Point pattern analysis on spatially aggregated data

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

The University of Tokyo

Ikuho Yamada
The University of Tokyo

Research Article

Keywords: point pattern analysis, spatially aggregated data, spatial point pattern, spatial label pattern

Posted Date: October 1st, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7586973/v1

Additional Declarations: No competing interests reported.

Point pattern analysis on spatially aggregated data

Abstract

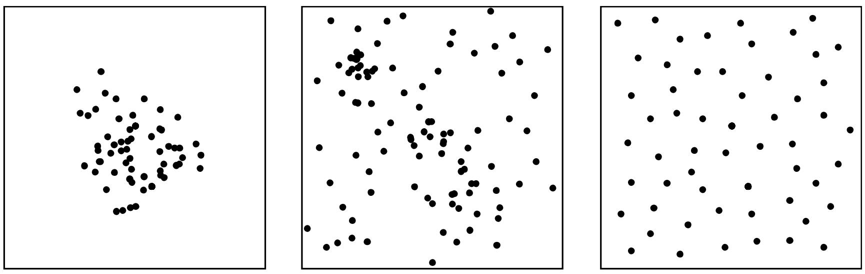
Point pattern analysis is a basic but essential analysis in geography and other fields. Point pattern analysis evaluates the spatial pattern of points using their locational data. Locational data, however, are not always available, especially when points represent individuals. Spatial units aggregate the information of individuals to keep their confidentiality, and existing methods of point pattern analysis cannot fully evaluate the spatial point pattern on spatially aggregated data. To fill the research gap, we propose a new method of point pattern analysis on spatially aggregated data. We consider the spatial patterns of points and labels, the latter of which is often called "marked" points in spatial statistics. We propose two statistics to evaluate these patterns, defined based on spatially aggregated data. We test the validity of the statistics through computational experiments. The results indicate the effectiveness of the statistics in a wide variety of situations.

Keywords: point pattern analysis, spatially aggregated data, spatial point pattern, spatial label pattern

1. Introduction

Point pattern analysis is a basic but essential analysis in geography and other academic fields that treat spatial objects represented as points. Geography analyzes the spatial pattern of retail stores (Rogers (1974); Rabino and Mastrangelo (2002); Cui and Han (2015)), restaurants (Ishizaki (1995); Prayag et al. (2012)), and hotels (Wall et al. (1985); Luo and Yang (2013)). Ecology is interested in the spatial pattern of birds' nests (Peterson and Gauthier (1985); Bisson et al. (2002)), forest trees (Warren (1972); Penttinen et al. (1992)), and so forth. Epidemiology discusses the spatial pattern of disease cases (Lawson (2013); Souris (2019)). Criminology analyzes the spatial pattern of crimes to detect the offenders' home location and to prevent crimes (Brantingham and Brantingham (1981); Wortley et al. (2008)).

An important topic in point pattern analysis is what we call *spatial point pattern* (Diggle (1983); Boots and Getis (1988)). Figures 1a-1c show examples of three typical spatial point patterns. Points are clustered in Figure 1a, while points are randomly distributed in Figure 1b. Points are dispersed in Figure 1c, which is often called a regular or repulsive pattern. Point pattern analysis evaluates point patterns, i.e., whether an observed pattern is clustered, dispersed, or random.



(a) (b) (c)

Figure 1 Spatial point patterns. (a) Clustered, (b) random, (c) dispersed.

Point pattern analysis also treats the spatial pattern of points with a binary label, often called "marked" points in spatial statistics. We call it *spatial label pattern* hereafter. Examples include restaurants with/without parking lots, houses with/without gardens, and trees with/without birds' nests. Figure 2 shows examples of spatial label patterns. Given a spatial pattern of points, we consider whether labeled points are relatively clustered, dispersed, or random (Cuzick and Edwards (1990); Diggle and Chetwynd (1991)).

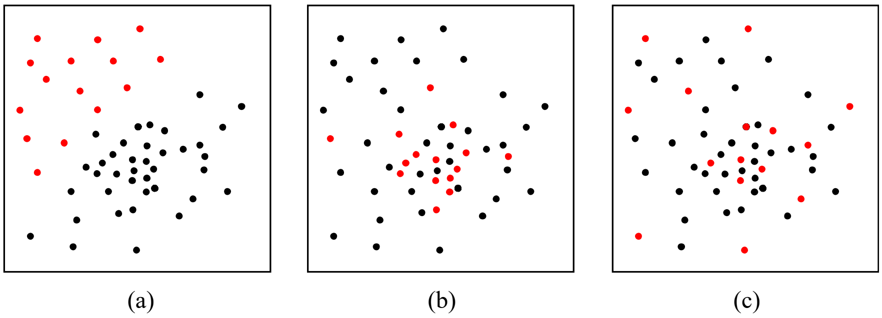


Figure 2 Spatial label patterns. (a) Clustered, (b) random, (c) dispersed.

The above analyses assume that the locational data of points are available. This does not always hold in reality, especially when points represent individuals. The information of individuals is generally aggregated across spatial units such as zip code zones and census tracts to keep their confidentiality. We can consider, however, the spatial patterns of points even if the data are spatially aggregated. Figure 3 shows hypothetical population data aggregated by census tracts. Figure 3a shows that individuals are clustered in central areas while individuals are dispersed in Figure 3b.

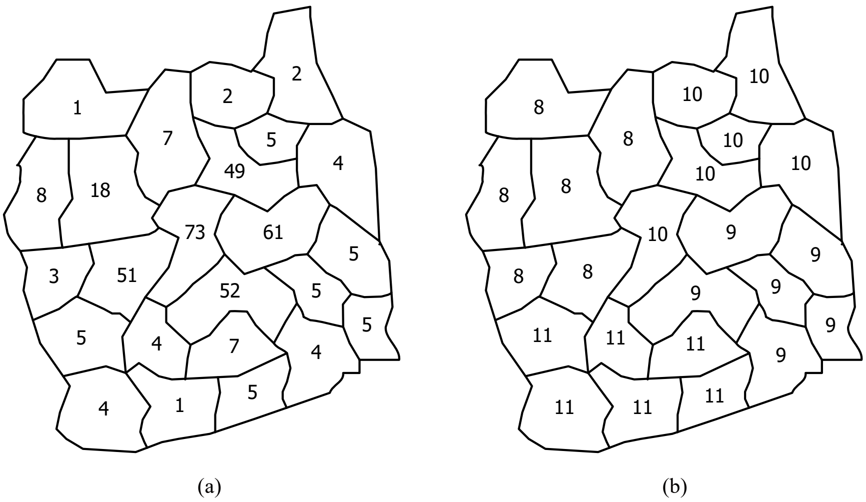


Figure 3 Hypothetical population data aggregated by census tracts. (a) Clustered pattern, (b) dispersed pattern.

The same applies to the spatial label pattern. Figure 4 shows another hypothetical population data, indicating the number of people over 65 in red. Our interest lies in whether people over 65 are relatively clustered compared to the overall population pattern. Figure 4a indicates that people over 65 are clustered in five central units, while dispersed in Figure 4b (the average proportion of people over 65 is about 1/4).

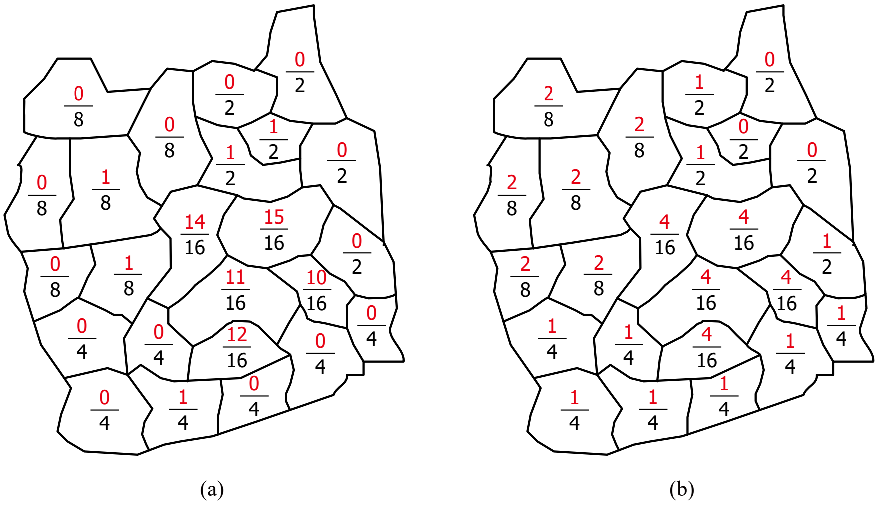


Figure 4 Hypothetical population data aggregated by census tracts. Red values indicate the number of people over 65. (a) Clustered pattern, (b) dispersed pattern.

As seen above, visual analysis lets us consider the spatial patterns of points and labels even if the point data are spatially aggregated. A disadvantage of visual analysis is that the results depend on the analyst and are somewhat subjective. Quantitative analysis is necessary, which leads to more objective conclusions. Existing methods, unfortunately, cannot fully evaluate the spatial pattern of points and labels on spatially aggregated data, as discussed in the next section. To fill the research gap, we propose a new method of point pattern analysis. We aim to classify statistically point patterns into three categories, i.e., clustered, dispersed, and random patterns, on spatially aggregated data. Section 2 reviews existing studies related to the topic of this paper. Section 3 proposed two statistics for analyzing spatially aggregated data. Section 4 tests the validity of the statistics by computational experiments. Section 5 summarizes the conclusion and discusses the topics of future research.

2. Related works

This paper considers the four requirements that should be fulfilled by the analytical method. 1) The method is directly applicable to spatially aggregated data. 2) The method can treat both spatial point and label patterns. 3) The method can statistically classify point patterns into three categories, i.e., clustered, dispersed, and random patterns. 4) The statistical power of the method is thoroughly discussed. The following discusses existing methods in terms of these requirements.

2.1 Analysis of spatial point pattern

The nearest neighbor distance method (Clark and Evans (1954); Pinder and Witherick (1972)) and Ripley's *K*-function (Ripley (1976); Ripley (1977)) are valuable tools for statistically evaluating spatial point patterns. They compare an observed point pattern with those generated under the complete spatial randomness and assess the statistical significance of the observed pattern. The quadrat method is also used in the analysis of spatial point patterns (Thompson (1958); Rogers (1974)). It places a lattice on the studied region, counts the number of points in each cell, and compares it with that obtained under the uniform distribution. The pattern is judged as non-uniform if the null hypothesis is rejected. The above methods are effective if the locational data of points are available. Unfortunately, point data are often aggregated across spatial units, especially when points represent individuals. The above methods are not directly applicable to spatially aggregated data.

Another approach is to use spatial autocorrelation indices such as Moran's I and Geary's C (Cliff (1969); Griffith (1987)). Moran's I, for instance, will show a large positive value in Figure 3a, which suggests the potential of Moran's I to detect clustered point patterns. However, though points are dispersed, Moran's I will also show a large positive value in Figure 3b. Moran's I and Geary's C cannot always distinguish clustered and dispersed point patterns. In addition, their statistical power in point pattern analysis based on spatially aggregated data is unknown. Though these indices proved effective for spatial autocorrelation analysis, it is unclear whether they continually work on data such as shown in Figure 3.

Spatial scan statistic aims to detect point clusters (Kulldorff and Nagarwalla (1995); Kulldorff (1997)). It draws circles of various sizes and locations, compares the point density inside and outside the circles, and extracts circles of higher point density. Spatial scan statistic is used to detect the clusters of disease cases (Glaz et al. (2001); Lawson (2006)), hot spots of crimes (Nakaya and Yano (2010); Shiode (2011)), hot spots of traffic accidents (Sparks (2012); Song et al. (2018)), and so forth. Spatial scan statistic, unfortunately, does not meet our demand. Our interest lies in the global pattern of points, as shown in Figures 1 and 2, while spatial scan statistic aims to detect local clusters. In addition, spatial scan statistic cannot detect dispersed point patterns.

2.2 Analysis of spatial label pattern

Cuzick and Edwards (1990) develops a statistical method to evaluate the spatial label pattern. Their method considers the random labeling as the null hypothesis, i.e., the randomization of labels without changing the location of points. The statistic is the number of labeled points within the ith nearest points from every labeled point. The colocation quotient proposed by Leslie and Kronenfeld (2011) is an extension of Cuzick and Edwards (1990). It can treat the spatial relationship between more than two types of labels, which has been extended by Cromley et al. (2014), Kronenfeld and Leslie (2015), and Kronenfeld and Leslie (2015).

These statistics, however, cannot detect dispersed label patterns such as that shown in Figure 4b since the statistics used in these methods aim to detect only clustered label patterns. Another limitation is that they basically assume point data. This prohibits them from being applied to spatially aggregated data.

As seen above, existing methods do not fully satisfy our demand. A primary weakness is that they are not effective for detecting dispersed patterns. We thus develop a new statistical method for point pattern analysis based on spatially aggregated data.

3. Method

Suppose a region Ξ consisting of L spatial units $=\{U_1, U_2, ..., U_L\}$. Each unit contains points whose numbers are given by $\{n_1, n_2, ..., n_L\}$. The total number of points is denoted by N. Area(O) indicates the area of spatial object O.

3.1 Analysis of spatial point pattern

This subsection considers the analysis of spatial point patterns. We aim to evaluate whether an observed point pattern is clustered, dispersed, or random. To this end, we randomly place w circles of radius r in such a way that they overlap Ξ. The ith circle is denoted by C_i. We define a measure

$$\mu_i = \sum_j \frac{\text{Area}\left(C_i \cap U_j\right)}{\text{Area}\left(U_j\right)} n_j + \left(\pi r^2 - \sum_j \text{Area}\left(C_i \cap U_j\right)\right)\mu_0, \tag{1}$$

where μ_0 is the average density of points in Ξ:

$$\mu_0 = \frac{N}{\sum_i \text{Area}\left(U_i\right)}. \tag{2}$$

The measure μ_i represents the estimated number of points in C_i. The second term considers the case where C_i is not fully contained in Ξ. This term extrapolates the data outside Ξ by using the average density of points. When C_i is fully contained in Ξ, the second term equals zero, and μ_i equals the first term of Equation (1).

The measure μ_i represents the overall degree of point clustering. If points are clustered, μ_i greatly varies among the circles. If points are dispersed, μ_i will show similar values. To evaluate the variation in μ_i, we use the median absolute deviation with a slight modification (Andrews et al. (1972); Hampel (1974); Rousseeuw and Croux (1993)). The original version is

$$\varphi_S = \text{med}_i \left|\mu_i - \text{med}_i \mu_i\right|. \tag{3}$$

We replace the median of μ_i with its mean to increase the statistical power:

$$\varphi_S = \text{med}_i \left|\mu_i - \bar{\mu}\right|. \tag{4}$$

The statistic φ_S represents the variation in μ_i. We perform a Monte Carlo simulation to evaluate the statistical significance of observed φ_S. The null hypothesis is the complete spatial randomness, i.e., N points are randomly distributed in Ξ. We randomly locate N points, count the number of points in each unit, and calculate φ_S using the above procedure. The probability distribution of φ_S permits us to evaluate the statistical significance of the observed pattern, i.e., whether points are statistically clustered or dispersed.

The radius r works as a parameter representing the geographic scale of analysis (Lam and Quattrochi (1992); Ruddell and Wentz (2009)). A large value gives us a macroscale perspective, while a small value allows us to discuss the local spatial pattern in detail. Ripley's K-function shows that point patterns can be evaluated as clustered and dispersed at different scales. We thus recommend trying various values to assess the spatial point patterns from various scale perspectives. The choice of w depends on the computer environment. A large w increases the statistical power but also increases the computing time.

3.2 Analysis of spatial label pattern

This subsection considers the analysis of spatial label patterns. We aim to evaluate whether an observed label pattern is clustered, dispersed, or random. Each point is either labeled or unlabeled. The number of labeled points in U_i is l_i. We randomly place circles of radius r in such a way that they overlap with Ξ and contain at least a single point. The ith circle is denoted by C_i.

We define a measure

$$\eta_i = \sum_j \frac{\text{Area}\left(C_i \cap U_j\right)}{\text{Area}\left(U_j\right)} l_j + \left(\pi r^2 - \sum_j \text{Area}\left(C_i \cap U_j\right)\right)\eta_0, \tag{5}$$

where η_0 is the average density of labeled points in Ξ:

$$\eta_0 = \frac{\sum_i l_i}{\sum_i \text{Area}\left(U_i\right)}. \tag{6}$$

The measure η_i represents the estimated number of labeled points in C_i. The second term complements the data outside Ξ if C_i is only partially contained in Ξ. When C_i is fully contained in Ξ, the second term equals zero, and η_i equals the first term of Equation (5). The estimated proportion of labeled points in C_i is given by

$$\kappa_i = \frac{\eta_i}{\mu_i}.$$

The measure κ_i represents the overall degree of label clustering. The measure κ_i greatly varies among the circles if labeled points are clustered, while the variation is small when labeled points are dispersed. We define a statistic representing the variation by

$$\varphi_L = \text{med}_i \left|\kappa_i - \bar{\kappa}\right|.$$

The statistic φ_L is large if κ_i greatly varies, while it is small when the variation is small. We perform a Monte Carlo simulation to test the statistical significance of φ_L. The null hypothesis is that points are randomly labeled, where we randomize labeled points without changing the number of points in each unit. The simulation gives us the probability distribution of φ_L, which allows us to evaluate the statistical significance of the observed pattern, i.e., whether labels are statistically clustered or dispersed.

4. Application

This section tests the validity of the statistics φ_S and φ_L through computational experiments. Subsections 4.1 and 4.2 evaluate φ_S and φ_L, respectively.

4.1 Analysis of spatial point pattern

To evaluate the validity of φ_S, we generate clustered/dispersed point patterns and use φ_S to test whether it successfully judges them as clustered/dispersed. This gives us the statistical power of φ_S, a measure of its effectiveness. The outline of experiments is as follows:

Algorithm ASPP (Analysis of Spatial Point Pattern)

Step 1: Define a spatial unit system in Ξ.
Step 2: Locate 1000 points in Ξ.
Step 3: Count the number of points in each spatial unit.
Step 4: Calculate the statistical significance of φ_S.
Step 5: Evaluate the point pattern.
Step 6: Repeat Steps 2-5 1000 times.
Step 7: Calculate the proportion of point patterns evaluated to be significant at a five percent level.

Step 1 defines a spatial unit system used for spatial aggregation as in Step 3. We used two Voronoi diagrams based on 100 generators as spatial unit systems. Generators are randomly distributed in Voronoi diagram V_1, while they are clustered around the center of Ξ in Voronoi diagram V_2. Spatial units gradually become larger from the center to the outer areas in V_2. We adopted this unit system since we often observe similar systems in the real world, i.e., spatial units are smaller in urban areas, while larger in suburban and rural areas. Figure A1 in the appendix shows these Voronoi diagrams.

Step 2 locates 1000 points in Ξ. We generated clustered point patterns at Step 2 to test the ability of φ_S to detect clustered patterns. We used the Thomas process with a slight modification (Thomas (1949); Daley and Vere-Jones (1988)). Thomas process first generates "mother points" and locates "daughter points" around the mother points. Daughter points form a clustered point pattern. The following is the detail of our approach at Step 2:

Algorithm GCPP (Generate Clustered Point Patterns)

Step 2a: Locate mother points randomly in Ξ.
Step 2b: Choose randomly a mother point.

Step 2c: Locate a daughter point around the chosen mother point according to a normal distribution.

Step 2d: Repeat Steps 2b-2c until 1000 points are located.

The number of mother points and the standard deviation of the normal distribution are denoted by M and σ, respectively. We represent the clustered point pattern obtained by Algorithms ASPP and GCPP as $PP_C(M, \sigma)$. Figure A2 shows examples of clustered point patterns.

To evaluate the ability of φ_S to detect dispersed point patterns, we generated dispersed point patterns at Step 2. We used the Matern's Type II point process, which prohibits points from being located within a predetermined distance (Matern (1960); Moller and Waagepetersen (2003)). The following is the details of this process:

Algorithm GDPP (Generate Dispersed Point Patterns)

Step 2a: Locate a point in Ξ.

Step 2b: If the point is not located within the distance $d_{\min}$ from all the existing points, keep the point.

Step 2c: Repeat Steps 2a-2b until 1000 points are located.

We denote dispersed point patterns generated by Algorithms ASPP and GDPP as $PP_D(d_{\min})$. Figure A3 shows examples of dispersed point patterns.

We use either Algorithm GCPP or GDPP as Step 2 of Algorithm ASPP. Steps 3-5 evaluate each point pattern, and we repeat these steps 1000 times, as shown in Step 6. Step 7 calculates the proportion of point patterns assessed to be significant at a five percent level. This is the statistical power of φ_S, which we denote Power(φ_S). 1-Power(φ_S) is equal to the probability of Type II error.

To perform the above computational experiments, we wrote a program in C++ and ran it on an i9-12900U CPU 2.40 GHz, RAM 128 GB computer running Windows 10 Professional. All the experiments finished within ten minutes. Concerning the number of circles w, we compared the results obtained where w=100, 500, and 1000 in some cases. We found that w=100 and 500 gave different results, but the difference was insignificant between w=500 and 1000. The following shows the results where w=500.

Table 1 shows Power(φ_S) in clustered point patterns. The statistical power of 0.8 is often said to be desirable in statistics (Zodpey (2004); Myors et al. (2010); Kraemer and Blasey (2015)). Table 1 shows that φ_S generally satisfies this requirement, especially when r=0.05 and 0.10. The point pattern becomes less clustered with an increase of M and σ, which decreases Power(φ_S). We can confirm this in Table 1, e.q., $PP_C(3, 0.1)$ gives better result than $PP_C(3, 0.4)$ and $PP_C(4, 0.1)$ where r=0.10. Power(φ_S) also decreases with an increase in r. This is probably because a large r generates circles not fully contained in Ξ that require the extrapolation by μ_0 in Equation (1). Voronoi diagrams used for spatial aggregation does not seem to affect the results. V_1 gave better results in some cases, while V_2 was better in other cases.

Table 2 shows Power(φ_S) in dispersed point patterns. Power(φ_S) is larger than 0.9 in all cases, which supports the ability of φ_S to detect dispersed point patterns. A decrease in d_{min} generates less dispersed point patterns, which decreases Power(φ_S). The statistic φ_S, however, is still effective when d_{min}=0.0005. The choice of Voronoi diagrams does not seem to affect the results again.

Table 1 Power(φ_S) in clustered point patterns $PP_C(M, \sigma)$. M and σ indicate the number of mother points and the spatial dispersion of daughter points, respectively. The upper rows use Voronoi diagram V_1, while the lower rows use V_2.

Point pattern	$PP_C(3, 0.1)$	$PP_C(3, 0.2)$	$PP_C(3, 0.4)$	$PP_C(4, 0.1)$	$PP_C(4, 0.2)$	$PP_C(4, 0.4)$	$PP_C(5, 0.1)$	$PP_C(5, 0.2)$	$PP_C(5, 0.4)$
r=0.05	1.000	0.998	0.936	1.000	1.000	0.994	1.000	0.984	0.876
r=0.10	1.000	0.984	0.814	0.998	0.970	0.888	1.000	0.974	0.834
r=0.20	1.000	0.922	0.762	0.886	0.746	0.626	0.984	0.846	0.474
r=0.05	1.000	0.998	0.934	1.000	0.992	0.884	1.000	0.998	0.848
r=0.10	1.000	0.994	0.804	1.000	0.966	0.782	0.998	0.940	0.812
r=0.20	0.996	0.928	0.756	0.992	0.866	0.644	0.982	0.860	0.606

Table 2 Power(φ_S) in dispersed point patterns $PP_D(d_{min})$. d_{min} indicates the minimum distance between points. The upper rows use Voronoi diagram V_1, while the lower rows use V_2.

Point pattern	$PP_D(0.0020)$	$PP_D(0.0010)$	$PP_D(0.0005)$
r=0.05	0.970	0.956	0.946
r=0.10	0.958	0.944	0.928
r=0.20	0.956	0.924	0.919
r=0.05	0.954	0.950	0.946
r=0.10	0.954	0.949	0.930
r=0.20	0.952	0.946	0.920

4.2 Analysis of spatial label pattern

This subsection evaluates the validity of φ_L through computational experiments. We generate clustered/dispersed label patterns and use φ_L to test whether it successfully judges them as clustered/dispersed. The outline of experiments is as follows:

Algorithm ASLP (Analysis of Spatial Label Pattern)

Step 1: Define a spatial unit system in Ξ.

Step 2: Choose a point pattern consisting of 1000 points in Ξ.

Step 3: Label a certain proportion of points.

Step 4: Count the labeled and unlabeled points in each spatial unit.

Step 5: Calculate the statistical significance of φ_L.

Step 6: Evaluate the point pattern.

Step 7: Repeat Steps 3-6 1000 times.

Step 8: Calculate the proportion of label patterns evaluated as significant at a five percent level.

Like Algorithm ASPP, Algorithm ASLP generates two Voronoi diagrams at step 1. Step 2 chose the clustered point pattern $PP_C(5, 0.2)$, the dispersed point pattern $PP_D(1.0)$, and a random point pattern, which we denote PP_R. To test the ability of φ_L to detect clustered label patterns, Step 3 generates clustered label patterns. We adopted a procedure similar to the Thomas process. We first choose mother points and label them. Let δ_i be a variable representing the relative distance from the ith mother point, initially set to one. We randomly select a mother point (assume it is the ith mother point), label its δ_ith nearest point, and add a random variable between 0 and Δ to δ_i. As we repeat this process, labeled points gradually spread around each mother point. The following is the detail of our approach at Step 3:

Algorithm GCLP (Generate Clustered Label Patterns)

Step 3a: Choose mother points from a given point pattern and label them.
Step 3b: Set δ_i=1 for all the mother points.
Step 3c: Choose a mother point and its δ_ith nearest point.
Step 3d: If the point is already labeled, go to Step 3c.
Step 3e: Label the point.
Step 3f: $\delta_i=\delta_i+\text{Rand}(\Delta)$.
Step 3g: Go to Step 3c.
Step 3h: Repeat Steps 3c-3g until enough points are labeled.

Function Rand(Δ) generates a random value between zero and Δ. Let p be the proportion of points to be labeled. We denote the clustered label pattern as $LP_C(p, \Delta)$. Figure A4 shows examples of clustered label patterns generated by Algorithms ASLP and GCLP.

We use a procedure similar to Matern's Type II process to generate dispersed label patterns. We randomly choose a point. If it is not labeled, we check whether the point is not nearer to all the labeled points than their ith nearest points. Variable i plays a role in keeping the relative distance between labeled points. If the point is not closer to all the labeled points, we label it. If not, we calculate Rand(1.0) and label the point if the obtained value exceeds 0.5. We introduce this random process to relax the strict requirement imposed by i. The following is the details of Step 3:

Algorithm GDLP (Generate Dispersed Label Patterns)

Step 3a: Choose a point.
Step 3b: If the point is already labeled, go to Step 3a.
Step 3c: If the point is not nearer to all existing labeled points than their ith nearest points,
label the point and go to Step 3a.

Step 3d: If Rand(1.0) is larger than 0.5, label the point and go to Step 3a.

Step 3e: Go to Step 3a.

Step 3f: Repeat Steps 3a-3e until enough number of points are labeled.

We denote the dispersed label pattern as $LP_D(p, i)$, where p is the proportion of points to be labeled. Figure A5 shows examples of dispersed label patterns generated by Algorithms ASLP and GDLP.

We use either Algorithm GCLP or GDPL at Step 3 of Algorithm LPA. We repeat Steps 3-6 1000 times to obtain the proportion of label patterns evaluated to be significant at a five percent level at Step 8. It is the statistical power of φ_L, which is denoted by Power(φ_L).

We wrote a program in C++ and ran it in the same computer environment as the previous subsection. Tables 3 and 4 show Power(φ_L) in clustered and dispersed label patterns, respectively. Power(φ_L) is larger than 0.8 in most cases, which supports the validity of φ_L. Labels become less clustered with an increase of Δ in Table 3, while labels become less dispersed with a decrease of i in Table 4. Power(φ_L) reduces in both cases, which is consistent with the results shown in Tables 1 and 2. An increase in r also reduces Power(φ_L). A difference from Tables 1 and 2 lies in Power(φ_L) when r=0.20, i.e., Power(φ_L) is larger than Power(φ_S) in many cases. This is probably because label pattern analysis excludes empty circles, which are more likely to occur around the boundary area. As mentioned in the previous subsection, they are often only partially contained in Ξ and reduce Power(φ_S). Label pattern analysis excludes such circles; thus, Power(φ_L) is still large when r=0.20. Concerning the proportion of labeled points, p=0.50 generally shows better results than p=0.25. The difference, however, does not seem significant. The spatial pattern of points and Voronoi diagrams do not significantly affect the results.

Table 3 Power(φ_L) in clustered label patterns $LP_C(p, \Delta)$. p and Δ indicate the proportion of labeled points and the maximum value of randomly generated values, respectively. The upper rows use Voronoi diagram V_1, while the lower rows use V_2.

Point pattern	PP_C(5, 0.2)	PP_C(5, 0.2)	PP_C(5, 0.2)	PP_D(1.0)	PP_D(1.0)	PP_D(1.0)	PP_R	PP_R	PP_R
Label pattern	LP_C(0.25, 2)	LP_C(0.25, 3)	LP_C(0.25, 4)	LP_C(0.25, 2)	LP_C(0.25, 3)	LP_C(0.25, 4)	LP_C(0.25, 2)	LP_C(0.25, 3)	LP_C(0.25, 4)
r=0.05	0.953	0.952	0.933	0.964	0.962	0.962	0.990	0.988	0.984
r=0.10	0.898	0.831	0.829	0.956	0.930	0.909	0.949	0.947	0.907
r=0.20	0.891	0.804	0.803	0.882	0.880	0.882	0.913	0.887	0.882
r=0.05	0.937	0.922	0.929	0.962	0.929	0.953	0.985	0.975	0.968
r=0.10	0.887	0.844	0.832	0.897	0.856	0.816	0.898	0.896	0.883
r=0.20	0.854	0.813	0.804	0.810	0.807	0.796	0.840	0.809	0.834

Point pattern	PP_C(5, 0.2)	PP_C(5, 0.2)	PP_C(5, 0.2)	PP_D(1.0)	PP_D(1.0)	PP_D(1.0)	PP_R	PP_R	PP_R
Label pattern	LP_C(0.50, 2)	LP_C(0.50, 3)	LP_C(0.50, 4)	LP_C(0.50, 2)	LP_C(0.50, 3)	LP_C(0.50, 4)	LP_C(0.50, 2)	LP_C(0.50, 3)	LP_C(0.50, 4)
r=0.05	0.936	0.934	0.918	0.998	0.970	0.930	0.993	0.972	0.931
r=0.10	0.969	0.883	0.856	0.948	0.902	0.893	0.971	0.948	0.918
r=0.20	0.894	0.833	0.824	0.894	0.832	0.816	0.905	0.852	0.785
r=0.05	0.939	0.970	0.917	0.996	0.917	0.905	1.000	0.960	0.916
r=0.10	0.960	0.951	0.848	0.924	0.869	0.859	0.976	0.862	0.851
r=0.20	0.823	0.817	0.813	0.841	0.829	0.819	0.824	0.805	0.803

Table 4 Power(φ_L) in clustered label patterns $LP_D(p, i)$. p and i indicate the proportion of labeled points and the minimum relative distance between labeled points, respectively. The upper rows use Voronoi diagram V_1, while the lower rows use V_2.

Point pattern	PP_C(5, 0.2)	PP_C(5, 0.2)	PP_C(5, 0.2)	PP_D(1.0)	PP_D(1.0)	PP_D(1.0)	PP_R	PP_R	PP_R
Label pattern	LP_D(0.25, 4)	LP_D(0.25, 3)	LP_D(0.25, 2)	LP_D(0.25, 4)	LP_D(0.25, 3)	LP_D(0.25, 2)	LP_D(0.25, 4)	LP_D(0.25, 3)	LP_D(0.25, 2)
r=0.05	0.985	0.972	0.967	0.960	0.957	0.952	0.954	0.949	0.948
r=0.10	0.927	0.913	0.906	0.931	0.930	0.918	0.923	0.887	0.858
r=0.20	0.862	0.860	0.837	0.923	0.898	0.874	0.859	0.856	0.803
r=0.05	0.990	0.978	0.972	0.981	0.969	0.955	0.976	0.961	0.944
r=0.10	0.936	0.935	0.908	0.923	0.919	0.864	0.901	0.899	0.869
r=0.20	0.864	0.851	0.848	0.847	0.825	0.815	0.817	0.810	0.798

Point pattern	PP_C(5, 0.2)	PP_C(5, 0.2)	PP_C(5, 0.2)	PP_D(1.0)	PP_D(1.0)	PP_D(1.0)	PP_R	PP_R	PP_R
Label pattern	LP_D(0.50, 4)	LP_D(0.50, 3)	LP_D(0.50, 2)	LP_D(0.50, 4)	LP_D0.50, 3)	LP_D(0.50, 2)	LP_D(0.50, 4)	LP_D(0.50, 3)	LP_D(0.50, 2)
r=0.05	0.974	0.947	0.931	0.953	0.931	0.923	0.939	0.926	0.902
r=0.10	0.934	0.895	0.889	0.929	0.901	0.880	0.902	0.897	0.864
r=0.20	0.868	0.832	0.811	0.900	0.858	0.840	0.876	0.855	0.809
r=0.05	0.943	0.943	0.903	0.934	0.928	0.920	0.929	0.928	0.887
r=0.10	0.892	0.852	0.844	0.860	0.857	0.821	0.877	0.849	0.838
r=0.20	0.851	0.838	0.810	0.813	0.811	0.803	0.838	0.815	0.806

5. Conclusion

This paper has proposed a new method of point pattern analysis on spatially aggregated data. A strength of our method is that it is effective even if the locational data of points are unavailable, which is not satisfied by existing methods. We proposed two statistics φ_S and φ_L, for analyzing spatial point and label patterns, respectively. Computational experiments showed that the statistical power of these statistics is large enough in most cases, which supports the effectiveness of the statistics. The statistics φ_S and φ_L successfully detected clustered and dispersed patterns of points and labels.

The proposed method meets the four requirements mentioned in Section 2. 1) It is directly applicable to spatially aggregated data. 2) It can treat both spatial point and label patterns. 3) It can statistically point patterns into three categories, i.e., clustered, dispersed, and random patterns. 4) The statistical power of the method is thoroughly discussed. The proposed method, however, is not free from limitations. We will discuss them and extensions for future research.

Firstly, more efficient methods are necessary for evaluating the statistical significance of φ_S and φ_L. We employed Monte Carlo simulations in the experiments, and fortunately, all the experiments finished within ten minutes. However, an increase of points and spatial units clearly increases the computing time. The best solution is to derive analytical forms of the probability distribution of φ_S and φ_L under the null hypothesis. However, it is not clear whether we can derive them by a mathematical procedure. Another option is to improve the efficiency of Monte Carlo simulation by developing faster subroutines used in our programs. This might be a realistic solution.

Secondly, the minimum number of points per spatial unit needs further discussion. In our experiments, we used 1000 points and 100 spatial units, i.e., one unit contains ten points on average. In some cases, we tried from 1 to 10 points per spatial unit in some cases for φ_S, and five points seemed enough to assure the statistical power. The minimum number of points, however, heavily depends on the spatial point pattern and the radius of circles. Further experiments are necessary to cover a wider variety of situations.

Thirdly, label pattern analysis considers two types of points, i.e., labeled and unlabeled. The colocation quotient (Leslie and Kronenfeld (2011)), on the other hand, can treat more than two types of points. Given many types of points, the colocation quotient evaluates the spatial proximity between every pair of point types. The statistic φ_L, unfortunately, is not directly applicable to the cases of more than two types of points. We should extend our method to treat these cases in future research.

Fourthly, an extension of Moran's I seems worth considering. Moran's I cannot evaluate Figure 3b as dispersed since it uses the randomization of the number of points as the null hypothesis. Suppose that the complete spatial randomness is the null hypothesis instead. Moran's I calculated under the complete spatial randomness will probably be larger than that of Figure 3b, and hence Moran's I will evaluate Figure 3b as dispersed. Though we have not yet discussed this approach in detail, it seems an interesting extension.

Data and code availability statement

The programs used in Section 4 are available on Figshare at

https://figshare.com/articles/dataset/Point_pattern_analysis/25999918

Acknowledgments

This research was supported by JSPS KAKENHI Grant Numbers 19H02375, 22H00245, and 23H01634.

References

ANDREWS, D., *et al.*, 1972. Robust estimates of location: survey and advances. Princeton, NJ: Princeton University Press.

BISSON, I. A., FERRER, M. and BIRD, D. M. 2002. Factors influencing nest-site selection by Spanish Imperial Eagles. *Journal of Field Ornithology,* 73(3), 298-302.

BOOTS, B. N. and GETIS, A., 1988. *Point pattern analysis.* Newbury Park, CA: Sage Publications.

BRANTINGHAM, P. J. and BRANTINGHAM, P. L., 1981. *Environmental criminology.* Beverly Hills, CA: Sage Publications.

CLARK, P. J. and EVANS, F. C. 1954. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology,* 35(4), 445-453.

CLIFF, A. D. ORD, J.K. 1969. The problem of spatial autocorrelation. *In:* SCOTT, A. J. ed. *London Papers in Regional Science.* London: Pion, 25-55.

CROMLEY, R. G., HANINK, D. M. and BENTLEY, G. C. 2014. Geographically weighted colocation quotients: specification and application. *The Professional Geographer,* 66(1), 138-148.

CUI, C. and HAN, Z., Spatial patterns of retail stores using POIs data in Zhengzhou, China. Ed. *2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM)*, 2015, 88-92.

CUZICK, J. and EDWARDS, R. 1990. Spatial Clustering for Inhomogeneous Populations. *Journal of the Royal Statistical Society. Series B (Methodological),* 52(1), 73-104.

DALEY, D. J. and VERE-JONES, D., 1988. An Introduction to the Theory of Point Processes, Springer Series in Statistics. New York: Springer-Verlag.

DIGGLE, P. J., 1983. *Statistical analysis of spatial point patterns.* London: Academic Press.

DIGGLE, P. J. and CHETWYND, A. G. 1991. Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics,* 47, 1155-1163.

GLAZ, J., POZDNYAKOV, V. and WALLENSTEIN, S., 2001. *Scan statistics.* Berlin: Springer.

GRIFFITH, D. A., 1987. *Spatial Autocorrelation: A Primer.* Washington, DC: Association of American Geographers.

HAMPEL, F. R. 1974. The influence curve and its role in robust estimation. *Journal of the American Statistical Association,* 69(346), 383-393.

ISHIZAKI, K. 1995. Spatial competition and marketing strategy of fast food chains in Tokyo. *Geographical review of Japan, Series B.,* 68(1), 86-93.

KRAEMER, H. C. and BLASEY, C., 2015. *How many subjects?: Statistical power analysis in research.* Thousand Oaks, CA: Sage Publications.

KRONENFELD, B. J. and LESLIE, T. F. 2015. Restricted random labeling: testing for between-group interaction after controlling for joint population and within-group spatial structure. *Journal of Geographical Systems,* 17(1), 1-28.

KULLDORFF, M. 1997. A spatial scan statistic. *Communications in Statistics - Theory and Methods,* 26(6), 1481-1496.

KULLDORFF, M. and NAGARWALLA, N. 1995. Spatial disease clusters: Detection and inference. *Stat Med,* 14(8), 799-810.

LAM, N. S.-N. and QUATTROCHI, D. A. 1992. On the Issues of Scale, Resolution, and Fractal Analysis in the Mapping Sciences*. *The Professional Geographer,* 44(1), 88-98.

LAWSON, A. B. 2006. Disease cluster detection: a critique and a Bayesian proposal. *Stat Med,* 25(5), 897-916.

LAWSON, A. B., 2013. *Statistical methods in spatial epidemiology.* Chichester: John Wiley & Sons.

LESLIE, T. F. and KRONENFELD, B. J. 2011. The Colocation Quotient: A New Measure of Spatial Association Between Categorical Subsets of Points. *Geographical Analysis,* 43(3), 306-326.

LUO, H. and YANG, Y. 2013. Spatial pattern of hotel distribution in China. *Tourism and Hospitality Research,* 13(1), 3-15.

MATéRN, B. 1960. Stochastic models and their application to some problems in forest surveys and other sampling investigations. *Medd. Statens Skogsforskningsinstitut,* 49(5), 1-144.

MOLLER, J. and WAAGEPETERSEN, R. P., 2003. *Statistical inference and simulation for spatial point processes.* Boca Raton, FL: CRC Press.

MYORS, B., MURPHY, K. R. and WOLACH, A., 2010. *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests.* London: Routledge.

NAKAYA, T. and YANO, K. 2010. Visualising Crime Clusters in a Space‐time Cube: An Exploratory Data‐analysis Approach Using Space‐time Kernel Density Estimation and Scan Statistics. *Transactions in GIS,* 14(3), 223-239.

PENTTINEN, A., STOYAN, D. and HENTTONEN, H. M. 1992. Marked point processes in forest statistics. *Forest science,* 38(4), 806-824.

PETERSON, B. and GAUTHIER, G. 1985. Nest site use by cavity-nesting birds of the Cariboo Parkland, British Columbia. *The Wilson Bulletin*, 319-331.

PINDER, D. A. and WITHERICK, M. 1972. The principles, practice and pitfalls of nearest-neighbour analysis. *Geography,* 57(4), 277-288.

PRAYAG, G., LANDRé, M. and RYAN, C. 2012. Restaurant location in Hamilton, New Zealand: Clustering patterns from 1996 to 2008. *International Journal of Contemporary Hospitality Management,* 24(3), 430-450.

RABINO, G. and MASTRANGELO, L., 2002. Point pattern analysis: an application to the loyalty networks of chain-stores. *42nd Congress of the European Regional Science Association.* Dortmund, Germany.

RIPLEY, B. D. 1976. The second-order analysis of stationary point processes. *Journal of Applied Probability,* 13(2), 255-266.

RIPLEY, B. D. 1977. Modelling spatial patterns. *Journal of the Royal Statistical Society. Series B (Methodological),* 39, 172-212.

ROGERS, A., 1974. *Statistical analysis of spatial dispersion: the quadrat method.* London: Pion.

ROUSSEEUW, P. J. and CROUX, C. 1993. Alternatives to the median absolute deviation. *Journal of the American Statistical Association,* 88(424), 1273-1283.

RUDDELL, D. and WENTZ, E. A. 2009. Multi‐tasking: Scale in geography. *Geography Compass,* 3(2), 681-697.

SHIODE, S. 2011. Street‐level spatial scan statistic and STAC for analysing street crime concentrations. *Transactions in GIS,* 15(3), 365-383.

SONG, J., WEN, R. and YAN, W., Identification of traffic accident clusters using Kulldorff's space-time scan statistics. Ed. *2018 IEEE International Conference on Big Data (Big Data)*, 2018, 3162-3167.

SOURIS, M., 2019. *Epidemiology and geography: principles, methods and tools of spatial analysis.* John Wiley & Sons.

SPARKS, R. 2012. Spatially clustered outbreak detection using the EWMA scan statistics with multiple sized windows. *Communications in Statistics-Simulation and Computation,* 41(9), 1637-1653.

THOMAS, M. 1949. A generalization of Poisson's binomial limit for use in ecology. *Biometrika,* 36(1/2), 18-25.

THOMPSON, H. 1958. The statistical study of plant distribution patterns using a grid of quadrats. *Australian Journal of Botany,* 6(4), 322-342.

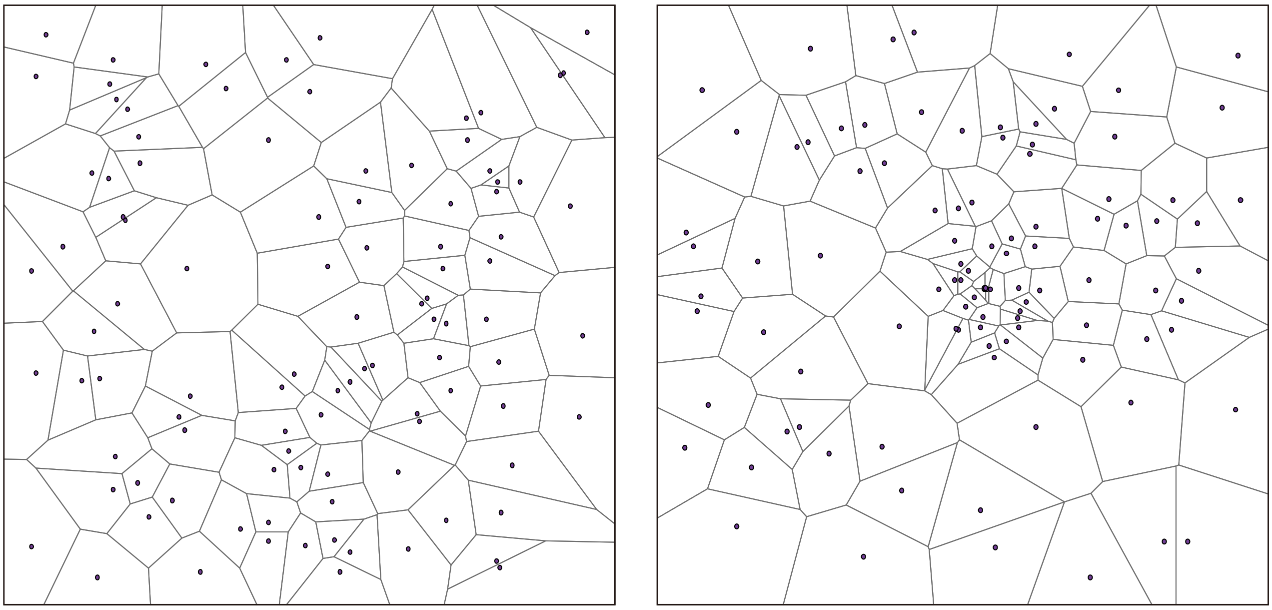
WALL, G., DUDYCHA, D. and HUTCHINSON, J. 1985. Point pattern analyses of accomodation in Toronto. *Annals of Tourism Research,* 12(4), 603-618.

WARREN, W. 1972. Point processes in forestry. *Stochastic Point Processes,* 801, 816.

WORTLEY, R., MAZEROLLE, L. G. and ROMBOUTS, S., 2008. *Environmental criminology and crime analysis.* Boca Raton, FL: Routledge.

ZODPEY, S. P. 2004. Sample size and power analysis in medical research. *Indian journal of dermatology venereology and leprology,* 70, 123-128.

Appendix



(a) (b)

Figure A1 Voronoi diagrams used for spatial aggregation. Generators are (a) randomly distributed in V_1, (b) clustered around the center in V_2.

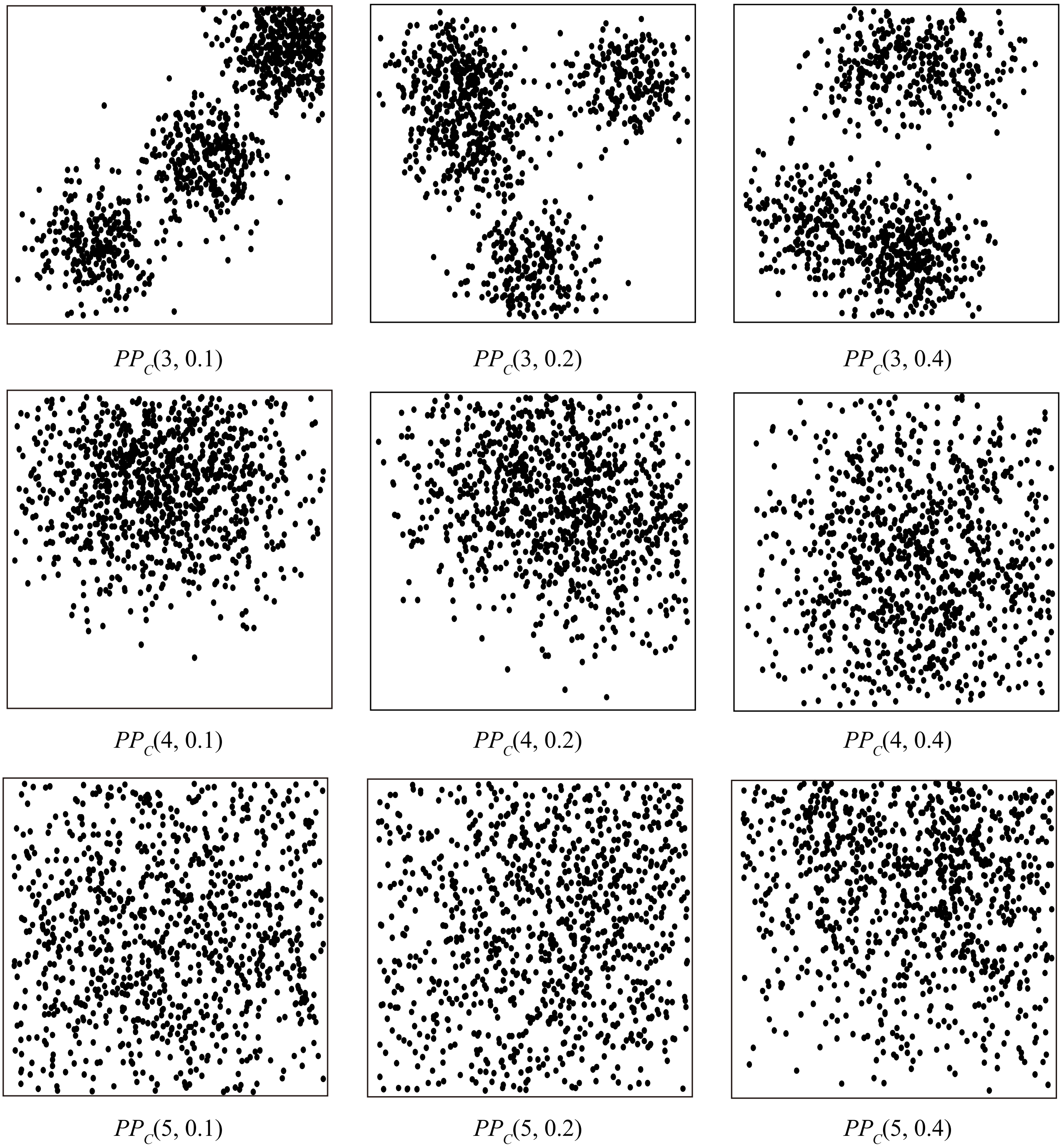
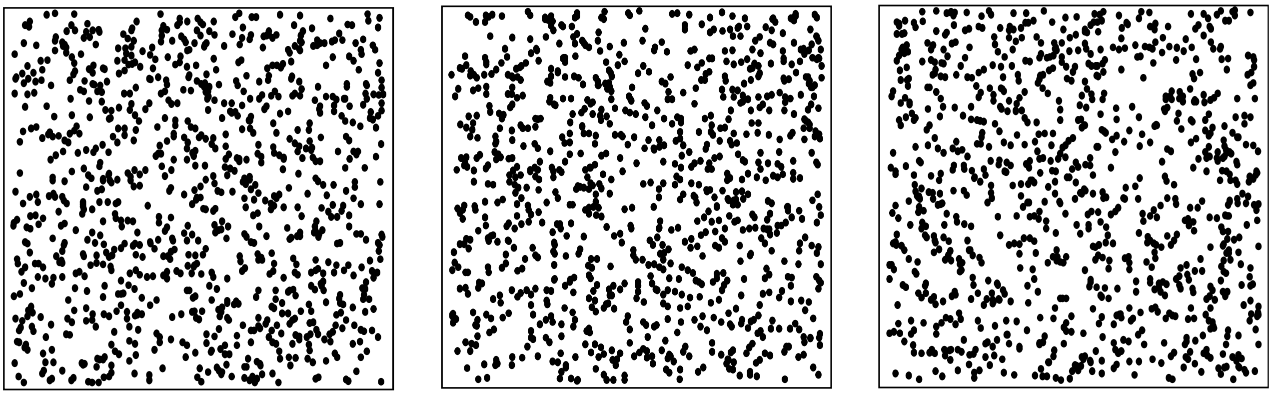
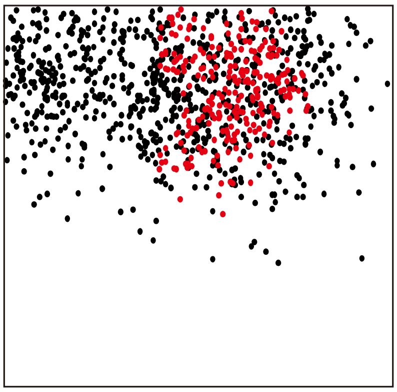


Figure A2 Clustered point patterns $PP_C(M, \sigma)$ generated by Algorithms ASPP and GCPP. M and σ represent the number of mother points and the standard deviation of the normal distribution, respectively.

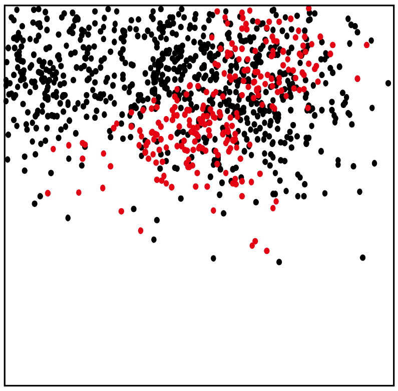


$PP_D(0.0020)$ $PP_D(0.0010)$ $PP_D(0.0005)$

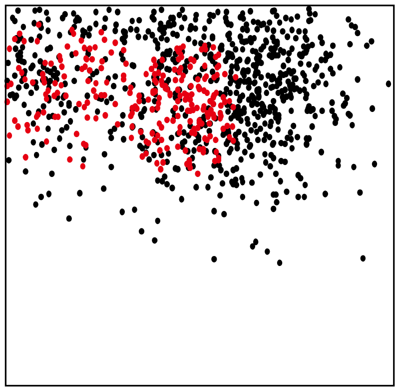
Figure A3 Dispersed point patterns $PP_D(d_{min})$ generated by Algorithms ASPP and GDPP. d_{min} indicates the minimum distance between points.



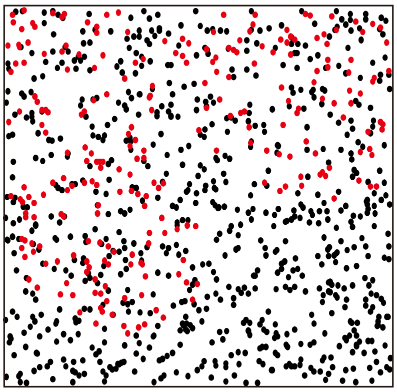
$PP_C(5, 0.2)$, $LP_C(0.25, 2)$



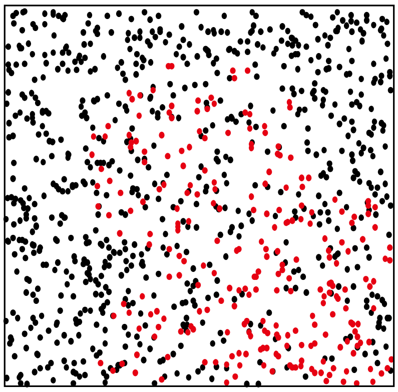
$PP_C(5, 0.2)$, $LP_C(0.25, 3)$



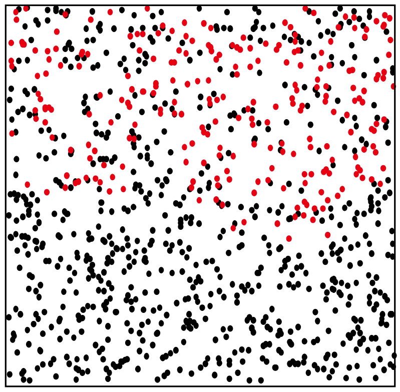
$PP_C(5, 0.2)$, $LP_C(0.25, 4)$



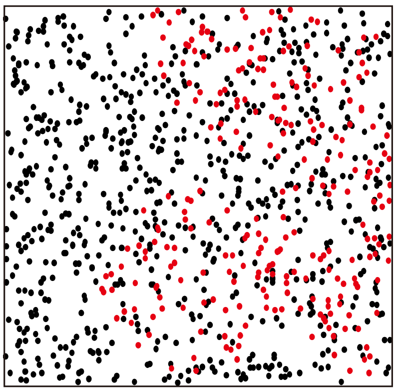
$PP_D(1.0)$, $LP_C(0.25, 2)$



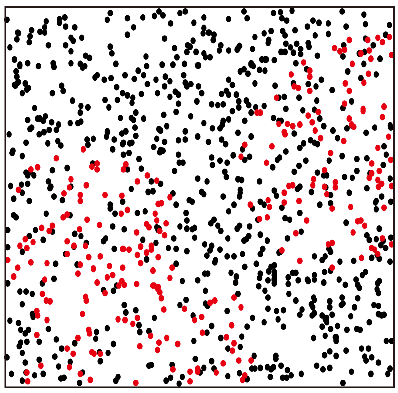
$PP_D(1.0)$, $LP_C(0.25, 3)$



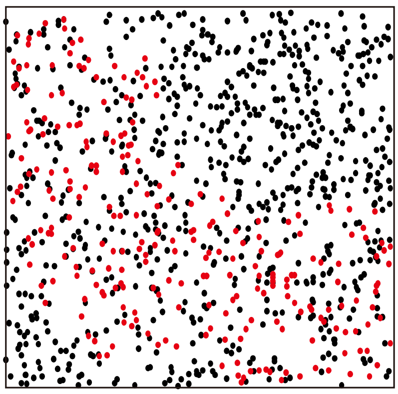
$PP_D(1.0)$, $LP_C(0.25, 4)$



PP_R, $LP_C(0.25, 2)$



PP_R, $LP_C(0.25, 3)$



PP_R, $LP_C(0.25, 4)$

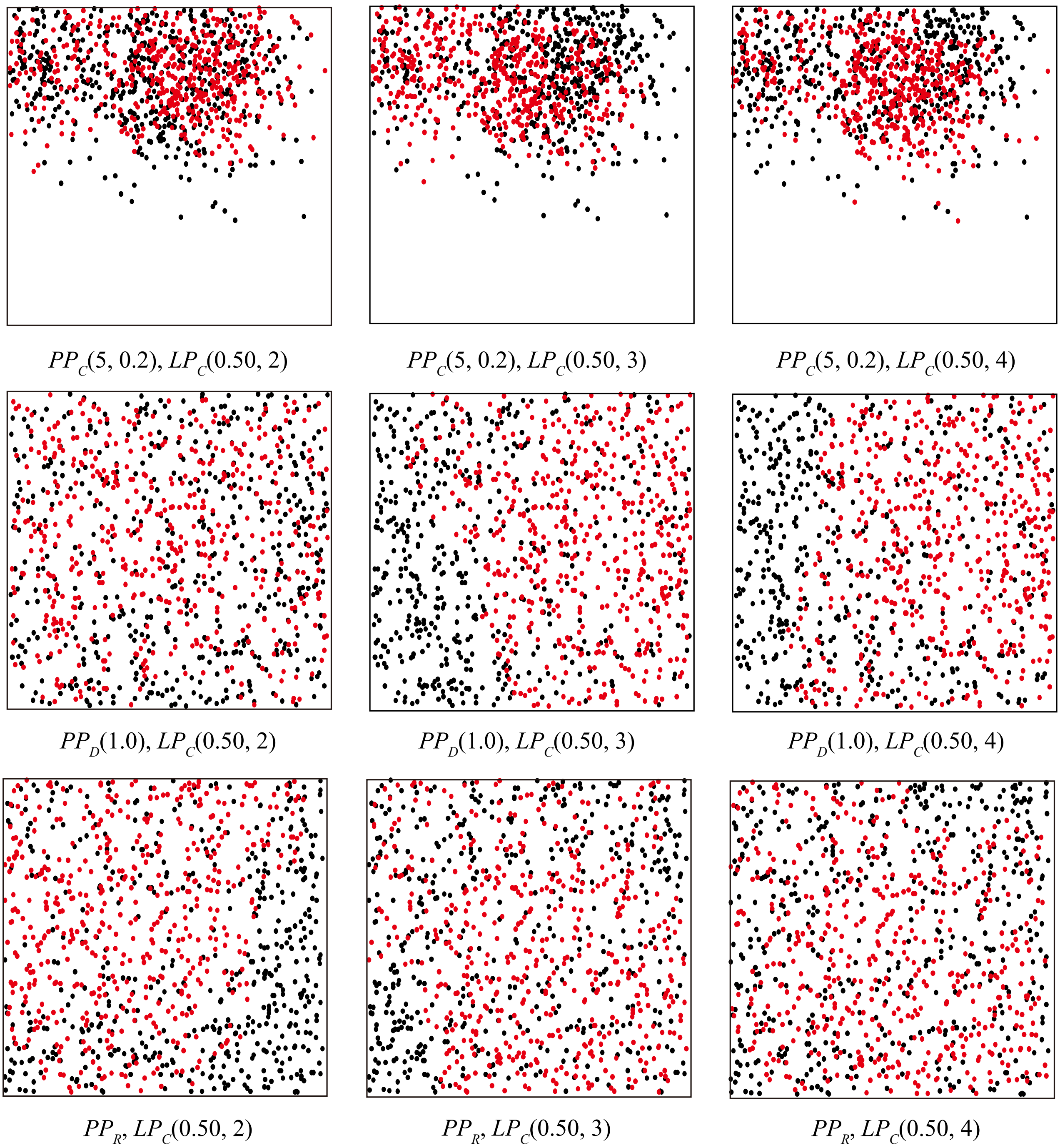
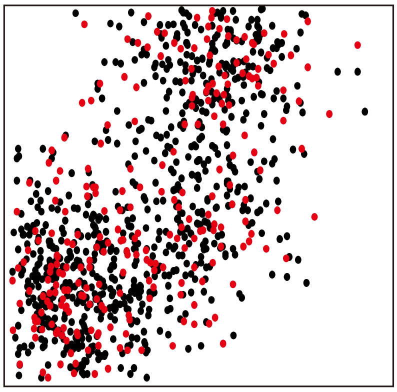
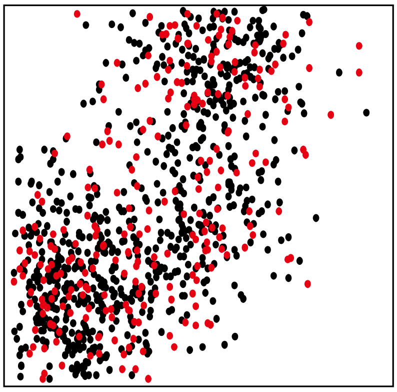


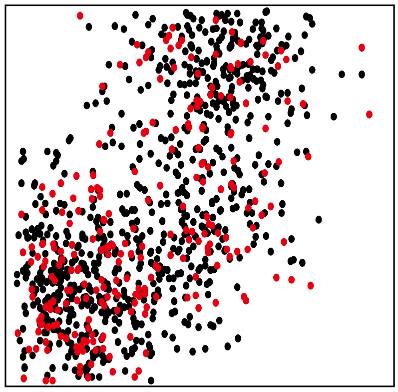
Figure A4 Clustered label patterns $LP_C(p, \Delta)$ generated by Algorithms ASLP and GCLP. Red points represent labeled points. p is the proportion of points to be labeled.



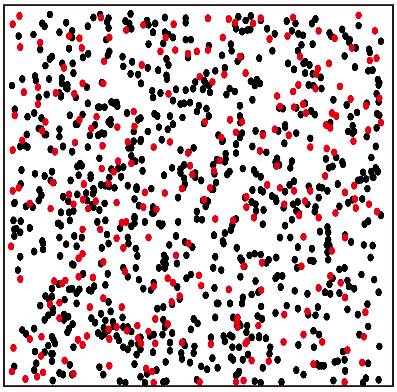
$PP_C(5, 0.2)$, $LP_D(0.25, 4)$



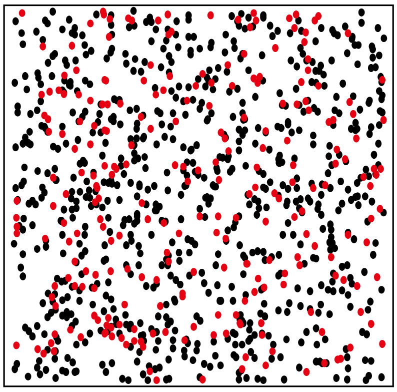
$PP_C(5, 0.2)$, $LP_D(0.25, 3)$



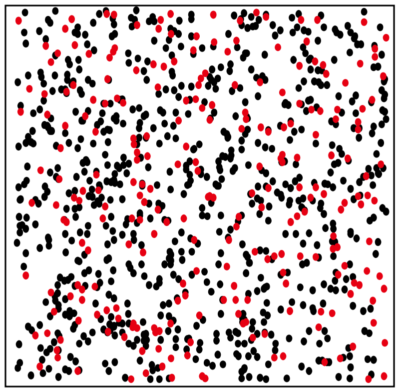
$PP_C(5, 0.2)$, $LP_D(0.25, 2)$



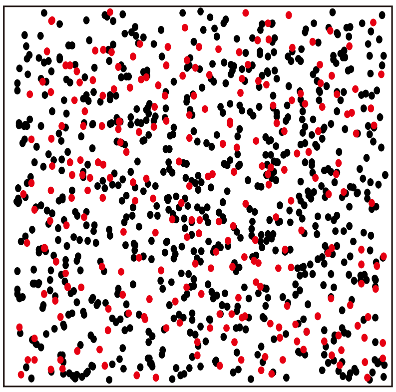
$PP_D(1.0)$, $LP_D(0.25, 4)$



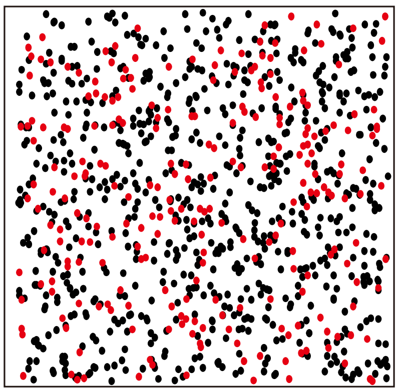
$PP_D(1.0)$, $LP_D(0.25, 3)$



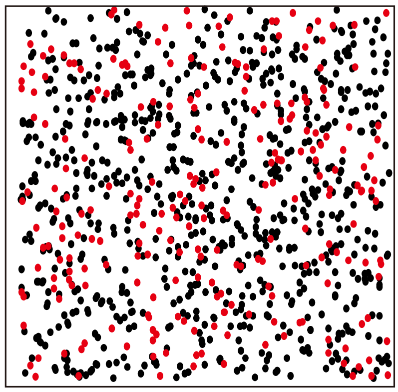
$PP_D(1.0)$, $LP_D(0.25, 2)$



PP_R, $LP_D(0.25, 4)$



PP_R, $LP_D(0.25, 3)$



PP_R, $LP_D(0.25, 2)$

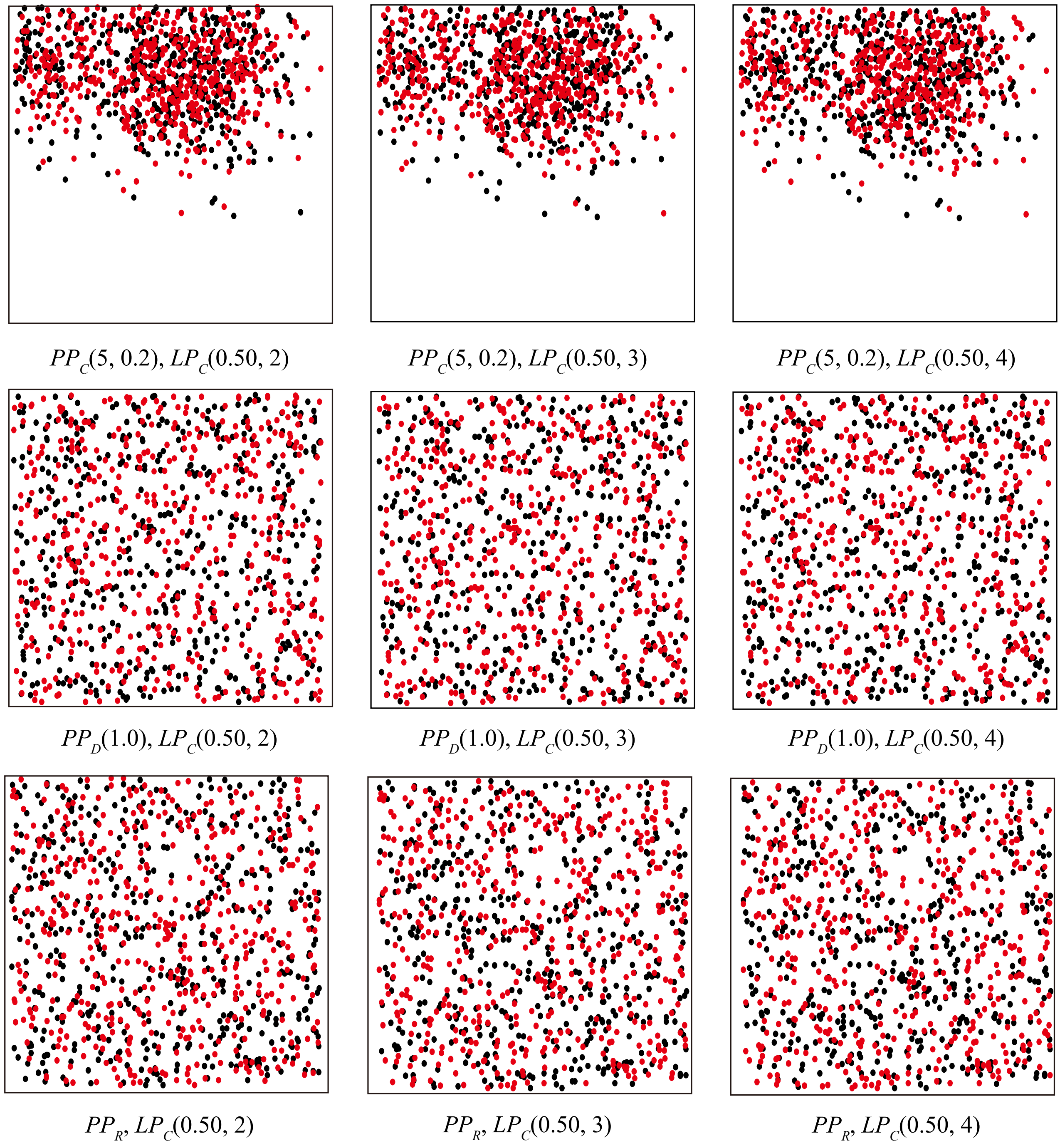


Figure A5 Dispersed label patterns $LP_D(p, i)$ generated by Algorithms ASLP and GDLP. Red points represent labeled points. p and i are the proportion of points to be labeled and the minimum distance between labeled points, respectively.