \frac{N_P}{N} \tag{3}$$

and

$$s_Q = \frac{N_Q}{N}, \tag{4}$$

respectively, and updated with a decrease in unlabeled points. The probability of choosing a point to be labeled is proportional to its weight. We denote the weight of Z_i of labels P and Q as w_{Pi} and w_{Qi}, respectively. The probabilities of Z_i being labeled P and Q are given by

$$t_{Pi} = \frac{w_{Pi}}{\sum_j w_{Pj}} \tag{5}$$

and

$$t_{Qi} = \frac{w_{Qi}}{\sum_j w_{Qj}}, \tag{6}$$

respectively. We update these probabilities in the labeling process so that the summations of t_{Pi} and t_{Qi} are both equal to one. We repeat the above step until all the points are labeled. The following is the algorithm of the labeling process. Lines 5.4 and 6.4 update the probabilities of label choice, while lines 8 and 9 update the probabilities of point choice.

Algorithm 1 Algorithm PL (Point Labeling).

1. Calculate s_P and s_Q.
2. Calculate t_{Pi} and t_{Qi} for all the points.
3. Repeat lines 4-9 until $N_P+N_Q=0$.
4. Choose a label based on s_P and s_Q.
5. If P is chosen, do lines 5.1-5.4 and go to line 7
   - 5.1 Choose a point based on t_{Pi}'s, and assume it is Z_k.
   - 5.2 Label Z_k as P.
   - 5.3 $N_P= N_P$-1.
   - 5.4 $$s_P = \frac{N_P}{N_P + N_Q}.$$
6. Else if Q is chosen, do lines 6.1-6.4 and go to line 7
   - 6.1 Choose a point based on t_{Qi}'s, and assume it is Z_k.
   - 6.2 Label Z_k as Q.
   - 6.3 $N_Q= N_Q$-1.
   - 6.4 $$s_Q = \frac{N_Q}{N_P + N_Q}.$$
7. $t_{Pk}=t_{Qk}=0$.
8. $$t_{Pi} = \frac{t_{Pi}}{\sum_j t_{Pj}}.$$
9. $$t_{Qi} = \frac{t_{Qi}}{\sum_j t_{Qj}}.$$
10. End

We call the above process the *weighted random labeling* hereafter. Points are labeled according to a probability distribution. We call ordinary random labeling the *unweighted random labeling*. All the points have the same weight and thus have the same probability of labeling. The weighted random labeling differs from the weighted random sampling in that the former assigns two labels in parallel while the latter assigns only one. Our approach is a generalized form of weighted random sampling and thus can be easily extended to treat more than two labels simultaneously.

				Z_1	Z_2	Z_3	Z_4	Z_5	Z_6
			t_{Pi}	$\frac{3}{12}$	$\frac{1}{12}$	$\frac{4}{12}$	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{1}{12}$
			t_{Qi}	$\frac{9}{60}$	$\frac{11}{60}$	$\frac{8}{60}$	$\frac{11}{60}$	$\frac{10}{60}$	$\frac{11}{60}$
Step 1	P	Z_3	t_{Pi}	$\frac{3}{12}$	$\frac{1}{12}$	$\frac{4}{12}$	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{1}{12}$
			t_{Qi}	$\frac{9}{60}$	$\frac{11}{60}$	$\frac{8}{60}$	$\frac{11}{60}$	$\frac{10}{60}$	$\frac{11}{60}$
Step 2	Q	Z_4	t_{Pi}	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{0}{8}$	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$
			t_{Qi}	$\frac{9}{52}$	$\frac{11}{52}$	$\frac{0}{52}$	$\frac{11}{52}$	$\frac{10}{52}$	$\frac{11}{52}$
Step 3	Q	Z_2	t_{Pi}	$\frac{3}{7}$	$\frac{1}{7}$	$\frac{0}{7}$	$\frac{0}{7}$	$\frac{2}{7}$	$\frac{1}{7}$
			t_{Qi}	$\frac{9}{41}$	$\frac{11}{41}$	$\frac{0}{41}$	$\frac{0}{41}$	$\frac{10}{41}$	$\frac{11}{41}$
Step 4	P	Z_1	t_{Pi}	$\frac{3}{6}$	$\frac{0}{6}$	$\frac{0}{6}$	$\frac{0}{6}$	$\frac{2}{6}$	$\frac{1}{6}$
			t_{Qi}	$\frac{9}{30}$	$\frac{0}{30}$	$\frac{0}{30}$	$\frac{0}{30}$	$\frac{10}{30}$	$\frac{11}{30}$
Step 5	P	Z_5	t_{Pi}	$\frac{0}{3}$	$\frac{0}{3}$	$\frac{0}{3}$	$\frac{0}{3}$	$\frac{2}{3}$	$\frac{1}{3}$
			t_{Qi}	$\frac{0}{21}$	$\frac{0}{21}$	$\frac{0}{21}$	$\frac{0}{21}$	$\frac{10}{21}$	$\frac{11}{21}$
Step 6	Q	Z_6	t_{Pi}	$\frac{0}{1}$	$\frac{0}{1}$	$\frac{0}{1}$	$\frac{0}{1}$	$\frac{0}{1}$	$\frac{1}{1}$
			t_{Qi}	$\frac{0}{11}$	$\frac{0}{11}$	$\frac{0}{11}$	$\frac{0}{11}$	$\frac{0}{11}$	$\frac{11}{11}$

Fig. 1 Weighted random labeling where Z_1–Z_6 denotes six points. Three are labeled P, while the others are labeled Q. Labeling progresses from top to bottom. The red indicates the point labeled at each step, while the blue represents the already labeled points. The second and third columns indicate the label and point chosen at each step (color figure online)

Figure 1 shows an example of the process of weighted random labeling. There are six points, three labeled P and the others labeled Q. Labeling progresses from top to bottom. The red indicates the point labeled at each step, while the blue represents the already labeled points. The second and third columns indicate the label and point chosen at each step.

We calculate the probability that $\alpha(r, X)$ or a larger value is obtained under the null hypothesis and denote it as $\beta(r, X)$. We then define a measure

$$\gamma(r, X) = 1 - 2\beta(r, X). \tag{7}$$

The range of $\gamma(r, X)$ is from -1 to 1. Positive values indicate that P points are clustered in $C(r, X)$, while negative values indicate that points are sparse.

Figure 2 shows point patterns where the weighted random labeling is expected to lead to the correct judgment of point clusters. Numbers indicate the weight of points to be labeled P. Circles indicate the local studied area $C(r, X)$. Red and black points represent P and Q points, respectively. The red points in Figure 2a look spatially clustered, but it is because of large weight values. It is a pseudo cluster corresponding to Type I errors in statistical tests. The red points in Figure 2b are weakly clustered and may not be regarded as a clustered pattern. However, their weight is very small, implying that these points are less likely to be labeled P. We should regard Figure 2b as a clustered pattern corresponding to Type II error. We can similarly discuss dispersed point patterns shown in Figure 2c and 2d. We should judge Figure 2c as not dispersed while Figure 2d as dispersed.

We place a lattice on Ξ and calculate $\gamma(r, X)$ at every lattice point. By visualizing the obtained $\gamma(r, X)$ as a map, we can discuss the spatial variation in the clusters of P points. Like Ripley's K-function, the radius r works as a parameter representing the geographic scale of analysis (Lam and Quattrochi 1992; Ruddell and Wentz 2009).

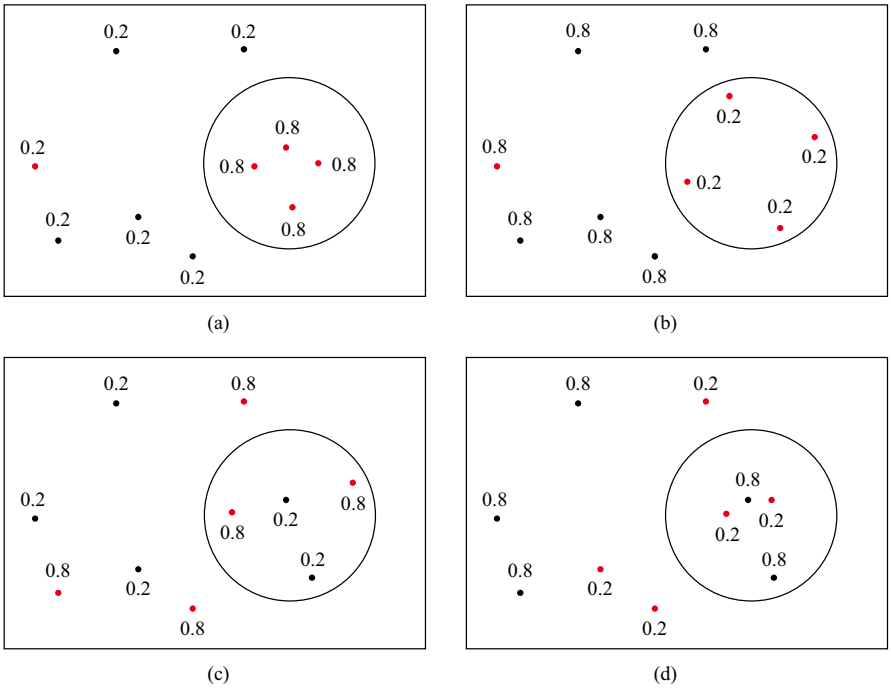


Fig. 2 Examples of point patterns where the weighted random labeling is expected to lead to correct judgment of point clusters. Numbers indicate the weight of points to be labeled P. Circles indicate the local studied area $C(r, X)$. Red and black points are P and Q points, respectively. **a** Red points look spatially clustered, but it is because of their large weights, **b** red points are weakly clustered, but their weights are small, **c** red points are dispersed due to large weights, **d** red points are weakly dispersed, but their weights are small (color figure online)

A large value gives us a macroscale perspective, while a small value permits us to analyze the local spatial pattern in detail.

We then move to the global analysis. Our question is whether P points are clustered across the region Ξ. If P points are clustered, $\gamma(r, X)$ varies across locations, while $\gamma(r, X)$ is uniform when points are dispersed. We thus consider the variance of $\gamma(r, X)$:

$$\lambda(r) = \sum_X \left\{\gamma(r, X) - \overline{\gamma}(r, X)\right\}^2. \tag{8}$$

A large $\lambda(r)$ indicates that P points are clustered, while a small value indicates a dispersed pattern. We randomize the labels using the earlier method to evaluate the statistical significance of $\lambda(r)$. We denote $\Lambda(r)$ as the probability that $\lambda(r)$ or a larger value is obtained under the null hypothesis. We then define a measure

$$\varphi(r) = 1 - 2\Lambda(r). \tag{9}$$

The measure $\varphi(r)$ ranges from -1 to 1. Like $\lambda(r)$, a large $\varphi(r)$ indicates a clustered pattern of points, while a small value indicates a dispersed pattern.

4 Applications

To test the validity of the proposed method, we perform two applications. One uses a hypothetical dataset, while the other uses a real dataset. We wrote two programs in C++ and ran them on an i9-12900U CPU 2.40 GHz, RAM 128 GB computer running Windows 10 Professional.

4.1 Application to hypothetical dataset

This subsection evaluates the proposed method using point distributions, each of which consists of 1000 points in a square of side 1.0. We generated 1000 distributions and evaluated their clustering degree by the nearest neighbor method (Clark and Evans 1954; Diggle 1983). We chose five distributions whose spatial clustering degree was evaluated as the 10, 30, 50, 70, and 90 percent high, denoted by D_{10}, D_{30}, D_{50}, D_{70}, and D_{90}. Concerning r, we tried five values $r = 0.02$, 0.04, 0.06, 0.08, and 0.10, which lead to $5 \times 5 = 25$ settings. The Gaussian distribution of mean 0 and variance 1 generated ten sets of weights for each setting, and we obtained 1000 labeling patterns according to the weights. We chose five significant and five insignificant clustering label patterns at a five percent level based on $\varphi(r)$. To evaluate the statistical significance of these patterns, we performed the Monte Carlo simulation at a five percent level based on the unweighted and weighted random labeling.

Table 1 shows the number of types I (false positive) and II (false negative) errors in 10,000 experiments in each setting. Acceptable levels of type I and II errors are often said to be 5 and 20 percent, respectively (Swinscow and Campbell 2002; Suresh and Chandrashekara 2012). Experiments generally satisfy these requirements except for the type I error of the unweighted random labeling in Table 1a. The result

Table 1 The number of errors in 10,000 experiments in each setting. (a) Type I errors, (b) Type II errors

(a) Type I errors

	Unweighted random labeling					Weighted random labeling				
r	D_{10}	D_{30}	D50	D_{70}	D_{90}	D_{10}	D_{30}	D_{50}	D_{70}	D_{90}
0.02	847	901	509	1096	1425	472	431	505	523	412
0.04	1211	817	722	866	1653	458	535	439	405	437
0.06	1051	1133	617	1167	1106	557	498	391	425	494
0.08	845	1320	1400	1579	1654	487	461	444	468	356
0.10	816	1034	1302	1890	1594	442	448	486	555	423

(b) Type II errors

	Unweighted random labeling					Weighted random labeling				
r	D_{10}	D_{30}	D_{50}	D_{70}	D_{90}	D_{10}	D_{30}	D_{50}	D_{70}	D_{90}
0.02	1521	1120	1279	1412	1857	1376	1000	1245	1056	1177
0.04	1400	1722	937	839	1817	1143	1080	1173	1055	1044
0.06	895	1048	1024	1532	1825	969	971	1133	1038	1146
0.08	989	931	1134	1341	2006	961	954	1169	949	1082
0.10	1227	775	1191	1809	1713	1224	985	1253	1154	935

clearly shows that the weighted random labeling reduces statistical errors. Type I errors were reduced in all 25 settings in Table 1a. Type II errors were reduced in 17 settings in Table 1b, statistically significant by the binomial test, where the p-value was 0.022.

4.2 Application to a real dataset

This subsection analyzes the spatial pattern of pubs in Shinjuku-ku, Tokyo. Our aim was to evaluate whether pubs are clustered among all the restaurants. We used telephone directory data provided by the NTT TownPage cooperation and building footprint data provided by the Zenrin cooperation. Figure 3 shows the restaurant distribution in Shinjuku-ku. This area contains 4187 restaurants, and 1382 of them are pubs.

Pubs prefer small buildings. We thus considered the floor size as the weight for evaluating pub clusters. Figure 4 shows the histogram of the floor size of pubs. We fitted the lognormal distribution to these data by the maximum likelihood method and obtained the distribution represented by the red line in the figure, where $(\mu, \sigma^2) = (2.474, 0.462)$. We defined the probability that ith building is assigned to other types of restaurants by

$$t_{Qi} = 1 - \frac{w_{Pi}}{\sum_j w_{Pj}}.$$

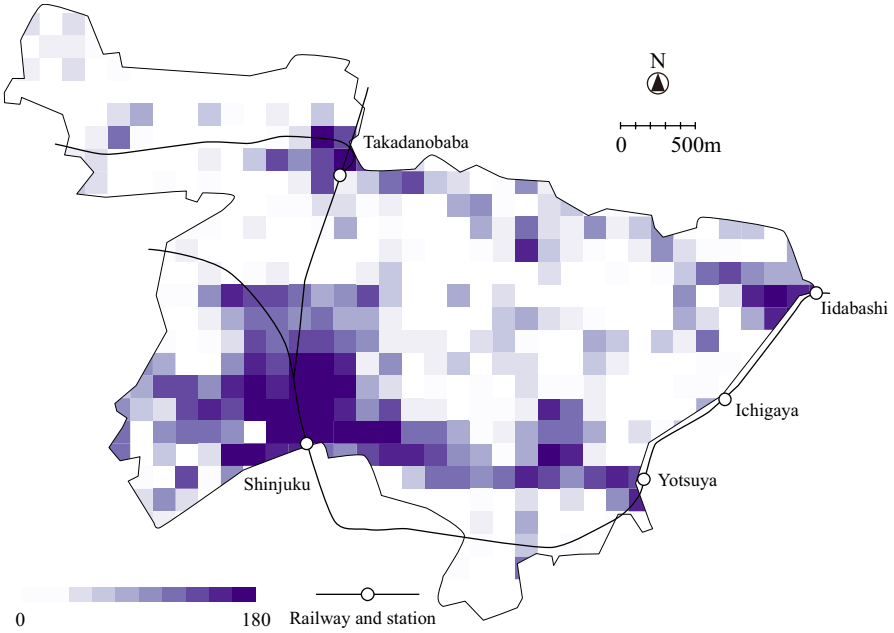


Fig. 3 The distribution of restaurants in Shinjuku-ku

We first performed the local analysis. We performed the Monte Carlo simulation 10,000 times to obtain $\gamma(r, X)$ at 6173 lattice points. The calculations were completed within 100 min in all the cases. The following shows the results when $r = 500$, 250, and 125 m.

Figure 5 shows the distribution of $\gamma(r, X)$ where $r = 500$ m. The two figures show the unweighted and weighted random labeling results, respectively. Red colors indicate pub clusters, while blue colors are sparse areas. Both figures show that pubs are clustered around the Shinjuku and Yotsuya stations. In contrast, pubs are clustered around the Takadanobaba station only in Fig. 5a and the Iidabashi station only in Fig. 5b. Figure 5a does not consider the floor size of buildings, while Fig. 5b uses the floor size distribution as the weight. Pubs tend to be located in small buildings, as shown in Fig. 4. Figure 5 suggests that small buildings are clustered around the Takadanobaba station, while few are clustered around the Iidabashi station. The red color around Takadanobaba station in Fig. 5a appears because of the clusters of small buildings rather than because of the pubs. They are pseudo clusters.

Figure 6 shows the distribution of $\gamma(r, X)$ where $r = 250$ m. The geographic scale of analysis is smaller; thus, the figures provide detailed patterns of pub clusters. Red colors exist around the Takadanobaba station in Fig. 6a and the Iidabashi station in Fig. 6b. This is consistent with Fig. 5. One difference lies in the area around the

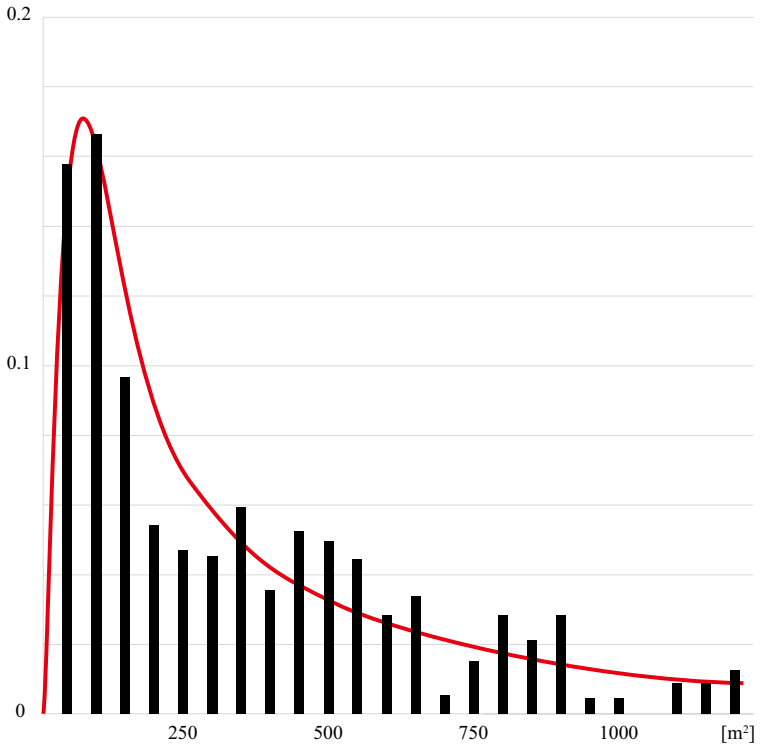
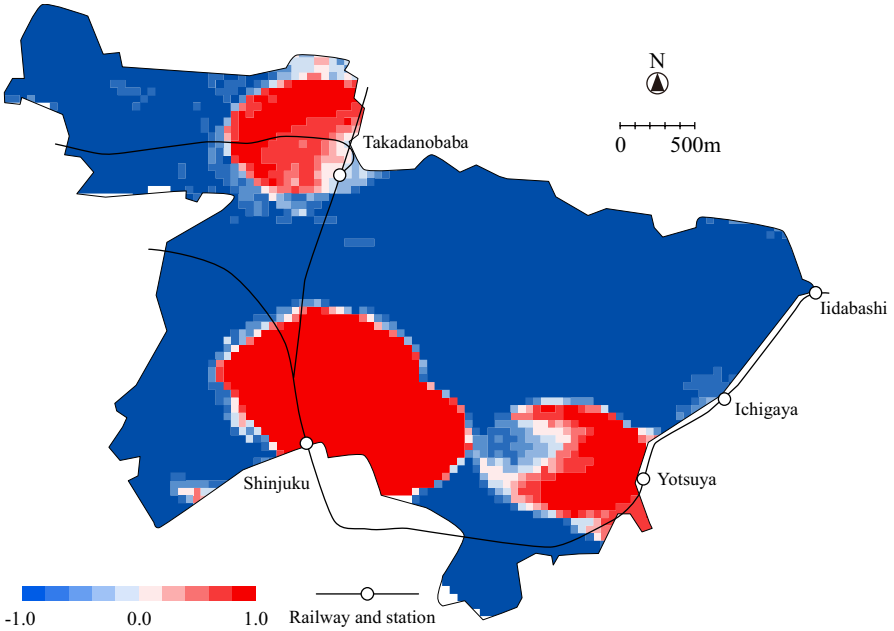


Fig. 4 Histogram of floor sizes of pubs and the lognormal distribution fitted to the floor size distribution

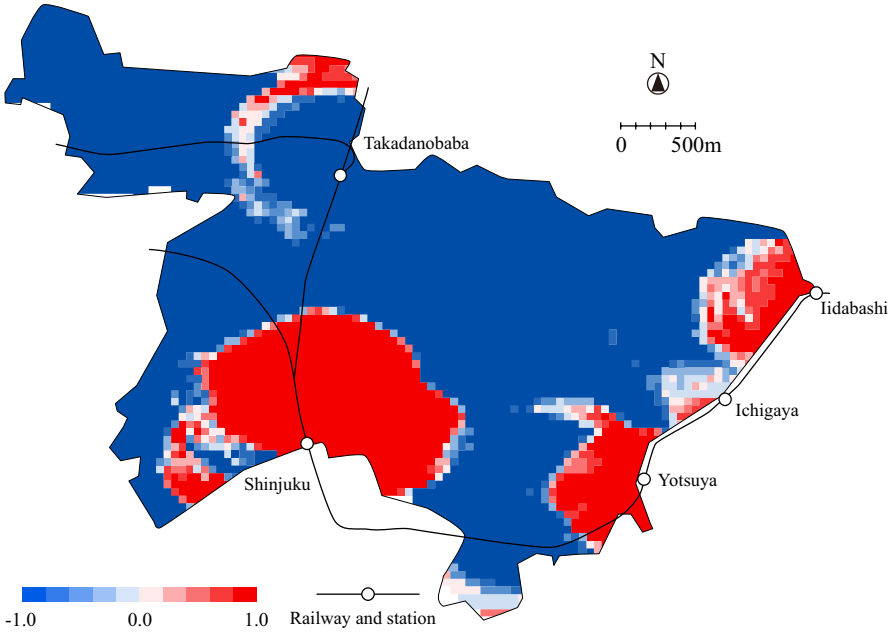
Takadanobaba station, as shown in Fig. 6b. The figure indicates that pubs are clustered west of the Takadanobaba station, which is unclear in Fig. 5b. Another difference is the blue colors around the Shinjuku station in Fig. 6b. The pubs are not clustered close to the Shinjuku station.

Figure 7 shows the distribution of $\gamma(r, X)$ where $r = 125$ m. Figure 7b shows a more detailed spatial pattern of pub clusters. Pub clusters around the Shinjuku station exhibit more complicated shapes. Pub clusters appear at the center of Shinjuku-ku and could not be detected in Figs. 5 and 4. Two clusters in the west of the Takadanobaba station are divided into three clusters, as shown in Fig. 7b.

Table 2 shows $\varphi(r)$, which represents the clustering tendency at the global scale in Shinjuku-ku. Large positive values indicate that the pubs are highly clustered at these scales. The values are different between the unweighted and weighted random labelings. This finding supports the importance of considering floor size when evaluating pub clusters.

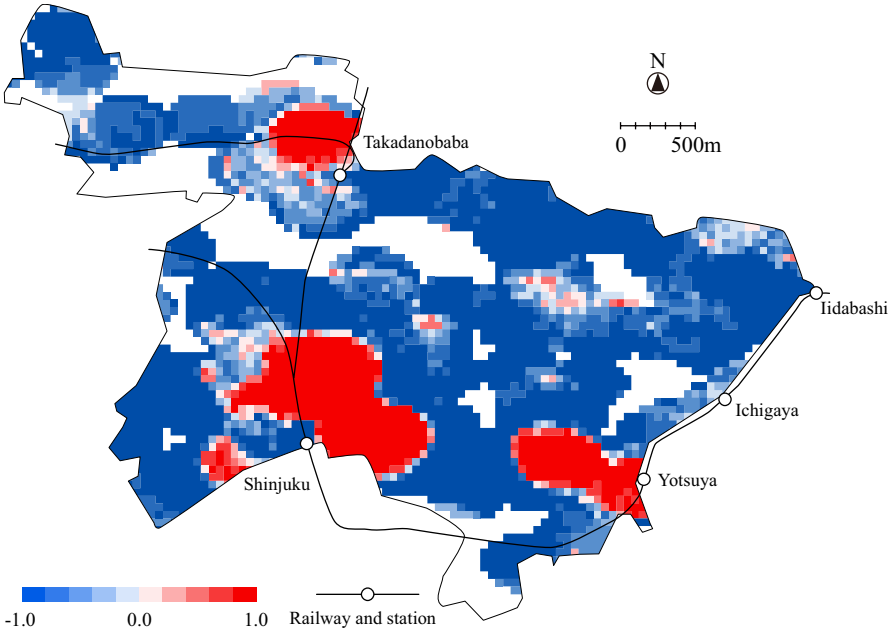


(a)

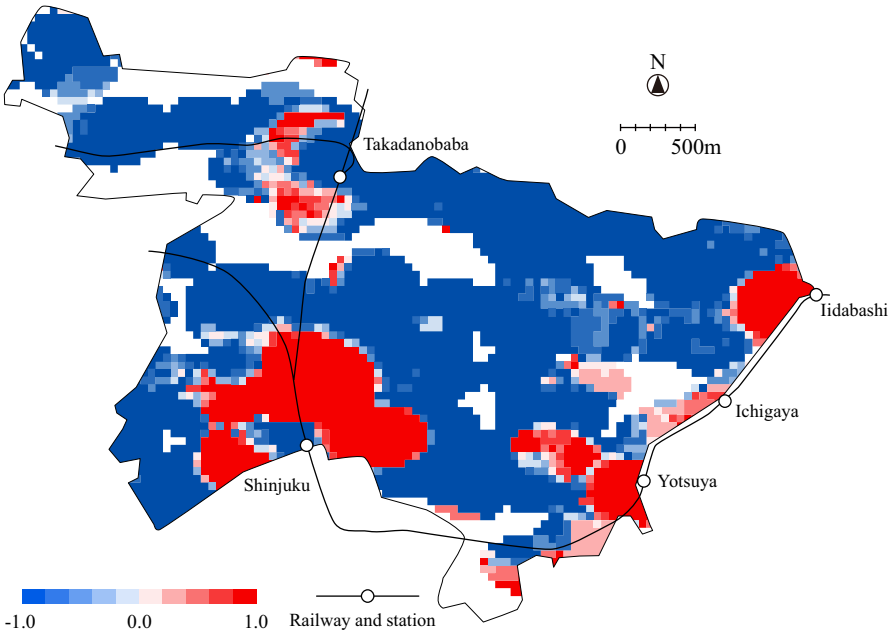


(b)

Fig. 5 The distribution of $\gamma(r, X)$ where $r = 500$ m. **a** Unweighted random labeling, **b** weighted random labeling

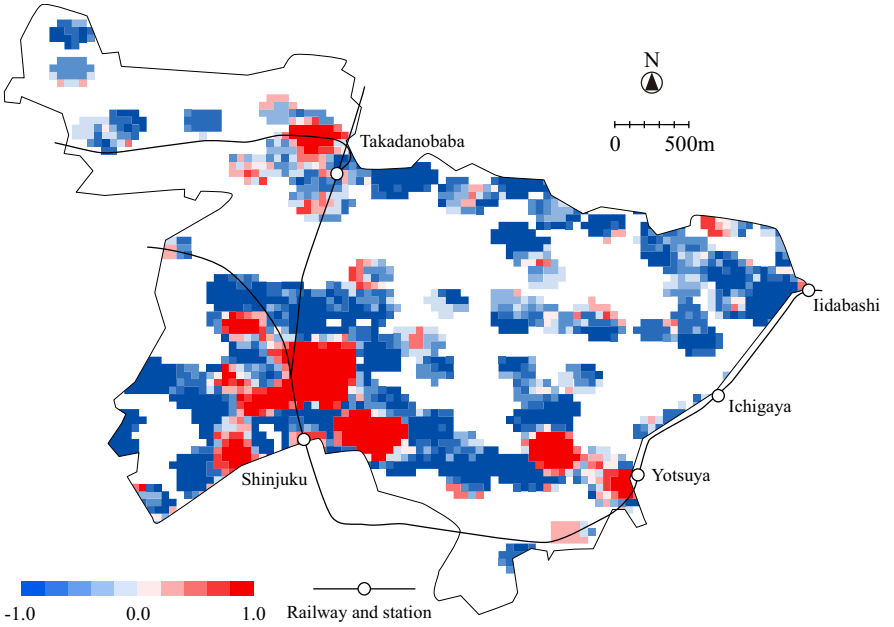


(a)

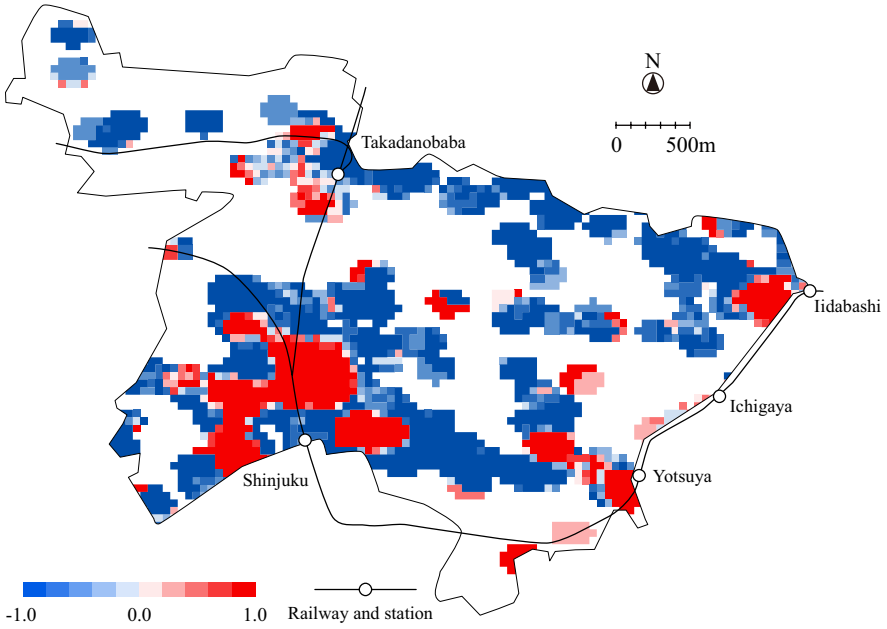


(b)

Fig. 6 The distribution of $\gamma(r, X)$ where $r = 250$ m. **a** Unweighted random labeling, **b** weighted random labeling



(a)



(b)

Fig. 7 The distribution of $\gamma(r, X)$ where $r = 125$ m. **a** Unweighted random labeling, **b** weighted random labeling

Table 2 The measure $\varphi(r)$ where $r = 500$, 250, and 125 m

r	Unweighted	Weighted
500	0.84	0.91
250	0.88	0.90
125	0.86	0.87

5 Conclusion

This paper proposed a new method for evaluating point clusters. The measure $\gamma(r, X)$ is useful for discussing the spatial variation in point clusters, while $\varphi(r)$ reflects the global tendency of point clusters. To test the validity of the method, we first applied it to a hypothetical dataset. The result statistically supports the advantage of the weighted random labeling. We then applied the method to the analysis of the spatial pattern of pubs in Shinjuku-ku, Tokyo. Empirical findings are useful and support the effectiveness of the proposed method.

An advantage of our method is that it considers all the three important points discussed in Sect. 1, i.e., spatial inhomogeneity, aspatial inhomogeneity, and analytical scale. The method, however, is not free of limitations. We discuss them and extensions for future research.

Firstly, this paper considers a numerical variable as the point attribute. SubSect. 3.1, on the other hand, also mentions categorical variables as the attribute. Categorical attributes of buildings include their structure, availability of parking lots, surrounding land use, and so forth. Weight calculation is easier than numerical variables. This, however, does not assure that the proposed method works successfully for categorical variables. Further applications are required to test the effectiveness of our method.

Secondly, this paper adopts an absolute measure to represent the geographical scale of analysis. As discussed in SubSect. 2.4, however, relative measures have their advantages. One method of relative approach is to replace the number of points in circle $C(r, X)$ with that within the kth nearest neighboring points. We do not have to modify the proposed method in this approach substantially. It is worth trying to use relative measures with resolving the difficult problem of choosing an appropriate k.

Thirdly, we should extend the proposed method to the spatiotemporal domain. Spatiotemporal point clusters have long been discussed in the literature (Diggle et al. 1995; Kulldorff et al. 1998; Alvarez et al. 2016). It may seem easily achievable by replacing the circle $C(r, X)$ with a cylinder. This approach, however, has two problems. Firstly, the scale of analysis depends on two variables, i.e., the radius and height of the cylinders. We will obtain various results, and the comparisons and interpretations of these results may be difficult. Secondly, the computing time will increase. An efficient algorithm is again necessary.

Fourthly, this paper considers the clusters of two labels represented as P and Q. Clusters, however, can occur where more than two labels exist. The colocation quotient developed by Leslie and Kronenfeld (2011) considers the colocation of more

than three types of points. We can improve our approach to treat more than two labels, as mentioned in SubSect. 3.2. An extension in this direction seems fruitful and interesting.

Fifthly, this paper assumes categorical labels. Consideration of numerical labels is a useful extension. A question is whether points of similar numerical values are clustered, which is equivalent to the question of spatial autocorrelation analysis. Existing spatial autocorrelation measures use unweighted randomization in statistical tests. Extending our method, we may be able to introduce weighted randomization in spatial autocorrelation analysis.

Acknowledgements The author thanks the reviewers for their constructive comments and suggestions. This research was supported by JSPS KAKENHI Grant Numbers 19H02375, 22H00245, and 23H01634.

Funding Open Access funding provided by The University of Tokyo.

Code availability The program used in the empirical study is available in Figshare at https://figshare.com/articles/dataset/Evaluation_of_point_clusters_within_an_inhomogeneous_population_using_weighted_random_labeling/24947337

References

Ahrens JH, Dieter U (1985) Sequential random sampling. ACM Trans Math Softw (TOMS) 11:157–169

Alvarez J, Goede D, Morrison R, Perez A (2016) Spatial and temporal epidemiology of porcine epidemic diarrhea (ped) in the midwest and southeast regions of the United States. Prev Vet Med 123:155–160

Baddeley A (2007) Spatial point processes and their applications. In: Baddeley A, Bárány I, Schneider R (eds) Stochastic geometry lecture notes in mathematic. s. Springer, Berlin

Van Balen J, Booy C, Van Franeker J, Osieck E (1982) Studies on hole-nesting birds in natural nest sites. Ardea 55:1–24

Boots BN, Getis A (1988) Point pattern analysis. Sage Publications, Newbury Park, Calif, CA

Boulesteix A-L, Strobl C (2007) Maximally selected chi-squared statistics and non-monotonic associations: an exact approach based on two cutpoints. Comput Stat Data Anal 51:6295–6306

Brantingham PJ, Brantingham PL (1981) Environmental criminology. Sage Publications, Beverly Hills, CA

Brown JM, Stewart JC, Stump TE, Callahan CM (2011) Risk of coronary heart disease events over 15 years among older adults with depressive symptoms. Am J Geriatr Psychiatry 19:721–729

Chetwynd AG, Diggle PJ, Marshall A, Parslow R (2001) Investigation of spatial clustering from individually matched case-control studies. Biostatistics 2:277–293

Clark PJ, Evans FC (1954) Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology 35:445–453

Clark PJ, Evans FC (1955) On some aspects of spatial pattern in biological populations. Science 121:397–398

Cliff AD, Ord JK (1981) Spatial processes: models & applications. Pion, London

Colditz GA, Willett WC, Stampfer MJ, Manson JE, Hennekens CH, Arky RA, Speizer FE (1990) Weight as a risk factor for clinical diabetes in women. Am J Epidemiol 132:501–513
Cressie N (2015) Statistics for spatial data. Wiley, New York
Cuzick J, Edwards R (1990) Spatial clustering for inhomogeneous populations. J Royal Stat Soc Ser B Method 52:73–104
Dabiri Z, Blaschke T (2019) Scale matters: a survey of the concepts of scale used in spatial disciplines. Eur J Remote Sens 52:419–434
Dawson JA (2012) Retail geography. Routledge, New York
Devroye L (2006) Nonuniform random variate generation. Handb Oper Res Manag Sci 13:83–121
Diggle PJ (1975) Robust density estimation using distance methods. Biometrika 62:39–48
Diggle PJ (1983) Statistical analysis of spatial point patterns. Academic press, London
Diggle PJ (2013) Statistical analysis of spatial and spatio-temporal point patterns. Chapman and Hall/CRC, Boca Raton, FL
Diggle PJ, Chetwynd AG, Häggkvist R, Morris SE (1995) Second-order analysis of space-time clustering. Stat Methods Med Res 4:124–136
Diggle PJ, Rowlingson BS (1994) A conditional approach to point process modelling of elevated risk. J R Stat Soc Ser A Stat Soc 157:433–440
Elliot P, Wakefield JC, Best NG, Briggs DJ (2000) Spatial epidemiology: methods and applications. Oxford University Press, Oxford, UK
Feldman AL, Griffin SJ, Ahern AL, Long GH, Weinehall L, Fhärm E, Norberg M, Wennberg P (2017) Impact of weight maintenance and loss on diabetes risk and burden: a population-based study in 33,184 participants. BMC Public Health 17:1–10
Gatrell AC, Bailey TC, Diggle PJ, Rowlingson BS (1996) Spatial point pattern analysis and its application in geographical epidemiology. Trans Inst Br Geogr 21:256–274
Haining RP (2003) Spatial data analysis: theory and practice. Cambridge University Press, Cambridge, UK
Hirotsu C (1986) Cumulative chi-squared statistic as a tool for testing goodness of fit. Biometrika 73:165–173
Hübschle-Schneider L, Sanders P (2022) Parallel weighted random sampling. ACM Trans Math Softw (TOMS) 48:1–40
Jacquez GM (1996) Ak nearest neighbour test for space-time interaction. Stat Med 15:1935–1949
Jacquez GM, Kaufmann A, Meliker J, Goovaerts P, Avruskin G, Nriagu J (2005) Global, local and focused geographic clustering for case-control data with residential histories. Environ Health 4:1–19
Kirkman MS, Briscoe VJ, Clark N, Florez H, Haas LB, Halter JB, Huang ES, Korytkowski MT, Munshi MN, Odegard PS (2012) Diabetes in older adults. Diabetes Care 35:2650
Kulldorff M (1997) A spatial scan statistic. Commun Stat Theory Methods 26:1481–1496
Kulldorff M (2006) Tests of spatial randomness adjusted for an inhomogeneity: a general framework. J Am Stat Assoc 101:1289–1305
Kulldorff M, Athas WF, Feurer EJ, Miller BA, Key CR (1998) Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, new Mexico. Am J Public Health 88:1377–1380
Kulldorff M, Nagarwalla N (1995) Spatial disease clusters: detection and inference. Stat Med 14:799–810
Lagazio C, Marchi M, Biggeri A (1996) The association between risk of disease and point sources of pollution: a test for case-control data. Stat Appl 8:343–356
Lam NS-N, Quattrochi DA (1992) On the issues of scale, resolution, and fractal analysis in the mapping sciences*. Prof Geogr 44:88–98
Lawson AB (2013) Statistical methods in spatial epidemiology. Wiley, Chichester
Leslie TF, Kronenfeld BJ (2011) The colocation quotient: a new measure of spatial association between categorical subsets of points. Geogr Anal 43:306–326
Lynch HJ, Moorcroft PR (2008) A spatiotemporal ripley's k-function to analyze interactions between spruce budworm and fire in British Columbia, Canada. Can J for Res 38:3112–3119
Oshan TM, Wolf LJ, Sachdeva M, Bardin S, Fotheringham AS (2022) A scoping review on the multiplicity of scale in spatial analysis. J Geogr Syst 24:293–324
Pearce N (2016) Analysis of matched case-control studies. BMJ, London, p 352
Peterson B, Gauthier G (1985) Nest site use by cavity-nesting birds of the Cariboo Parkland, British Columbia. Wilson Bull 97:319–331
Quattrochi DA, Goodchild MF (1997) Scale in remote sensing and gis. CRC Press, Boca Raton, FL
Ripley BD (1976) The second-order analysis of stationary point processes. J Appl Probab 13:255–266

Ripley BD (1979) Tests ofrandomness' for spatial point patterns. J Royal Stat Soc Ser B Methodol 41:368–374
Rogerson PA (2006) Statistical methods for the detection of spatial clustering in case–control data. Stat Med 25:811–823
Ruddell D, Wentz EA (2009) Multi-tasking: scale in geography. Geogr Compass 3:681–697
Scott P (1970) Geography and retailing. Transaction Publishers, Chicago
Song C, Kulldorff M (2003) Power evaluation of disease clustering tests. Int J Health Geogr 2:1–8
Suresh K, Chandrashekara S (2012) Sample size estimation and power analysis for clinical research studies. J Hum Reprod Sci 5:7–13
Swinscow TDV, Campbell MJ (2002) Statistics at square one, 10th edn. BMJ, London
Tango T (2007) A class of multiplicity adjusted tests for spatial clustering based on case–control point data. Biometrics 63:119–127
Tao R, Thill J-C (2019) Flow cross k-function: a bivariate flow analytical method. Int J Geogr Inf Sci 33:2055–2071
Upton G, Fingleton B (1985) Spatial data analysis by example. Volume 1: point pattern and quantitative data. Wiley, Chichester, UK
Wortley R, Mazerolle LG, Rombouts S (2008) Environmental criminology and crime analysis. Routledge, Boca Raton, FL
Zhang J, Atkinson P, Goodchild MF (2014) Scale in spatial information and analysis. CRC Press, Boca Raton, FL