Stoch Environ Res Risk Assess (2014) 28:1277–1306
DOI 10.1007/s00477-013-0824-9

ORIGINAL PAPER

Simulation and characterization of multi-class spatial patterns from stochastic point processes of randomness, clustering and regularity

Elvan Ceyhan

Published online: 16 November 2013

Abstract Spatial pattern analysis of data from multiple classes (i.e., multi-class data) has important implications. We investigate the resulting patterns when classes are generated from various spatial point processes. Our null pattern is that the nearest neighbor probabilities being proportional to class frequencies in the multi-class setting. In the two-class case, the deviations are mainly in two opposite directions, namely, segregation and association of the classes. But for three or more classes, the classes might exhibit mixed patterns, in which one pair exhibiting segregation, while another pair exhibiting association or complete spatial randomness independence. To detect deviations from the null case, we employ tests based on nearest neighbor contingency tables (NNCTs), as NNCT methods can provide an omnibus test and post-hoc tests after a significant omnibus test in a multi-class setting. In particular, for analyzing these multi-class patterns (mixed or not), we use an omnibus overall test based on NNCTs. After the overall test, the pairwise interactions are analyzed by the post-hoc cell-specific tests based on NNCTs. We propose various parameterizations of the segregation and association alternatives, list some appealing properties of these patterns, and propose three processes for the two-class association pattern. We also consider various clustering and regularity patterns to determine which one(s) cause segregation from or association with a class from a homogeneous Poisson process and from other processes as well. We perform an extensive Monte Carlo simulation study to investigate the newly proposed association patterns and to understand which stochastic processes might result in segregation or association. The methodology is illustrated on two real life data sets from plant ecology.

Keywords Complete spatial randomness · Nearest neighbor contingency table · Random labeling · Relative abundance · Spatial clustering

1 Introduction

The spatial interaction or clustering of points from multiple classes has important implications in various fields. For example, in ecology, the interaction between tree species might be of interest, while in microbiology the interaction between different cell types might be of concern. Spatial interaction among species (including association of species) also has important implications and potential for applicability in biodiversity theory (Illian and Burslem 2007). We investigate the multi-class patterns with respect to the null pattern of randomness in nearest neighbor (NN) structure which causes NN probabilities proportional to the class frequencies. This randomness could be resulting from complete spatial randomness (CSR) independence or random labeling (RL) among others. In the two-class case, the deviations from this null pattern are mainly in two opposite directions: segregation and association. Roughly defined, association is the pattern in which points from different classes are closer to each other, while segregation is the spatial pattern in which points from the same class are closer to each other. Also, in the presence of three or more classes, the classes or species might exhibit mixed patterns, where some classes could be segregated, while others are associated or conforming to CSR independence or RL.

In literature, spatial association is used for both univariate and multivariate spatial data, which could be from

E. Ceyhan (✉)
Department of Mathematics, Koç University, Sarıyer,
34450 Istanbul, Turkey
e-mail: elceyhan@ku.edu.tr

the same class or different classes. Among the types of spatial association are point, line, and areal spatial association. In point spatial association a distance measure or metric is employed; in line spatial association, distances and paths are used; and in areal spatial association, distance and contiguity are used. In literature there are various measures of association between two classes. That is, in a bivariate spatial process (X, Y), where both X and Y are point processes, one can use bivariate counterparts of F-, G-, K-, g-, and J statistics (Diggle 2003; Ripley 2004; Stoyan 1984; van Lieshout and Baddeley 1999; Harkness and Isham 1983). For example, Comas et al. (2009) employ the univariate and bivariate versions of the (inhomogeneous) pair correlation function to assess the spatial distribution of trees in a forest in Central Catalonia. Their univariate analysis indicates that *P. sylvestris* trees tend to be clustered at short distances, while *P. nigra* and *P. halepensis* seem to exhibit some regularity. Furthermore, they observe segregation between trees from different species. The literature is not as rich when one class, say X is a point process, while Y is a random set. Foxall and Baddeley (2002) generalize van Lieshout and Baddeley's J-function for such a case motivated from a geological application where Y is a set of "lineaments" (line segments believed to represent geological faults). Berman (1986) also proposes some (parametric) measures of association between a point process and another stochastic process (again motivated by an example from geology). Spatial association between two variables (as opposed to two classes) is also studied in literature (see, e.g., Waller et al. (2007) where the authors use geographically weighted regression and spatially varying coefficient models to investigate the association between violence and alcohol consumption). Spatio-temporal clustering also became a topic of interest in literature recently. For instance, Meliker and Jacquez (2007) investigate the spatio-temporal clustering of cases and controls in a residential setting using Q-statistics, which is developed by the authors' group and uses Monte Carlo randomization to attach significance to the observed interaction between the two groups. However, in this article, we are concerned with (non-temporal) point spatial association for multi-class data.

Spatial segregation is relatively easy to parameterize (see, e.g., Ceyhan 2008), while spatial association is not. Among many types of spatial clustering tests (Kulldorff 2006), we will use tests based on nearest neighbor contingency table (NNCTs) (Dixon 1994). We first state some desirable properties for the segregation/association patterns, and propose three bivariate association patterns and then perform extensive Monte Carlo simulations to investigate these patterns by the cell-specific and overall segregation tests based on NNCTs (Ceyhan 2008). Moreover, we investigate the mixed alternatives, where the pattern between classes in a pair could be segregation while the pattern between classes in another pair could be association or CSR independence (or RL). We also consider various regularity and clustering patterns and explore which one(s) result in segregation or association with respect to a homogeneous Poisson process (HPP) and also with respect to other processes. The proposed methodology provides guidelines to assess the multi-class interaction (as segregation or association with respect to CSR independence or RL) by NNCT-tests and in the presence of significant segregation/association provides guidelines to simulate the underlying patterns that might account for or explain the observed interaction between classes. A similar approach is taken by Uria-Diez et al. (2013) to understand the dependence (on abiotic and biotic factors) of spatial distribution of a plant. The authors first fit both homogeneous and inhomogeneous versions of spatial point process models to assess the underlying generative process for the plant cohorts over the 2 years (labeled as adults and seedlings, respectively). They also evaluate the spatial interaction between the adults and seedlings by Ripley's bivariate L-function which only suggested a weak positive association between adults and seedlings at small distance values. The fitted point process models for the three cohorts indicate that the level of clustering decreases from seedlings to adults. Our methodology in this article applies a process in the reverse order (in our example data analysis): we first assess the interaction by tests based on NNCTs and then attempt fitting point process models to understand why a particular multi-class interaction occurs.

We describe null and alternative patterns and tests based on NNCTs in Sect. 2, provide desirable properties and various parameterizations of the segregation and association alternatives in Sect. 3, spatial patterns resulting from various point processes in Sect. 4, discuss spatial patterns between three classes in Sect. 5, illustrate the methodology on two real life data sets from plant ecology in Sect. 6, and provide discussion and conclusions in Sect. 7.

2 Preliminaries

2.1 Null and alternative patterns

In a multi-class setting, the null pattern we consider is

> H_o: NN probabilities are proportional to class frequencies

which may result from (among other patterns) RL or complete spatial randomness (CSR) independence of points from two classes. RL is the pattern where class labels are randomly assigned to a given set of points; and under CSR independence, each class satisfies CSR independently of

Table 1 The NNCT for k classes

	NN class			Total
	Class 1	…	Class k	
Base class				
Class 1	N_{11}	…	N_{1k}	n_1
⋮	⋮	⋱	⋮	⋮
Class k	N_{k1}	…	N_{kk}	n_k
Total	C_1	…	C_k	n

other classes, i.e., each class is uniformly distributed in the region of interest provided their sample sizes are fixed, otherwise they are from a HPP. In the two-class case, there are two major types of alternatives which are resulting from deviations from this null hypothesis: *segregation* and *association*. *Segregation* occurs if the NN of an individual is more likely to be of the same class as the individual than to be from a different class in the sense that the probability that this individual has a NN from the same class is larger than the relative frequency of the same class (see, e.g., Pielou 1961). *Association* occurs if the NN of an individual is more likely to be from another class than to be of the same class as the individual in the sense that the probability that this individual has a NN from another class is larger than the relative frequency of the other class in question. See Ceyhan (2008) for more detail on the null and alternative patterns.

2.2 Cell-specific and overall segregation tests based on NNCTs

We employ tests based on NNCTs (referred to as NNCT-tests, henceforth) to explore the spatial patterns in a multi-class setting, since NNCT-tests provide an omnibus test of overall deviation from the null pattern in a multi-class setting (similar to an ANOVA F-test in a multi-group setting). We then resort to cell-specific tests as post-hoc tests after a significant overall test (as in pairwise t-tests after a significant F-test). To our knowledge, NNCT methodology is the only one with this property in spatial data analysis (i.e., having an omnibus overall test in a multi-class setting and then cell-specific tests as the pairwise post-hoc tests after the overall test). Below we provide a brief description of NNCTs, see Ceyhan (2008) for more details. NNCTs are constructed using the NN frequencies of classes. For k classes, we will have a $k \times k$ NNCT, which would have N_{ij} in cell (i, j) where N_{ij} is the number of times the NN of a class i point is from class j. We present the NNCT in Table 1 where C_j is the sum of column j; i.e., number of times class j points serve as NNs for $j \in \{1,2,\ldots,k\}$ and n_i is the sum of row i, i.e., sample size for class i for $i = 1,2,\ldots,k$. We adopt the convention that variables denoted by upper case letters are random quantities, while lower case letters represent fixed quantities. In a NNCT-analysis, row sums are assumed to be fixed (i.e., class sizes are given), while column sums are assumed to be random and depend on the NN relationships between the classes. Under segregation of class i from other classes, the diagonal cell counts, N_{ii}, would be larger than expected, while under association of class j with class i with $i \neq j$, the off-diagonal cell counts, N_{ij}, would be larger than expected.

Dixon's cell-specific segregation tests and four new cell-specific tests together with the corresponding overall segregation tests are defined and compared in Ceyhan (2008). In the same article, it has been shown that among the cell-specific and overall tests, Dixon's and type III tests have better size and power performance. Hence in this article, we only use these tests in our investigations.

The test statistic suggested by Dixon for cell (i, j) is given by

$$Z_{ij}^D = \frac{N_{ij} - \mathbf{E}[N_{ij}]}{\sqrt{\mathbf{Var}[N_{ij}]}}. \tag{1}$$

Under RL or CSR independence, the expected cell count for cell (i, j) is

$$\mathbf{E}[N_{ij}] = \begin{cases} n_i(n_i - 1)/(n-1) & \text{if } i = j, \\ n_i n_j/(n-1) & \text{if } i \neq j, \end{cases} \tag{2}$$

and the variance $\mathbf{Var}[N_{ij}]$ is given in Ceyhan (2008).

The type III cell-specific test suggested by Ceyhan (2008) is

$$Z_{ij}^{III} = \frac{T_{ij}^{III}}{\sqrt{\mathbf{Var}\left[T_{ij}^{III}\right]}}, \tag{3}$$

where

$$T_{ij}^{III} = \begin{cases} N_{ii} - \frac{(n_i-1)}{(n-1)} C_i & \text{if } i = j, \\ N_{ij} - \frac{n_i}{(n-1)} C_j & \text{if } i \neq j. \end{cases} \tag{4}$$

The explicit forms of expectation and variance of T_{ij}^{III} are provided in Ceyhan (2008).

In the multi-class case with k classes, Dixon (2002a) suggests the following quadratic form combining the k^2 cell-specific tests and obtains the overall segregation test:

$$C_D = (\mathbf{N} - \mathbf{E}[\mathbf{N}])' \Sigma_D^- (\mathbf{N} - \mathbf{E}[\mathbf{N}]) \tag{5}$$

where $\mathbf{N}$ is the $k^2 \times 1$ vector of k rows of the NNCT concatenated row-wise, $\mathbf{E}[\mathbf{N}]$ is the vector of $\mathbf{E}[N_{ij}], \Sigma_D$ is the $k^2 \times k^2$ variance–covariance matrix for the cell count vector $\mathbf{N}$ with diagonal entries equal to $\mathbf{Var}[N_{ij}]$ and off-diagonal entries being $\mathbf{Cov}[N_{ij}, N_{kl}]$ for $(i, j) \neq (k, l)$. The explicit forms of the variance and covariance terms are provided in Dixon (2002a). Also, Σ_D^- is a generalized inverse of Σ_D (Searle 2006) and $'$ stands for the transpose of a vector or matrix. Then under RL, C_D asymptotically has a $\chi^2_{k(k-1)}$ distribution.

When we combine the type III cell-specific tests, we obtain type III overall test as follows. Let $\mathbf{T^{III}}$ be the vector of k^2 T^{III}_{ij} values, i.e.,

$$\mathbf{T^{III}} = \left(T^{III}_{11}, T^{III}_{12}, \ldots, T^{III}_{1k}, T^{III}_{21}, T^{III}_{22}, \ldots, T^{III}_{2k}, \ldots, T^{III}_{kk}\right)',$$

and let $\mathbf{E}\left[\mathbf{T^{III}}\right]$ be the vector of $\mathbf{E}\left[T^{III}_{ij}\right]$ values. Note that $\mathbf{E}\left[\mathbf{T^{III}}\right] = \mathbf{0}$. As the type III overall segregation test, we use the following quadratic form:

$$C_{III} = \left(\mathbf{T^{III}}\right)' \Sigma^-_{III} \left(\mathbf{T^{III}}\right) \tag{6}$$

where Σ_{III} is the $k^2 \times k^2$ variance–covariance matrix of $\mathbf{T^{III}}$.

Under RL, the explicit forms of the variance–covariance matrix for Dixon's and type III overall tests are provided in Ceyhan (2008). Furthermore, under RL, C_{III} asymptotically has a $\chi^2_{(k-1)^2}$ distribution.

3 Parameterizations of the alternative patterns

3.1 Desirable properties of an association/segregation pattern

In a two-class setting with classes 1 and 2, let an association (a segregation) alternative be parameterized by $\nu_a > 0$ ($\nu_s > 0$), where association (segregation) gets stronger as ν_a (ν_s) gets larger and $\nu_a = \nu_s = 0$ corresponds to the null hypothesis. In particular, ν_a could be the excess probability of a class 2 point being a NN to a class 1 point than expected under H_o, so with increasing ν_a, NNs of class 1 points would be more and more likely to be from class 2, which would imply stronger association of class 2 with class 1. Similarly, ν_s could be the excess probability of a class 1 point being a NN to a class 1 point than expected under H_o, so that with increasing ν_s, NNs of class 1 points would be more and more likely to be from class 1, which would imply stronger segregation of class 1 from class 2. Although any increasing function of the NN probabilities would parameterize these alternatives equivalently, we will consider ν_a and ν_s as the probabilities described above henceforth.

Let T_{n_1,n_2} be a consistent statistic used to test spatial patterns of segregation/association against H_o. Under a sensible association/segregation alternative, we would observe the following properties:

(P1) As ν_a (ν_s) increases, the power estimate of the test statistic, T_{n_1,n_2}, also increases.

(P2) Under an association (a segregation) alternative, level of association (segregation) is independent of the relative abundance of classes. That is, association (segregation) is not confounded by the differences between the class sizes.

(P3) Under an association (a segregation) alternative, the power estimate of the test statistic, T_{n_1,n_2}, would increase as both n_1, n_2 with $n_1 \approx n_2$ increase.

(P4) Under an association (a segregation) alternative with one class being of fixed size, say n_1 is fixed, the power estimate of the test statistic, T_{n_1,n_2}, increases as n_2 increases.

(P5) Under an association (a segregation) alternative, the power estimate of the test statistic, T_{n_1,n_2}, increases as the total sample size, $n = n_1 + n_2$, increases.

In particular, NNCT-tests are consistent for testing spatial patterns (see Ceyhan 2008). Hence, an association (a segregation) pattern should enjoy the above properties with respect to NNCT-tests. When checking the properties of an alternative pattern, we calculate the power estimates using the asymptotic critical values based on the standard normal approximation for the cell-specific tests and the corresponding χ^2-distributions for the overall tests. When the asymptotic approximations fail, Monte Carlo randomized versions of the tests should be employed.

In a multi-class setting, the null pattern, segregation and association can be characterized by the multi-class extensions of the distributions of NN distances and distances between randomly selected points and points from the class of interest i.e., "random point"–"class point" distance, which is also referred to as "point-event" distance in literature (Dixon 2002b). Let X_i be the distance from a randomly selected point to the nearest point from class i and $F_i(x)$ be the corresponding cdf. Also let W_{ij} be the distance from a class i point to the nearest class j point, and $G_{ij}(w)$ be its cdf. If the process of class i is independent of the process of class j, then we have $F_i(x) = G_{ji}(x)$ and $F_j(x) = G_{ij}(x)$ and X_i and X_j are independent (Diggle and Cox 1983; Goodall 1965). Note that the above equalities are not equivalent (i.e., one does not necessarily imply the other) (see, e.g., Goodall 1965). If $F_i(x) = F_j(x)$, then the corresponding independence structure would imply the null case of NN probabilities being proportional to class frequencies. Furthermore, in the two-class setting, if $G_{11}(x) > G_{12}(x)$, then W_{11} is stochastically smaller than W_{12}. Hence it is more likely for a class 1 point to be a NN of a class 1 point, which implies segregation of class 1 from class 2. Similarly, if $G_{22}(x) > G_{21}(x)$, we have segregation of class 2 from class 1. On the other hand, if $G_{12}(x) > G_{11}(x)$, then class 1 points are more likely to be NN of class 2 points, so class 1 is associated with class 2. Likewise, if

$G_{21}(x) > G_{22}(x)$, then class 2 tends to be associated with class 1.

Remark 3.1 The properties introduced in this section are useful as guiding principles to obtain "robust" (to differences in relative abundances) and "consistent" (the pattern does not change, but maybe becomes more precise as the class sizes increase) segregation/association patterns. When a pattern is found to follow these properties, then we can assess current or new methodology using generated samples from them. Otherwise, e.g., the empirical size or power performance of the methods can be confounded by the defective properties of the pattern generated. □

3.2 A parametrization of the segregation patterns

Segregation pattern is relatively easy to parameterize. For example, Ceyhan (2008) parameterizes a segregation alternative by generating $X_i \overset{iid}{\sim} \mathcal{U}(S_1)$ and $Y_j \overset{iid}{\sim} \mathcal{U}(S_2)$ where $S_1 = (0, 1-s) \times (0, 1-s)$ and $S_2 = (s, 1) \times (s, 1)$ for $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$ and $s \in (0,1), \mathcal{U}(S)$ stands for uniform distribution on S, and "iid" stands for "independent identically distributed as". Hence the segregation alternative is

$$H_S : s > 0. \tag{7}$$

Here "$s = 0$" case corresponds to the CSR independence pattern. Notice that, the level of segregation increases as s increases; that is, $\nu_s(s)$ gets larger as s increases, and so does the power of the (consistent) tests. Hence P1 is satisfied with this segregation parameterization. Also properties P2–P5 hold, as shown in Ceyhan (2008). For example, the empirical power estimates under H_S: $s = 1/6$ for various sample size combinations are presented in Fig. 1. For each sample size combination, 10,000 Monte Carlo replications are performed. The empirical power estimates for each sample size combination are joined by solid or dotted lines for better visualization (which is adopted throughout the article). Notice that the properties P3–P5 are empirically verified for this type of segregation. In particular, considering sample sizes (10,10), (30,30), (50,50), and (100,100), we observe that as $n_1 = n_2 = n$ increases, the power estimates increase as well, hence P3 follows. Considering sample sizes (10,10), (10,30), (10,50), or (10,30), (30,30) or (30,30), (30,50), or (30,50), (50,50) or (50,100), (100,100) we observe that P4 follows. Also, in the presented order of sample size combinations, total sample size increases except from (10,50) to (30,30), and power estimates increase as n increases, hence P5 follows. Notice also that in Fig. 1, we only present the cell-specific tests for cells (1,1) and (2,2) (and we will stick to this choice in the two-class case henceforth), because the cell-specific tests for the other cells essentially carry the same information (but with opposite signs). That is, for Dixon's cell-specific test, we have $Z^D_{i1} = -Z^D_{i2}$ for $i = 1{,}2$ and for type III cell-specific test, we have $Z^{III}_{1j} = -Z^{III}_{2j}$ for $j = 1{,}2$. Notice also that type III tests tend to have higher power compared to Dixon's test under this type of segregation.

This type of segregation pattern can be generalized as follows. Let F_i be the distribution for class i and $S(F_i)$ be the corresponding support set for $i = 1, 2, \ldots, k$ with $k \geq 2$. For simplicity, consider $k = 2$ and supports being on the real plane, $\mathbb{R}^2$. Clearly, if $S(F_1)$ and $S(F_2)$ are disjoint a.s., then classes 1 and 2 are segregated. Furthermore, let $S_{1>2} = \{(x,y) \in \mathbb{R}^2 : F_1(x,y) > F_2(x,y)\}$ and $S_{2>1} = \{(x,y) \in \mathbb{R}^2 : F_2(x,y) > F_1(x,y)\}$ and let λ be the Lebesgue measure in $\mathbb{R}^2$. Then there is segregation between classes 1 and 2, if $\lambda(S_{1>2})$ or $\lambda(S_{2>1})$ is positive. Then if $\lambda(S_{i>j}) > 0$ and $\lambda(S_{j>i}) = 0$ for $i \neq j$, then class i is segregated from class j; if $\lambda(S_{i>j}) > 0$ and $\lambda(S_{j>i}) > 0$ for $i \neq j$, then both classes i and j are segregated from each other. Furthermore, letting X and Y are random variables from F_1 and F_2, respectively, if $P(X \in S_{1>2}) > P(Y \in S_{2>1})$, then class 1 is more segregated than class 2, and switching 1 and 2, we get the reverse relationship where the probabilities are with respect to the corresponding distributions. In particular, if, e.g., in the null case the classes have the same distribution $F_1 = F_2 = F$ with the same support, and when

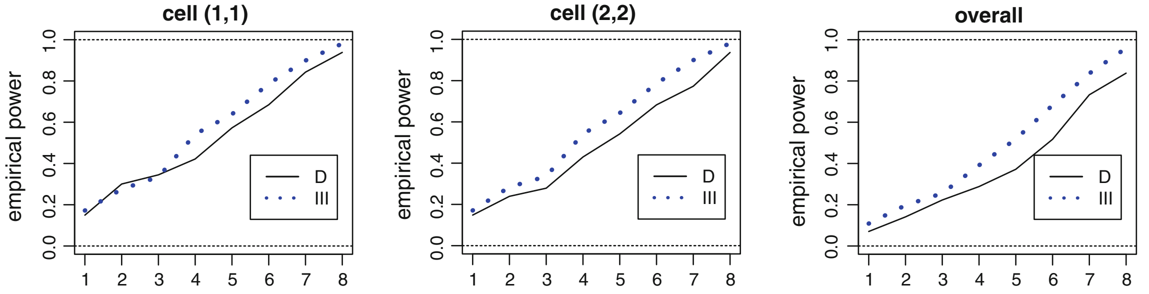

Fig. 1 The empirical power estimates for the cell-specific tests (*left* and *middle*) and the overall tests (*right*) under the segregation alternative H_S:$s = 1/6$ in the two-class case. The *horizontal axis labels* are for sample size combinations (n_1,n_2) with 1 = (10,10), 2 = (10,30), 3 =(10,50), 4 =(30,30), 5 = (30,50), 6 = (50,50), 7 = (50,100), 8 = (100,100). The *legend labeling*: *D* Dixon's, and *III* type III cell-specific or overall tests

we translate the support of one of the classes d units with $d \neq 0$ in any direction, then the resulting distribution would have the case that $\lambda(S_{1>2}) > 0$ and $\lambda(S_{2>1}) > 0$, so the classes would be segregated. In particular, the larger the value of $|d|$, the stronger the resulting segregation. Such a segregation parametrization satisfies all properties, P1–P5, for a consistent test.

In practice, segregation patterns might result (among others) from niche specificity (with different niches for different species), or inhibition of one species by another. To generate robust and consistent (in the sense of Remark 3.1) segregation alternatives, we may start with different supports or niches satisfying the above properties. On the other hand, if a given multi-class data exhibits segregation, the above properties together with support estimation might help us understand the generative pattern behind the data in question. This would be valuable for estimation and inferential purposes as well.

3.3 Various parameterizations of the association pattern

Association is not as easy to parameterize as the segregation pattern. For example, in literature Ceyhan (2008) parameterizes a type of association, which will be referred to as "Type C Association" in this article. We also propose two new parameterizations.

3.3.1 *Type C association*

Type C association is parameterized as follows (see also Ceyhan 2008): First generate $X_i \overset{iid}{\sim} \mathcal{U}((0,1)\times(0,1))$ for $i = 1, 2, \ldots, n_1$. Then generate Y_j associated with X's for $j = 1, 2, \ldots, n_2$ as follows. For each j, select an i randomly, and set $Y_j = X_i + R_j(\cos T_j, \sin T_j)'$ where $R_j \overset{iid}{\sim} \mathcal{U}(0, r)$ with $r \in (0, 1)$ and $T_j \overset{iid}{\sim} \mathcal{U}(0, 2\pi)$. Then the association alternatives are as

$$H_A : r \in (0, r_0) \tag{8}$$

for r_0 sufficiently small that $\nu_a(r)$ would be larger than expected. In Ceyhan (2008), $r = 1/4$, $1/7$, and $1/10$ are considered. By construction, the association of Y points with X points is stronger, compared to the association of X points with Y points.

Notice that association gets stronger as r decreases; and as r decreases, $\nu_a(r)$ gets larger. So type C association satisfies P1 and also Monte Carlo simulations suggest that P3 holds. However, it is shown empirically that P3, P4 and P5 fail for this association type. In particular, association is shown to be confounded by the differences in relative abundances of the classes (Ceyhan 2008).

In the two-class setting, we consider the following cases for the type C association:

$$H_A^I : r = 1/4, \quad H_A^{II} : r = 1/7, \quad H_A^{III} : r = 1/10,$$
$$H_A^{IV} : r = 1/20, \quad H_A^{V} : r = \frac{1}{2\sqrt{n_1}}, \quad H_A^{VI} : r = \frac{1}{4\sqrt{n_1}},$$
$$\text{and} \quad H_A^{VII} : r = \frac{1}{2\sqrt{n}}. \tag{9}$$

Notice that association gets stronger as r decreases for fixed n_1 and n_2; that is, association gets stronger from H_A^I to H_A^{IV} and from H_A^V to H_A^{VI}. The same happens from H_A^V to H_A^{VII} provided $n_2 > 3n_1$. Under each of $H_A^I - H_A^{IV}$, r is fixed, and as r decreases the association parameter $\nu_a(r)$ increases. Under each of H_A^V and H_A^{VI}, r depends on n_1 and under H_A^{VII}, r depends on n. So under each of H_A^V and H_A^{VI}, if n_1 increases, r decreases (so the level of association depends on n_1) and under H_A^{VII} if n increases, r decreases (so the level of association depends on the total sample size, n).

The alternatives $H_A^V - H_A^{VII}$ are motivated from the expected distance between points from HPP. In particular, let D be the distance from a randomly chosen point to the nearest other point in a HPP with intensity ρ. Then $\mathbf{E}[D] = 1/(2\sqrt{\rho})$ and $\mathbf{Var}[D] = (4-\pi)/(4\pi\rho)$ (Dixon 2002b). Then the choice r could be determined based on these two quantities. For example, in our case, under CSR independence intensity of class 1 with n_1 points would be $\widehat{\rho}_1 = n_1$, since area of the unit square is 1. Hence we have set $r = 1/(2\sqrt{n_1})$ and $r = 1/(4\sqrt{n_1})$ for H_A^V and H_A^{VI}, respectively. That is, under H_A^V, the displacements of Y_j around X_i would be limited by the average distance between X points under H_o, and under H_A^{VI}, the displacements of Y_j around X_i would be limited to half of the average distance between X points under H_o. Under CSR independence, combining both classes 1 and 2, we have n many points from the same HPP (conditional on n). Hence $\widehat{\rho}_T = 1/(2\sqrt{n})$, and we set $r = 1/(2\sqrt{n})$ in H_A^{VII}. That is, under H_A^{VII}, the displacements of Y_j around X_i would be limited to the average distance (under H_o) between X and Y points combined. In general, one can design type C association alternatives with $r = 1/(k\sqrt{\widehat{\rho}_1})$ or $r = 1/(k\sqrt{\widehat{\rho}_T})$ with $k \geq 2$, so that on the average displacement of Y points would be closer to X points compared to other X points. Also, one can set $r = 1/(2\sqrt{\widehat{\rho}_1}) - k\sqrt{(4-\pi)(4\pi\widehat{\rho})}$ with $0 < k$ and $k < \sqrt{\pi/(4-\pi)}$, since larger k values would imply $r < 0$, an impossibility. Again, on the average, displacement of Y points would be closer to X points compared to other X points.

In Fig. 2, we present the average test statistics for the cell-specific tests and the overall tests and the corresponding

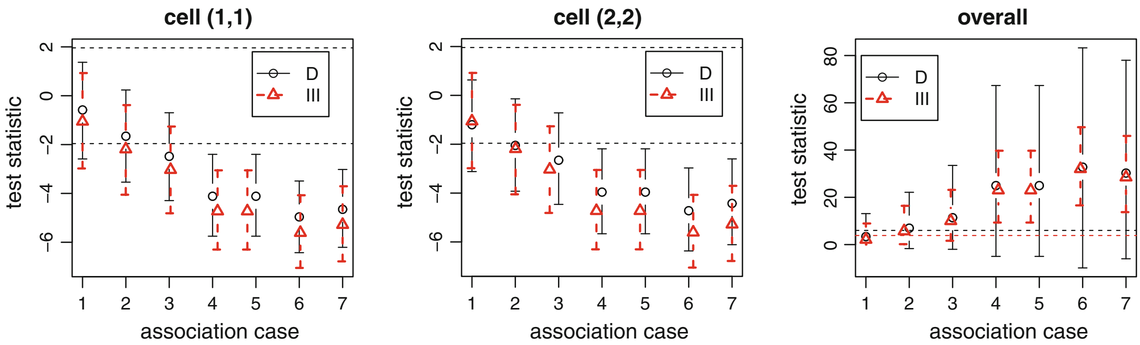

Fig. 2 The average test statistic values and the 95 % empirical confidence intervals for the cell-specific and overall NNCT-tests under the type C association alternatives $H_A^I - H_A^{VII}$ (labeled with the corresponding arabic numerals 1–7 in the *horizontal axis*) in the two-class case and *legend labeling* is as in Fig. 1. The *dashed horizontal lines* are the critical values at $\alpha = 0.025$ of the standard normal distribution (i.e., -1.96 and 1.96) for the cell-specific tests, and $\alpha = 0.05$ critical values of the χ^2 distribution with 1 and 2 degrees of freedom (i.e., 3.84 and 5.99, respectively) for the overall tests. The power estimates are jittered for better visualization

95 % empirical confidence intervals (CIs) based on 10,000 Monte Carlo replications for each case with $n_1 = n_2 = 100$. The $100(1 - \alpha)$ % empirical CIs are computed as follows: Out of the obtained values of each test statistic, we find the 100 $(\alpha/2)$th and $100(1 - \alpha/2)$th percentiles for the end points of the CIs. In our Monte Carlo setup, this corresponds to finding 250th and 9,750th values of the test statistics. Notice that the most severe association occurs under H_A^{VI} and H_A^{VII} with former being slightly stronger. For the cell-specific tests, type III test seems to be more sensitive, while for the overall tests Dixon's test is slightly more sensitive against association, but type III test has considerably less variation.

The sensitivity of the type C association to the balanced but increasing sample sizes is investigated empirically under H_A:$r = 1/10$ and $r = 1/(2\sqrt{n})$ with $n_1 = n_2 = n = 10, 20, \ldots, 100$ in Fig. 3. Notice that the power estimates increase with n, hence indicate that type C association satisfies property P3. Notice also that $H_A : r = 1/(2\sqrt{n})$ yields higher power estimates, since as n increases r decreases and also precision increases, hence association gets stronger since r decreases, and power increases since both association gets stronger and sample sizes increase. So we recommend type C association with $r = 1/(k\sqrt{\hat{\rho}})$ with appropriate choice of $k \geq 2$ to get an association pattern more robust to differences in relative abundances. Under these alternatives, properties P3 and P5 can not be evaluated for fixed r, since r depends on the sample sizes. However P1 can be verified with these alternatives, since ν_a increases (or r decreases) as n increases. On the other hand, P2–P5 can be evaluated under a specific alternative formulation as $H_A : r = 1/(k\hat{\rho})$. However, it should be kept in mind that the association parameter is not fixed, but only the structure of the alternative is fixed.

The sensitivity of type C association to the differences in relative abundances is investigated under H_A:$r = 1/10$ and $r = 1/(2\sqrt{n})$ with $n_1 = 10$ and $n_2 = 10, 20, \ldots, 100$. In Fig. 4, we only present the power estimates under H_A: $r = 1/10$, as the power trend under $H_A : r = 1/(2\sqrt{n})$ is very similar. Notice that as n_2 increases (i.e., the difference in relative abundance increases) the power estimates for cell-specific test for cell (2,2) and the overall test have a concave down trend (with increasing first, reaching a peak, and then decreasing). And the performance of Dixon's cell-specific test for cell (1,1) is severely affected by the differences in sample sizes. Hence, this figure suggests that properties P2, P4 and P5 fail for type C association. However, for $n_1 = 10$ and n_2 from 10 to 100, it is very likely that N_{11} gets smaller than the required cell counts in a NNCT for asymptotic approximation to be appropriate. In particular, it is recommended that cell counts should be at least 10 for Dixon's test and at least 5 for type III cell-specific tests (Ceyhan 2008). To avoid this confounding effect of asymptotic approximation, we try larger samples with $n_1 = 30$ and $n_2 = 30, 40, \ldots, 100$ for H_A:$r = 1/10$ and $r = 1/(2\sqrt{n})$. The corresponding power estimates are presented in Fig. 5. Notice that under H_A: $r = 1/10$, P2 and P4 seem to fail with Dixon's tests, but P2–P5 seem to hold with the type III tests. Hence type C association sometimes fails to satisfy properties P2, P4 and P5.

The sensitivity of type C association to the association parameter r is investigated and the test statistics together with 95 % empirical CIs for the NNCT-tests are plotted in Fig. 6. Notice that depending on the values of n_1 and n_2 and r, this parametrization yields association, H_o, or segregation. In particular, when $n_1 = n_2$ is large, the pattern does not deviate significantly from H_o for $r \approx 0.4$, and the pattern belongs to association for $r \leq 0.3$ and to segregation for $r \geq 0.5$. On the other hand, when $n_1 = 30$ and $n_2 = 100$, the pattern does not deviate significantly from H_o for $r \approx 0.6$ or 0.7, and the pattern implies association for $r \leq 0.5$ and mild segregation for $r \geq 0.8$. Hence we

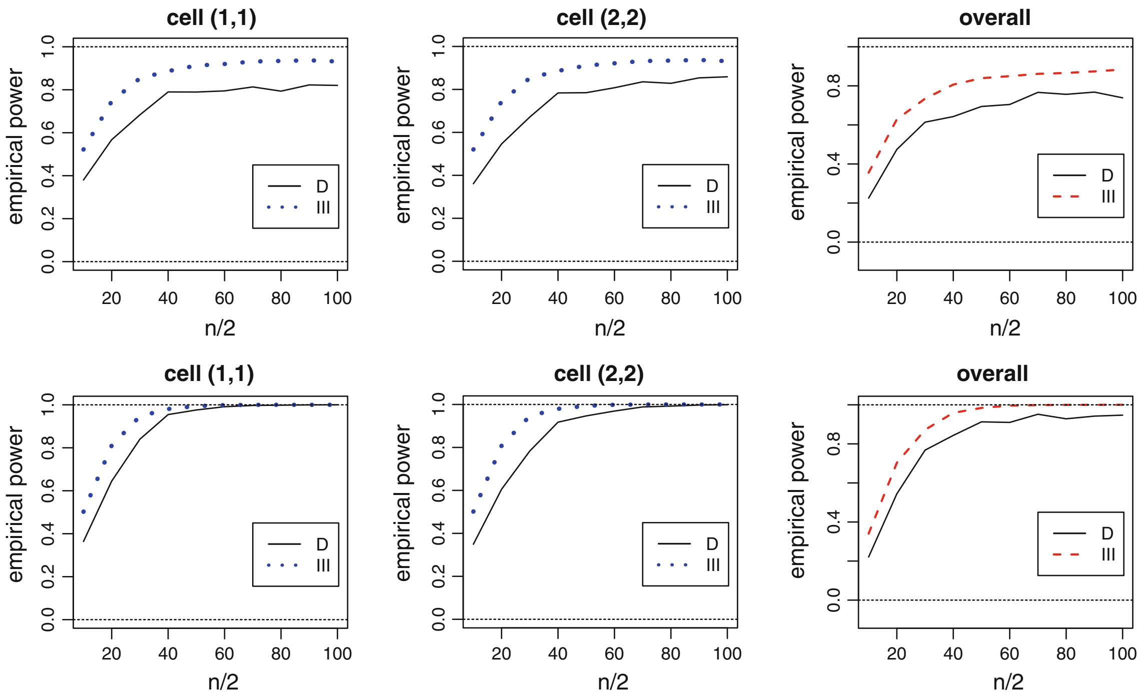

Fig. 3 The empirical power estimates for the NNCT-tests under the type C association alternatives $H_A^{III}: r = 1/10$ (*top row*) and $H_A^{VII}: r = 1/(2\sqrt{n})$ (*bottom row*) in the two-class case as a function of $n_1 = n_2 = n/2 = 10, 20, \ldots, 100$. The *horizontal axis labels* and *legend labeling* are as in Fig. 1

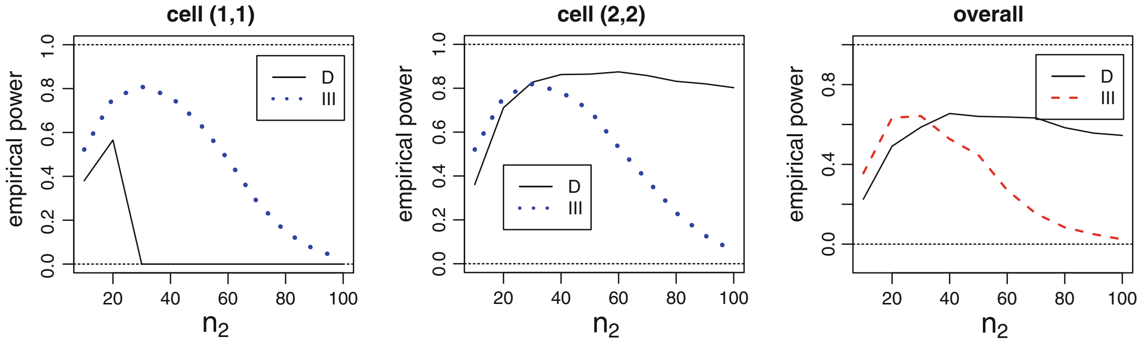

Fig. 4 The empirical power estimates for the NNCT-tests under the type C association alternatives in the two-class case under H_A:$r = 1/10$ with $n_1 = 10$ and $n_2 = 10, 20, \ldots, 100$. The *legend labeling* is as in Fig. 1

recommend to use $r \leq 0.25$ times length of the shorter edge of a rectangular study region, or use $r = 1/(k\sqrt{\hat{\rho}})$ with the choice of k implying $r \leq 0.25$ times length of the shorter edge to have alternative patterns more robust to differences in sample sizes.

Under type C association, although R_j and θ_j are generated uniformly in their respective ranges, Y_j are not uniformly distributed in the circles centered at X_i with radius r_0. To see this, without loss of generality, assume a given $X_i = (0,0)$. In polar coordinates, we have $R_i \sim \mathcal{U}(0, r_0)$ and $\theta_i \sim \mathcal{U}(0, 2\pi)$. Then probability density functions (pdfs) of R_i and θ_i are $f_R(r) = \frac{1}{r_0}$ and $f_\theta(\theta) = \frac{1}{2\pi}$, respectively. Hence by independence, the joint density of (R, θ) is

$$f_{R,\theta}(r, \theta) = \frac{1}{2\pi r_0} \quad \text{for } r \in [0, r_0] \text{ and } \theta \in [0, 2\pi].$$

Given $Y_j = (t, v)$ with $t = r\cos\theta$ and $v = r\sin\theta$, we have $r = \sqrt{t^2 + v^2}$ and $\theta = \arctan(v/t)$. Then Jacobian is

$$J = \begin{vmatrix} \frac{\partial r}{\partial t} & \frac{\partial r}{\partial v} \\ \frac{\partial \theta}{\partial t} & \frac{\partial \theta}{\partial v} \end{vmatrix} = \frac{1}{\sqrt{t^2 + v^2}}.$$

Hence, $|J| = \frac{1}{\sqrt{t^2+v^2}}$. Therefore, joint pdf of (T, V) is

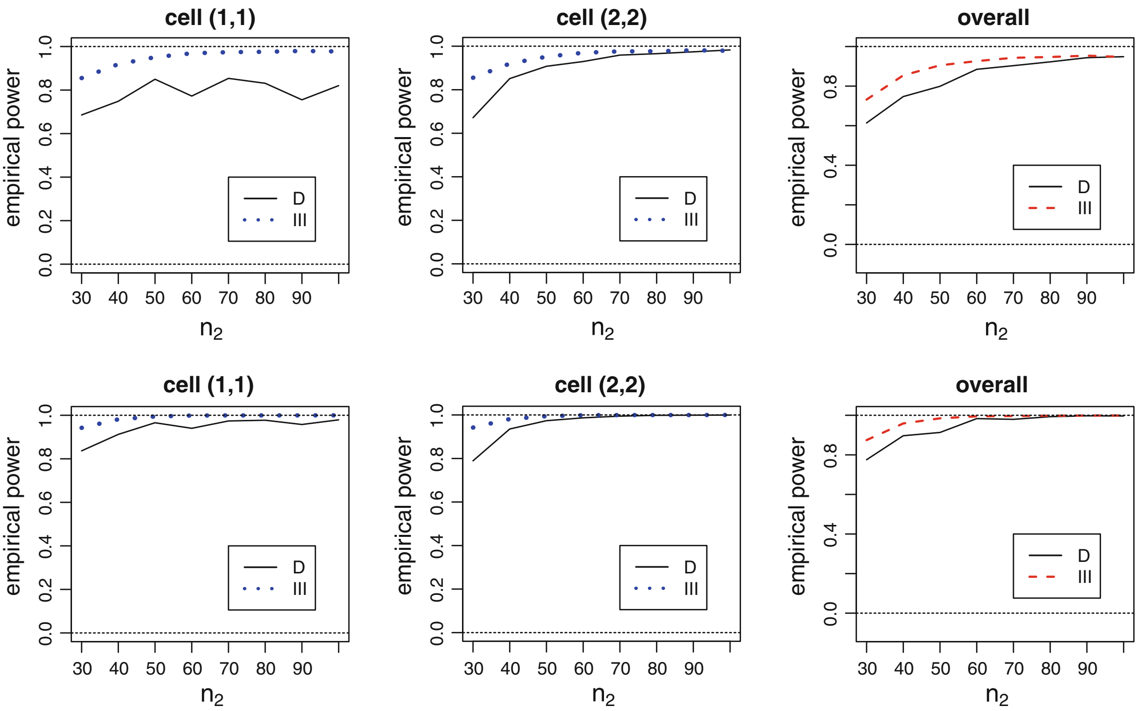

Fig. 5 The empirical power estimates for the NNCT-tests under the type C association alternatives $H_A{:}r = 1/10$ (*top row*) and $r = 1/(2\sqrt{n})$ (*bottom row*) in the two-class case with $n_1 = 30$ and $n_2 = 30, 40, \ldots, 100$. The *legend labeling* is as in Fig. 1

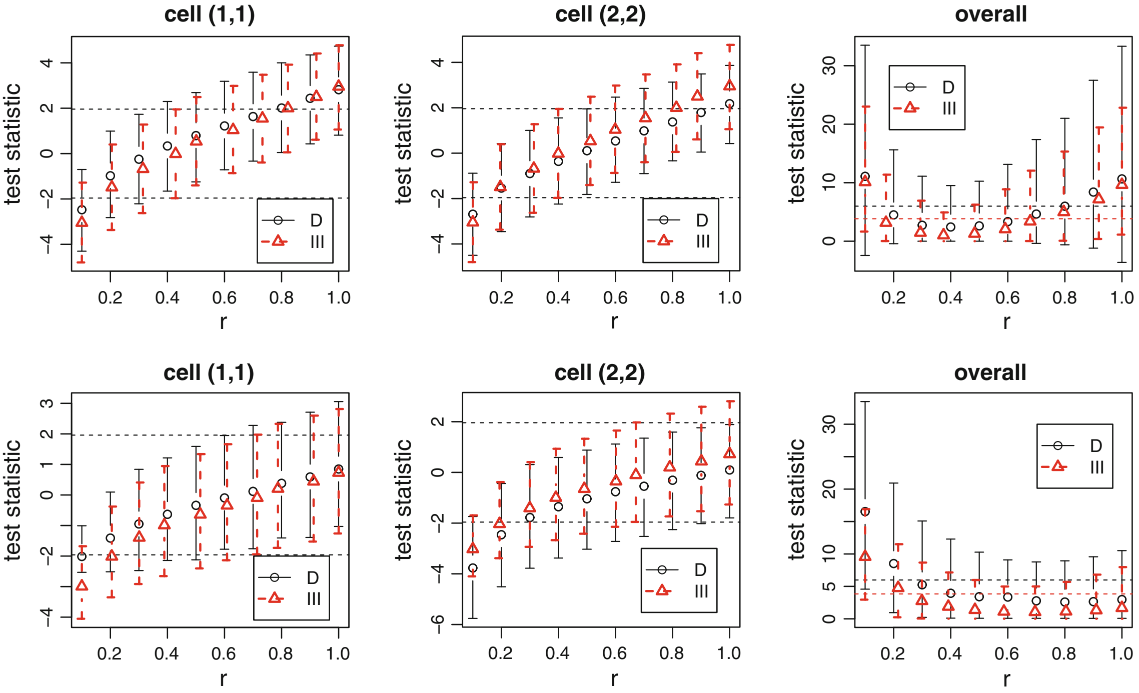

Fig. 6 The test statistics (means and 95 % empirical CIs) for the NNCT-tests under the type C association with $r = 0.1, 0.2, \ldots, 1.0$ for $n_1 = n_2 = 100$ (*top row*) and $n_1 = 30$ and $n_2 = 100$ (*bottom row*). The *dashed horizontal lines* for the cell-specific tests are critical z-scores for a two-sided test with $\alpha = 0.05$ (i.e., at -1.96 and 1.96), and the *dashed horizontal lines* for the overall tests are as in Fig. 2. The *legend labeling* is as in Fig. 1

$$f_{T,V}(t,v) = f_{R,\theta}\left(\sqrt{x^2+y^2}, \arctan(y/x)\right)|J|$$
$$= \frac{1}{2\pi r_0\sqrt{t^2+v^2}} \quad \text{for } 0 < t^2+v^2 \le r_0^2.$$

For Y_j to be uniform in the circle centered at 0 with radius r_0, $C(0, r_0)$, the joint pdf of (T, V) would have been $1/(\pi\ r_0^2)$.

3.3.2 Type U association

Since Y points are not uniform around randomly selected X points under type C association, we suggest another type of association pattern called *type U association* ("U" for uniform distribution) where Y points are generated uniformly around the X points. First generate $X_i \overset{iid}{\sim} \mathcal{U}((0,1)\times(0,1))$ for $i = 1, 2, \ldots, n_1$. Then generate Y_j associated with X's for $j = 1, 2, \ldots, n_2$ as follows. For each j, select an i randomly from $\{1, 2, \ldots, n_1\}$, and generate $U_j \overset{iid}{\sim} \mathcal{U}(0,1)$ and set $R_j = r_0\sqrt{U_j}$. Then set $Y_j = X_i + R_j\left(\cos T_j, \sin T_j\right)'$ where $T_j \overset{iid}{\sim} \mathcal{U}(0, 2\pi)$. Then the association alternatives are as

$$H_A : r \in (0, r_0) \tag{10}$$

for r_0 sufficiently small such that $\nu_a(r)$ would be larger than expected. Again, by construction, the association of Y points with X points is stronger, compared to the association of X points with Y points. Notice also that association gets stronger as r decreases whence $\nu_a(r)$ gets larger. So this type of association satisfies P1.

Under type U association, Y_j are uniformly distributed in the circles centered at X_i with radius r_0. To see this, without loss of generality, let $X_i = (0,0)$. Since $U_j \overset{iid}{\sim} \mathcal{U}(0,1)$, cumulative distribution function (cdf) of R_j is $F_R(r) = P(r_0\sqrt{U} \le r) = P(U_j \le r^2/r_0^2) = r^2/r_0^2$. Hence pdf of R_j is $f_R(r) = \frac{2r}{r_0^2}$ and $\theta_j \overset{iid}{\sim} U(0, 2\pi)$, and pdf of θ_i is $f_\theta(\theta) = \frac{1}{2\pi}$. By independence, the joint density of (R, θ) is

$$f_{R,\theta}(r,\theta) = \frac{r}{\pi r_0^2} \quad \text{for } r \in [0, r_0] \text{ and } \theta \in [0, 2\pi].$$

Switching from polar coordinates to Cartesian coordinates by letting $t = r\sin\theta$ and $v = r\cos\theta$, the Jacobian is $J = \frac{1}{\sqrt{t^2+v^2}}$. Therefore, joint pdf of (T, V) is

$$f_{T,V}(t,v) = f_{R,\theta}\left(\sqrt{x^2+y^2}, \arctan(y/x)\right)|J| = \frac{1}{\pi r_0^2}$$
$$\text{for } 0 < t^2+v^2 \le r_0^2.$$

Hence it follows that Y_j are uniform in $C(0,r_0)$ as claimed.

The sensitivity of the type U association to the balanced but increasing samples sizes is investigated empirically under H_A:$r = 1/10$ and $r = 1/(2\sqrt{n})$ with $n_1 = n_2 = n/2 = 10, 20, \ldots, 100$ and only the results for $r = 1/(2\sqrt{n})$ are presented in Fig. 7. Notice that the power estimates increase with n, which supports our assertion that type U association satisfies property P3.

The sensitivity of the type U association to the differences in relative abundances is investigated empirically with H_A:$r = 1/10$ and $r = 1/(2\sqrt{n})$ with $n_1 = 30$ and $n_2 = 30, 40, \ldots, 100$ in Fig. 8. Notice that under H_A:$r = 1/10$, properties P2, P4, and P5 seem to fail for both type III and Dixon's tests, however under $H_A : r = 1/(2\sqrt{n})$, these properties hold for type III tests.

The sensitivity of type U association to the association parameter, r, is investigated and the mean test statistics together with 95 % empirical CIs for the NNCT-test statistics are plotted in Fig. 9. Notice that depending on the values of n_1 and n_2 and r, this parametrization yields association, null, or segregation patterns. In particular, when $n_1 = n_2$ is large, the pattern does not deviate significantly from H_o for $r \approx 0.2$, and the pattern implies association for $r \le 0.1$ and segregation for $r \ge 0.4$. Moreover, when $n_1 = 30$ and $n_2 = 100$, we have a similar trend. For type U association, we recommend to use $r \le 0.10$ times length of the shorter edge of a rectangular study region, or use $r = 1/(k\sqrt{\rho})$ with the choice of k satisfying $r \le 0.10$ times length of the shorter edge to

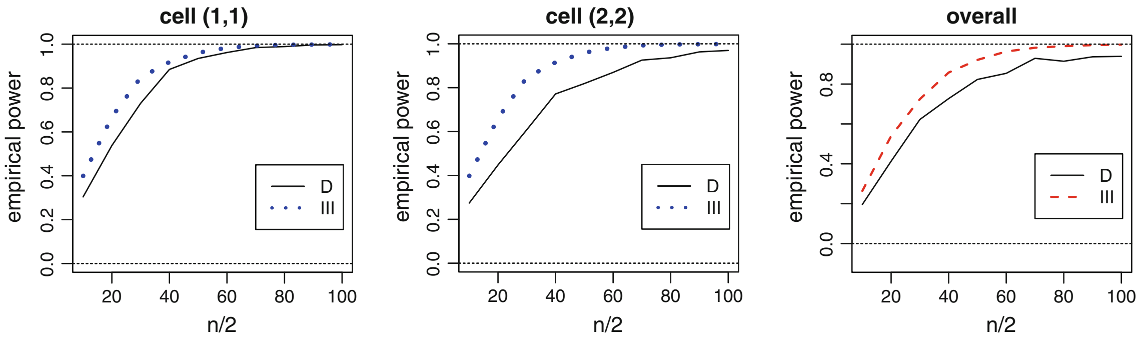

Fig. 7 The empirical power estimates for the NNCT-tests under the type U association alternatives $H_A : r = 1/(2\sqrt{n})$ with $n_1 = n_2 = n/2 = 10, 20, \ldots, 100$. The *horizontal axis labels* and *legend labeling* are as in Fig. 1

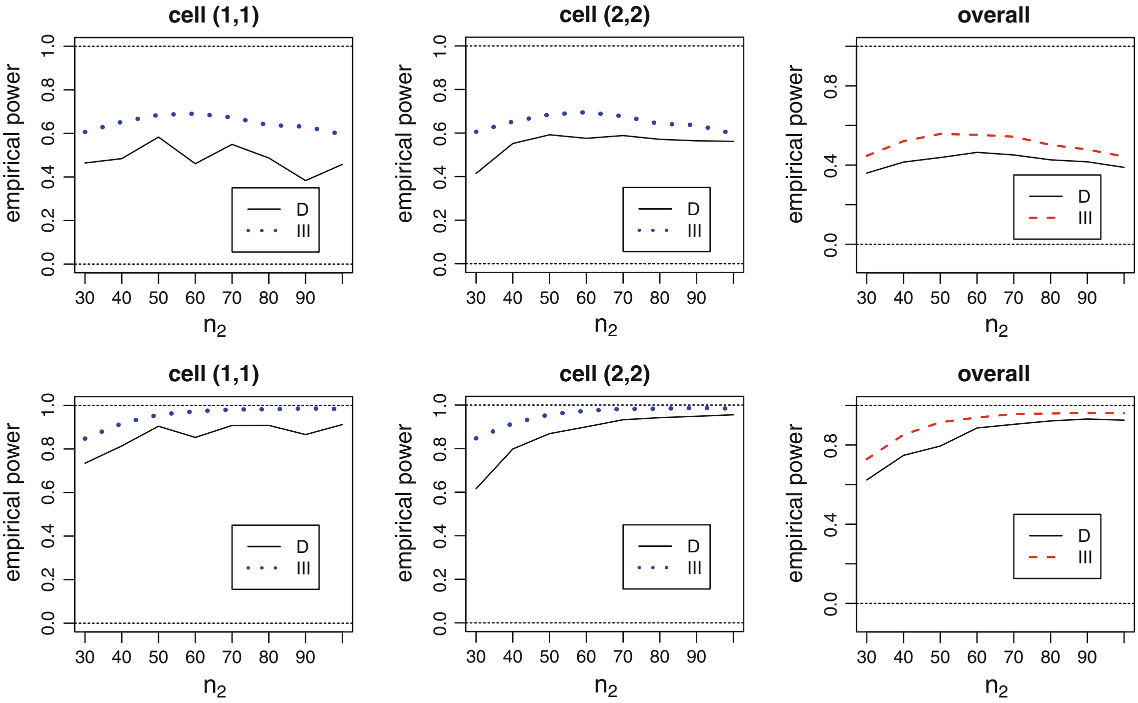

Fig. 8 The empirical power estimates for the NNCT-tests under the type U association alternatives H_A:$r = 1/10$ (*top row*) and $H_A : r = 1/(2\sqrt{n})$ (*bottom row*) with $n_1 = 30$ and $n_2 = 30, 40, \ldots, 100$. The *horizontal axis labels* and *legend labeling* are as in Fig. 1

have alternative patterns more robust to differences in sample sizes.

3.3.3 Type G association

We also introduce a type of association pattern, called *type G association* ("G" for Gaussian), where Y points are generated according to bivariate normal distribution around the X points. In particular, generate $X_i \overset{iid}{\sim} \mathcal{U}((0,1) \times (0,1))$ for $i = 1, 2, \ldots, n_1$. Then generate Y_j associated with X's for $j = 1, 2, \ldots, n_2$ as follows. For each j, select an i randomly from $\{1, 2, \ldots, n_1\}$, and generate $U_j \sim N(0,\sigma)$ and $V_j \sim N(0, \sigma)$, where $N(\mu, \sigma)$ stands for normal distribution with mean μ and standard deviation σ. Then set $Y_j = X_i + \left(U_j, V_j\right)'$. Then the association alternatives are as

$$H_A : \sigma \in (0, \sigma_0) \qquad (11)$$

for σ_0 sufficiently small such that $\nu_a(\sigma)$ would be larger than expected. Again, by construction, the association of Y points with X points is stronger, compared to the association of X points with Y points. Notice also that association gets stronger as σ decreases whence $\nu_a(\sigma)$ gets larger. So this type of association satisfies P1.

Under type G association, Y_j are distributed according to bivariate normal distribution with mean X_i and covariance matrix $\sigma^2 I_2$ where I_2 is the 2-by-2 identity matrix. In polar coordinates, to find the marginal distributions of R_j and θ_j, without loss of generality, we let $X_i = (0,0)$. Then the joint density of $Y_j = (T, V)$ is $f_{U,V}(t, v) = \frac{1}{2\pi\sigma^2}\exp(-\frac{t^2+v^2}{2\sigma^2})$. Making the change of variables, $t = r\cos\theta$ and $v = r\sin\theta$, we get the joint density of R, θ as

$$f_{R,\theta}(r, \theta) = \frac{r}{2\pi\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right)$$
$$\text{for } r \in (0, \infty) \text{ and } \theta \in [0, 2\pi).$$

So the marginal pdf of R is $f_R(r) = \frac{r}{\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right)$ for $r \geq 0$ and cdf of R is $F_R(r) = 1 - \exp\left(-\frac{r^2}{2\sigma^2}\right)$ for $r \geq 0$. On the other hand, the marginal pdf of θ is $f_\theta(\theta) = \frac{1}{2\pi}$, for $\theta \in [0, 2\pi)$ hence θ is uniformly distributed in (0,2 π).

The sensitivity of the type G association to the balanced but increasing sample sizes is investigated empirically with H_A:$\sigma = 1/10$ and $\sigma = 1/(2\sqrt{n})$ with $n_1 = n_2 = n/2 = 10, 20, \ldots, 100$ in Fig. 10. Notice that under H_A:$\sigma = 1/10$, the power estimates tend not to increase as n increases (in fact, it tends to decrease as n increases), hence P3 fails with both tests. However, under $H_A : \sigma = 1/(2\sqrt{n})$, P3 holds for both tests, and the trend in the power is more in line with P3 for type III tests.

The sensitivity of the type G association to the differences in relative abundances is investigated empirically with H_A:$\sigma = 1/10$ and $\sigma = 1/(2\sqrt{n})$ with $n_1 = 30$ and $n_2 = 30, 40, \ldots, 100$ in Fig. 11. Notice that under H_A:$\sigma = 1/10$, P2, P4, and P5 seem to fail; however under $H_A : \sigma = 1/(2\sqrt{n})$, these properties hold for type III tests.

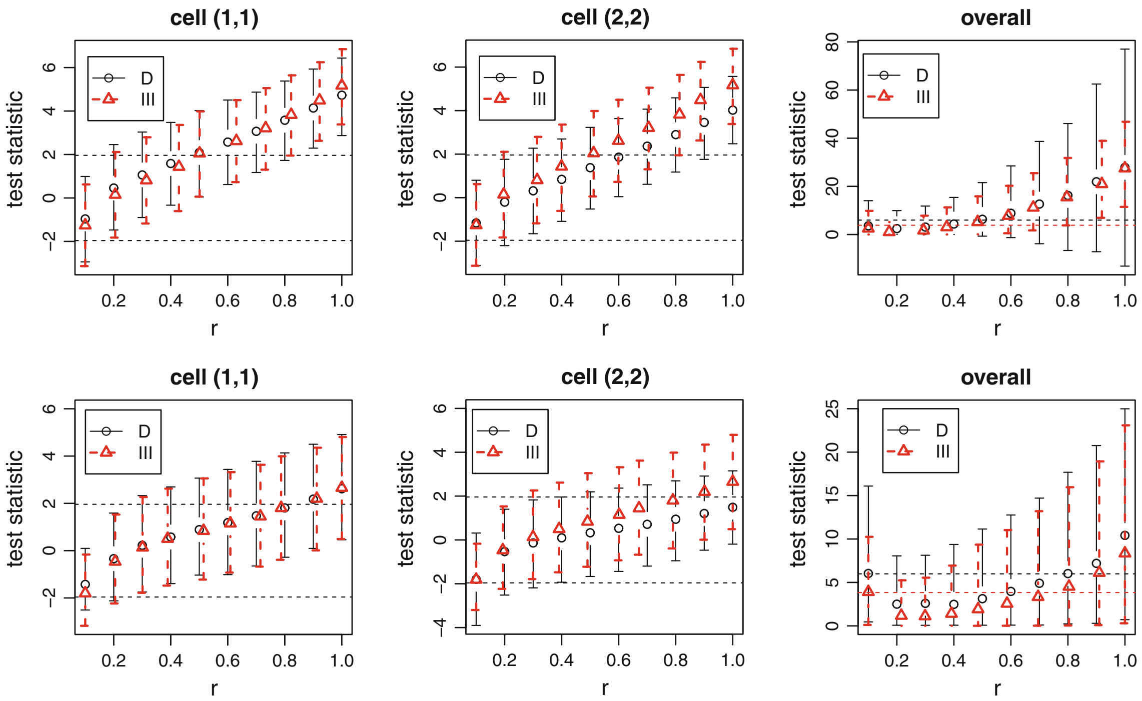

Fig. 9 The test statistics (means and 95 % empirical CIs) based on 10,000 Monte Carlo replications for the NNCT-tests under type U association with $r = 0.1, 0.2, \ldots, 1.0$ for $n_1 = n_2 = 100$ (*top row*) and $n_1 = 30$ and $n_2 = 100$ (*bottom row*). The *dashed horizontal lines* are as in Fig. 6 and the *legend labeling* is as in Fig. 1

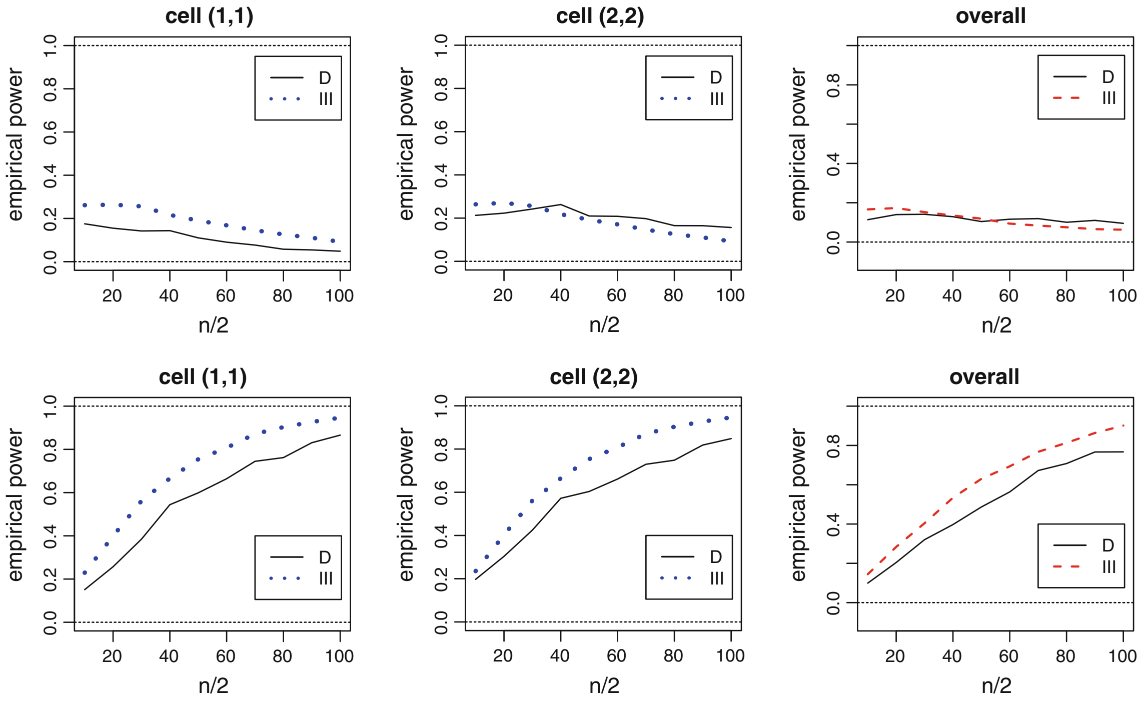

Fig. 10 The empirical power estimates for the NNCT-tests under the type G association alternatives H_A:σ = 1/10 (*top row*) and $H_A : \sigma = 1/(2\sqrt{n})$ (*bottom row*) with $n_1 = n_2 = n/2 = 10, 20, \ldots, 100$. The *horizontal axis labels* and *legend labeling* are as in Fig. 1

The sensitivity of type G association to the association parameter σ is investigated and the mean test statistics together with 95 % empirical CIs for the NNCT-test statistics are plotted in Fig. 12. Notice that depending on the values of n_1 and n_2 and σ, this parametrization yields association, null, or segregation patterns. In particular,

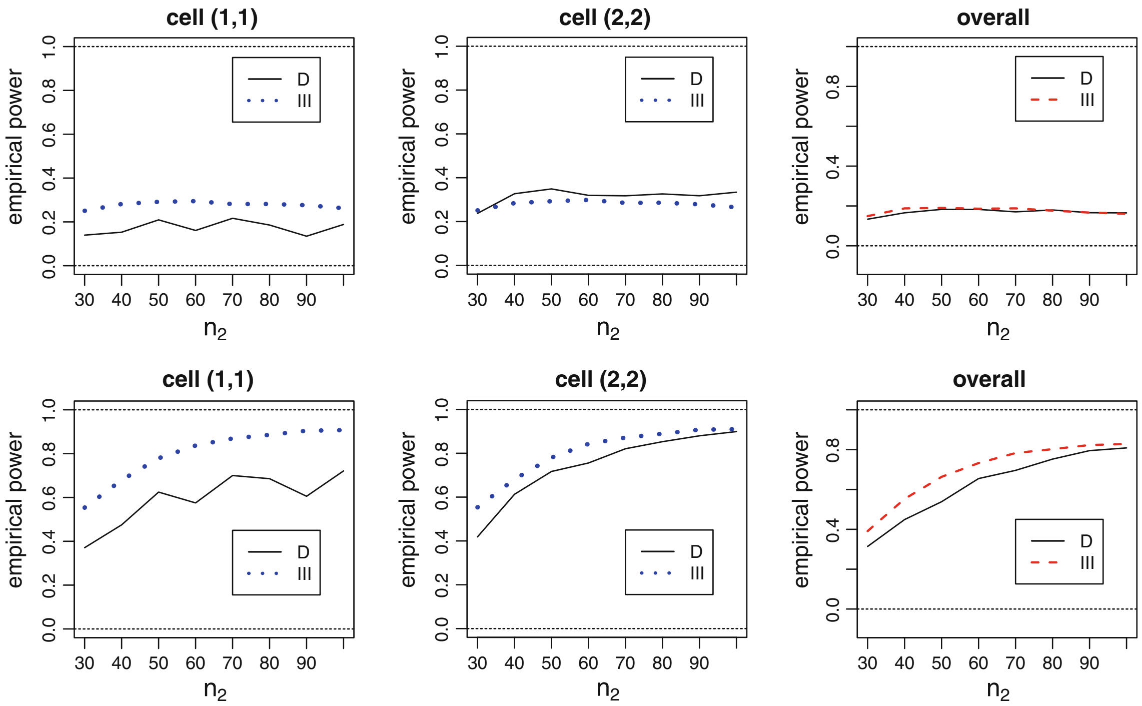

Fig. 11 The empirical power estimates for the NNCT-tests under the type G association alternatives H_A:$\sigma = 1/10$ (*top row*) and $H_A : \sigma = 1/(2\sqrt{n})$ (*bottom row*) with $n_1 = 30$ and $n_2 = 30, 40, \ldots, 100$. The *horizontal axis labels* and *legend labeling* are as in Fig. 1

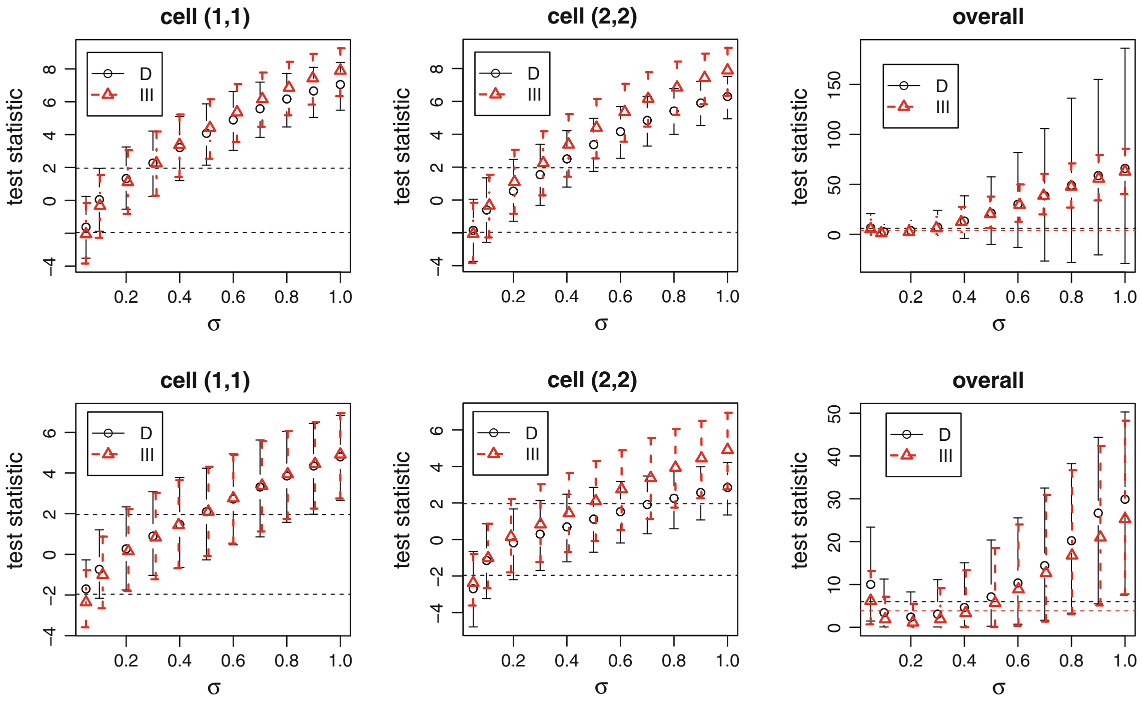

Fig. 12 The test statistics (means and 95 % empirical CIs) based on 10,000 Monte Carlo replications for the NNCT-tests under type G association with $\sigma = 0.05, 0.1, 0.2, \ldots, 1.0$ for $n_1 = n_2 = 100$ (*top row*) and $n_1 = 30$ and $n_2 = 100$ (*bottom row*). The *dashed horizontal lines* are as in Fig. 6 and the *legend labeling* is as in Fig. 1

when $n_1 = n_2$ is large, the pattern does not deviate significantly from H_o for $\sigma \approx 0.2$, and the pattern belongs to association for $\sigma \leq 0.1$ and to segregation for $\sigma \geq 0.4$. Moreover, when $n_1 = 30$ and $n_2 = 100$, we have a similar trend. For type G association, we recommend using $\sigma \leq 0.10$ times length of the shorter edge of a rectangular region, or using $\sigma = 1/(k\sqrt{\rho})$ with the choice of k satisfying $\sigma \leq 0.10$ times length of the shorter edge to have alternative patterns more robust to differences in sample sizes.

Remark 3.2 Among the above association pattern types, type C was defined previously in Ceyhan (2008) (in fact, only the alternatives $H_A^I - H_A^{III}$ in Eq. (10) were employed previously), but types G and U are newly introduced in this article. In practice, when a multi-class data set exhibits association, one can fit the best association pattern to it, and assess the goodness-of-fit of the proposed model to the real data. Furthermore, the fitted model provides a means to assess power and distribution of the test or estimated parameters empirically. Furthermore, to generate robust and consistent (in the sense of Remark 3.1) association patterns, these association types provide useful alternatives. □

4 Multi-class spatial patterns resulting from various point processes

For simplicity, we consider the two-class case. The extension of the discussion to three or more classes would be similar, possibly with more complicated interactions between the classes.

4.1 One class from HPP, the other from a clustering or regularity process

We also investigate which stochastic point process results in segregation or association with respect to a class of points from CSR. For this purpose, we generate X points from a HPP with intensity $\lambda = 100$. Hence, a realization of this process is from the CSR pattern. Then we generate Y points from the following spatial point processes (see Baddeley and Turner 2005 for more details):

(1) **Matérn Model I:** Y points are generated from the Matérn Model I inhibition process, denoted MatérnI(κ, r). First a HPP with intensity κ is generated. Then any point that lies closer than a distance r from another point is deleted. That is, pairs of points with distance less than r are removed. The Y points that remain constitute a realization of Matérn Model I process. We use $\kappa = 100$ and $r = 0.01, 0.02, \ldots, 0.05$ in our Monte Carlo simulations.

(2) **Matérn Model II:** Y points are generated from the Matérn Model II inhibition process, denoted MatérnII(κ, r). First a HPP with intensity κ is generated as in Matérn I case. Then each point is marked with an "arrival time", a number uniformly distributed in [0,1] independently of other points. Any point that lies closer than distance r from another point whose arrival time is smaller is deleted. Hence Matérn Model II has higher intensity for the same parameter values compared to Matérn Model I. We use $\kappa = 100$ and $r = 0.01, 0.02, \ldots, 0.10$ in our simulations.

(3) **Simple Sequential Inhibition (SSI) Process:** Y points are generated from the SSI process, denoted as *SSI*(r, n). In this case, we start with the unit square, and add points as follows. Each new point is generated independently uniformly in the unit square. If a new point has distance less than r units from an existing point, then it is not retained and another random point is generated. The algorithm ends when the desired number of points n is reached, or when the current point allocation does not change for a sufficiently large number of iterations. We use $n = 100$ and $r = 0.01, 0.02, \ldots, 0.10$ in our simulations.

(4) **Matérn Cluster Process (MCP), case 1:** In this process, Y points are generated from Matérn's cluster process in the unit square, denoted MCP(κ, μ, r). First "parent" points are generated from a Poisson process with intensity κ and then each parent is replaced by points independently uniformly generated inside the circle centered at the parent point with radius r, where number of these points follow a Poisson(μ) distribution. The parent points are not restricted to lie in the unit square. Here we take $\kappa = 5$, $\mu = 20$ and $r = 0.05$ and $r = 0.1, 0.2, \ldots, 1.0$ in our simulations.

(5) **MCP, case 2:** This is the same process as above, but we take $r = 0.1, \kappa = 1, 2, \ldots, 10$ and $\mu = \lfloor 100/\kappa \rfloor$, respectively, where $\lfloor x \rfloor$ stands for the floor of x. That is, we take $(\kappa, \mu) \in \{(1, 100), (2, 50), (3, 33)\ldots, (10, 10)\}$ in our simulations.

(6) **Thomas Cluster Process (TCP), case 1:** This clustering process, denoted TCP(κ, μ, σ), is a special case of Neyman–Scott process (NSP). In this process, "parent" points are independently uniformly generated from a Poisson process with intensity κ. Then each parent point is replaced by points whose positions being isotropic Gaussian displacements $N(0, \sigma^2\, I_2)$, where number of these points follow a Poisson(μ) distribution. Here we take $\kappa = 5$, $\mu = 20$ and $\sigma = 0.1, 0.2, \ldots, 1.0$ in our simulations.

(7) **TCP, case 2:** In this process, we take $\sigma = 0.1, \kappa = 1, 2, \ldots, 10$ and $\mu = \lfloor 100/\kappa \rfloor$, respectively. That is, we take $(\kappa, \mu) \in \{(1, 100), (2, 50), (3, 33)\ldots, (10, 10)\}$ in our simulations.

(8) **NSP, case 1:** In this case, Y points are generated from a NSP, denoted $NS(\kappa, r_0, cluster(r_1, r_2, \mu))$, where κ is the intensity of the Poisson process of cluster centers, r_0 is the maximum radius of a random cluster, and

cluster(r_1, r_2, μ) is a function generating random clusters. First, the "parent" points are generated from a Poisson process with intensity κ. Then each parent is replaced by a random cluster of points from the *cluster* function. In our case, we take the cluster function that generates N uniform points, with N having a Poisson(μ) distribution, in the circles centered at the parent points with radius r_2, and remove points whose distance to the parent points is less than r_1. That is, cluster function produces ring shaped clusters around the parent points. In our simulations, we take $\kappa = 5$ and $\mu = 27$ and $r_0 = r$, $r_2 = r$, and $r_1 = r/2$ where $r = 0.1, 0.2, \ldots, 1.0$ in our simulations. The choice of $\mu = 27$ is to have the expected number of generated Y points to be approximately 100.

(9) **NSP, case 2:** In this case, we take $r_0 = r$, $r_2 = r$, and $r_1 = r/2$ with $r = 0.1, \kappa = 1, 2, \ldots, 10$ and $\mu = \lfloor 135/\kappa \rfloor$, respectively. That is, we take $(\kappa, \mu) \in \{(1, 135), (2, 67), (3, 45) \ldots, (10, 13)\}$ in our simulations. The choice of $\mu = \lfloor 135/\kappa \rfloor$ is to have the expected number of generated Y points to be approximately 100.

Remark 4.1 In all the above cases, $N_{mc} = 10{,}000$ replications are performed for each parameter value. For example, in Matérn Model I, we generate 10,000 realizations of the process for each of $r = 0.01, 0.02, \ldots, 0.05$ with $\kappa = 100$. Furthermore, in order to remove the influence of the relative abundance differences, the parameters are chosen so that we have 100 points on the average for the X points in all cases, and 100 points for the Y points from the clustering processes.

Note that in the inhibition processes (i.e., cases (1)–(3), if we take $r = 0$, the processes boil down to a HPP. Moreover, in the usual NSP, if we take the cluster function that generates n uniform points in the circles centered at the parent points with radius r_0 and choose $\kappa = 5$ and $n = 20$ and $r_0 = 0.1, 0.2, \ldots, 1.0$, we obtain the MCP, case 1 above. If we take $r = 0.1, \kappa = 1, 2, \ldots, 10$ and $\mu = \lfloor 100/\kappa \rfloor$, respectively, we obtain the MCP, case 2 above. Furthermore, in the NSP cases in (8) above, if we take $r_1 = 0$ and $\mu = 100$, we also obtain MCP, case 1. □

With X points from a HPP and Y points generated from the above spatial point processes, we expect to have various patterns between X and Y points.

Processes (1)–(3) are inhibition processes, so the Y points generated by these processes would deviate towards regularity from CSR. Hence under these processes, X points are from the CSR process, and Y points are more regular than X points. As a result of this, with increasing level of regularity probability of NN of Y points being from Y points decreases, hence probability of NN of Y points being from X points increases. Thus we expect that as the level of inhibition or regularity increase, the level of association between X and Y points increases as well.

Processes (4)–(9) are cluster processes, so the Y points generated by these processes would deviate towards clustering from CSR. Hence under these processes, X points are from the CSR process, and Y points are more clustered than X points. With increasing level of clustering, probability of a NN of Y points being from Y points increases, hence we expect that as the level of clustering increases, the level of segregation between X and Y points increases as well.

The means and the 95 % empirical CIs around the means for the cell (2,2) statistics under the Processes (1)–(9) are presented in Fig. 13. Cell (1,1) statistics are very similar to the cell (2,2) statistics, hence are omitted. Furthermore, the overall test statistics do not provide the direction of deviation from CSR independence, hence are not presented either.

4.1.1 *Comparison of cell (2,2) statistics for each point process*

(1) In Matérn Model I, test statistics tend to be negative for $r \geq 0.02$, and get more negative as r increases which imply an increasing level of association between X and Y points. Hence as r increases, the level of regularity of Y points increases, and the level of association between X and Y points increases as well. However, in this setup there is only mild to moderate association, with strongest association occurring around $r \approx 0.05$.

(2) In Matérn Model II, test statistics tend to be negative which suggest association between X and Y points. Furthermore, they have a concave up trend, i.e., they decrease, reach a minimum, and then increase as r increases. The reason for such a trend could be that with increasing r, number of Y points tend to decrease. With fixed n_2, we expect that the level of association should increase (i.e., the test statistics should decrease) as r increases. The highest level of association occurs around $r = 0.06$.

(3) In SSI Process, test statistics have a similar trend as in Matérn Model II with the highest level of association occurring at $r = 0.08$.

(4) In MCP, case 1, the test statistics tend to be positive for $r \leq 0.5$ (hence are suggestive of segregation between X and Y points), and for $r > 0.5$, they are within the null region which suggest no significant deviation from H_o. So as r decreases, test statistics

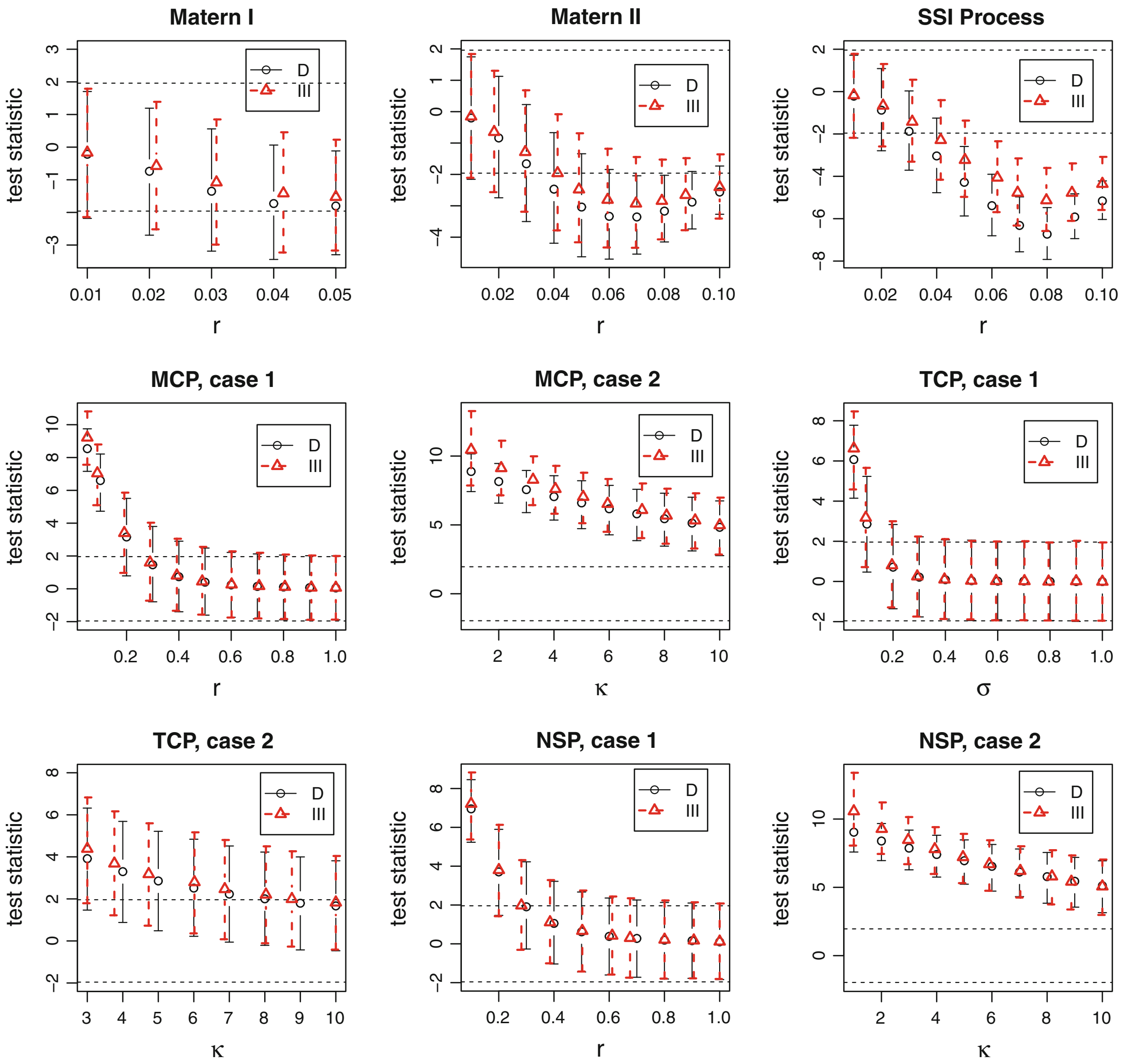

Fig. 13 The test statistics (means and 95 % empirical CIs) for cell (2,2) with X points from a HPP with intensity $\kappa = 100$ and Y points from various other clustering and regularity patterns under cases (1)–(9) of Sect. 4.1. *The dashed horizontal lines* are as in Fig. 6 and *legend labeling* is as in Fig. 1. *SSI* simple sequential inhibition, *MCP* Matérn cluster process, *TCP* Thomas cluster process, *NSP* Neyman–Scott process

tend to increase in the positive direction. Hence as r decreases, the level of clustering of Y points increases, and thus level of segregation between X and Y points increases.

(5) In MCP, case 2, as κ increases, the number of clusters tends to increase as well. The test statistics are all significantly positive for the ranges of the parameters we considered, hence there is strong segregation between X and Y points. However, we also observe that as κ increases, test statistics tend to decrease. Hence as the number of clusters tend to increase, the level segregation decreases (provided we have the same total number of points on the average in the same support). That is, in our setup, when κ increases, the number of points per cluster tends to decrease proportional with κ. However, if κ were increasing with fixed μ, we might have a different trend in the level of segregation.

(6) In TCP, case 1, the test statistics have a similar trend as in MCP, case 1, with smaller test statistic values for the ranges of parameters considered. So as σ decreases, test statistics tend to increase in the positive direction, and hence as σ decreases, the level of segregation between X and Y points increases.

(7) In TCP, case 2, the test statistics have a similar trend as in MCP, case 2, with smaller test statistic values for the ranges of parameters considered. Hence as the number of clusters tend to increase, the level segregation decreases (provided we have the same number of points on the average).

(8) In NSP, case 1, the trend is as in the MCP, case 1 with test statistics tending to be positive for $r \leq 0.6$.

(9) In NSP, case 2, the trend is as in the MCP, case 2.

4.2 Multi-class patterns with both classes from clustering or regularity processes

We also investigate the case that both classes are from stochastic point processes which result in clustering or regularity in a one-class setting for the particular class. Our goal is to understand which clustering or regularity patterns in a one-class setting will result in segregation or association in a multi-class setting. For convenience, we study the two-class case. We generate X and Y points from the following spatial point processes (see Baddeley and Turner 2005 for more details):

(1) **Matérn Model II:** Both X and Y points are independently generated from the Matérn Model II inhibition process, MatérnII(κ, r). That is, first $n_1 = n_2 = 100$ points are generated from a HPP with intensity κ. Then each point is marked by an "arrival time", a number uniformly distributed in [0,1] independently of other points. Any X point that lies closer than distance r from another X point whose arrival time is smaller is deleted. Likewise a similar thinning is applied to Y points. We use $\kappa = 100$ and $r = 0.01, 0.02, \ldots, 0.10$ in our simulations.

(2) **MCP, case 1:** In this case, both X and Y points are generated independently from Matérn's cluster process in the unit square, MCP(κ, μ, r) with different parent sets for X and Y points. That is, first "parent" points are generated for X points from a Poisson process with intensity κ and then each parent is replaced by X points independently uniformly generated inside the circle centered at the parent point with radius r, where number of these X points follows a Poisson(μ) distribution. Y points are generated similarly. Here we take $\kappa = 5, \mu = 20$ and $r = 0.1, 0.2, \ldots, 1.0$ for both classes in our simulations.

(3) **MCP, case 2:** In this case, both X and Y points are generated independently from Matérn's cluster process in the unit square, MCP(κ, μ, r), but with the same parent set for both of X and Y points. That is, first "parent" points are generated from a Poisson process with intensity κ and then X and Y points are generated around these same parents as in case 1 above. Here we take $\kappa = 10, \mu = 20$ and $r = 0.1, 0.2, \ldots, 1.0$ for each class as well in our simulations.

(4) **MCP, case 3:** In this case, both X and Y points are generated independently from Matérn's cluster process in the unit square, MCP(κ, μ, r) with different parent sets for X and Y points. Here we take $\kappa = 1, 2, \ldots, 10, \mu = \lfloor 100/\kappa \rfloor$ and $r = 0.1$ for both classes in our simulations.

(5) **MCP, case 4:** This is the same as case 3 above but with the same parent sets for both X and Y points, and $\kappa = 1, 2, \ldots, 10, \mu = \lfloor 100/\kappa \rfloor$ and $r = 0.1$ for each class in our simulations.

(6) ***X* points from Matérn II Process**, ***Y* points from Matérn Clustering Process:** In this case, X points are generated from MatérnII(κ_i, r_i) process which is a regularity (or inhibition) process, while Y points are from MCP(κ_c, μ, r_c) process which is a clustering process. For MatérnII(κ_i, r_i) process we take $\kappa_i = 100$ and $r_i = 0.01, 0.02, \ldots, 0.10$, and for MCP($\kappa_c$, μ, r_c) process we take $\kappa_c = 5$, $\mu = 20$, and $r_c = 0.05, 0.1, 0.2, 0.3, \ldots, 1.0$. That is, we have generated X and Y points for each combination of (r_i, r_c) values in our simulations.

In cases (1)–(5), 10,000 Monte Carlo replications are generated for each parameter value at each case, and in case (6), we generate 1,000 Monte Carlo replications for each combination of (r_i, r_c) values. Under case (1) above, both classes X and Y are from the same regularity or inhibition process, while under cases (2)–(5), both classes are from the same clustering process, with same or different parents. Under case (6), we have a mixed process, where class X is from a regularity process, while class Y is from a clustering process.

The means and the 95 % empirical CIs around the means for the cell (2,2) statistics under the processes (1)–(5) are presented in Fig. 14. Cell (1,1) statistics are very similar to the cell (2,2) statistics, hence are omitted. Furthermore, the overall test statistics do not provide the direction of deviation from the null hypothesis, hence are not presented either.

4.2.1 Comparison of cell (2,2) statistics for each point process

(1) In this case, the test statistics tend to be negative and start to be significantly negative for $r \geq 0.02$. As r increases the level of regularity for each class increases, and we also observe that the level of association between X and Y points increases as well. That is, when both X and Y points are from the same inhibition or regularity process, members of a class tend to repel conspecifics (i.e., their own kind), but do not repel members from the other class. Hence indirectly, this causes mixed NN pairs to be more likely in the pattern, hence the association between the classes. In other words, the association between

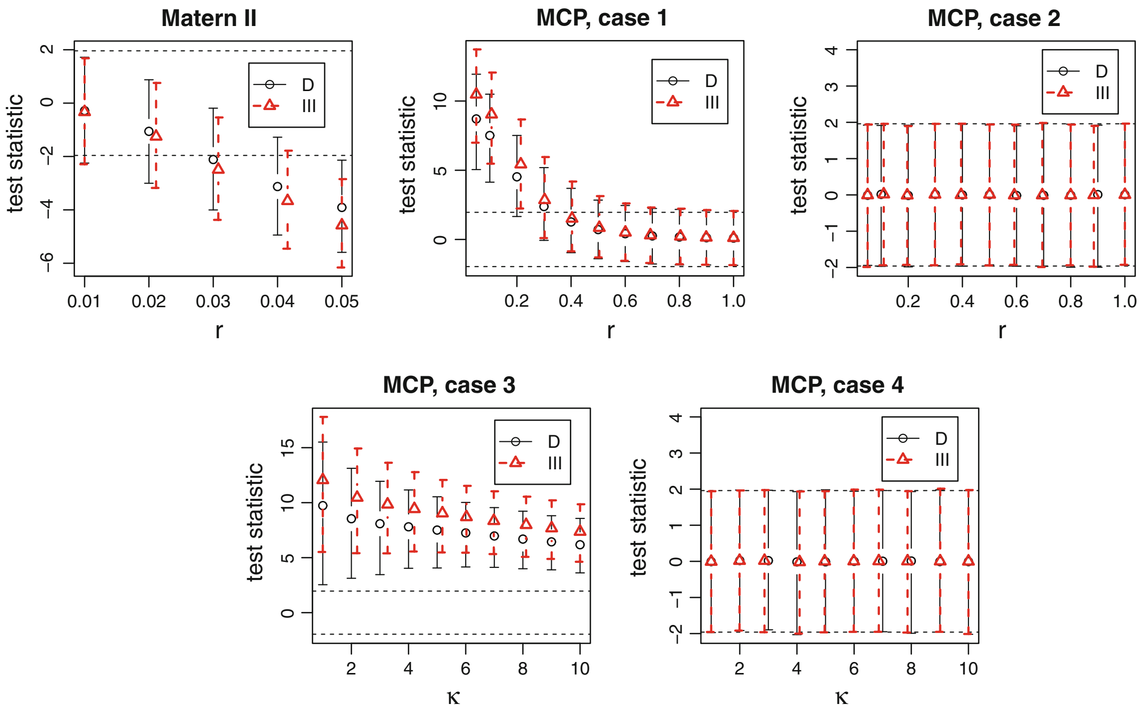

Fig. 14 The test statistics (means and 95 % empirical CIs) for cell (2,2) with *X* points and *Y* points from various clustering or regularity patterns under cases (1)–(5) of Sect. 4.2. The *dashed horizontal lines* are as in Fig. 6 and *legend labeling* is as in Fig. 1. *MCP* Matérn cluster process

X and *Y* is not necessarily because of the attraction between the classes, but because each class inhibits or repels its own kind, which renders its NNs to be from the other class more often than expected.

(2) In this case, the test statistics tend to be significantly positive for $r \leq 0.5$, and not significantly different from zero for the larger r values. Hence for $r \leq 0.5$, there is significant segregation between the classes, and for larger r values the pattern is not significantly different from the null pattern. Also we observe that as r decreases the level of segregation increases. By construction, as r decreases, each class would be more strongly clustered. Hence, when points are from the same clustering process with different parents, as the level of clustering increases, so does the level of segregation.

(3) In this case, both classes are from the same Matérn cluster process as in case (2) but with the same parent points. We observe that the test statistics are not significant in either direction, hence the patterns do not significantly deviate from the null hypothesis. The reason behind this is that, when the classes share the same parents, then it is as if the RL of points, after the points are generated from a clustering process. Hence the NN probabilities are expected to be proportional to the class sizes, which is our H_o.

(4) In this case, the test statistics are all significantly positive for all κ values considered, which suggests significant segregation between the classes. However, we also observe that as κ increases, the test statistics tend to decrease, and so does the level of segregation. Hence, when points are from the same clustering process with different parents, as the number of clusters increases, the level of segregation tends to decrease provided the average number of points per cluster, μ, is proportionally decreasing with κ to have the total average number of points fixed.

(5) In this case, we have the same conclusion as in case (3) above, since both classes are from the same Matérn clustering process with the same parent points.

(6) In this case, the mean test statistics for each combination of (r_i, r_c) values are presented in Fig. 15 as a gray-scale picture, where only test statistics for Dixon's cell (1,1) and cell (2,2) statistics are presented; the test statistics for type III cell-specific tests are similar, hence are omitted. In this case, we observe positive test statistics values for smaller r_c values, and negative test statistics for larger r_c values. Hence for smaller r_c values, the clustering of *Y* points dominate the pattern and causes segregation between the two classes; and for larger r_c values the clustering is weak, and the inhibition of *X* points dominate the pattern and suggest association between the two classes. However, considering the test statistics, we observe that positive

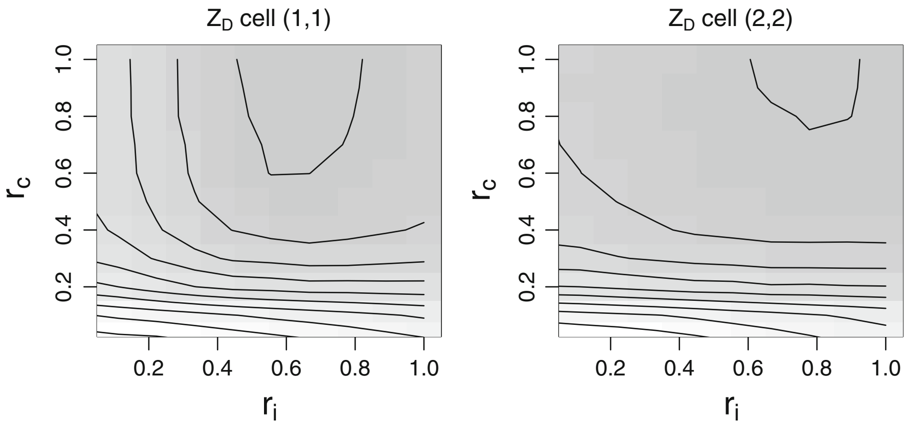

Fig. 15 The means of the Dixon's cell (1,1) and cell (2,2) test statistics under case (6) in Sect. 4.2 for combinations of $r_c = 0.05, 0.1, 0.2, \ldots, 1.0$ and $r_i = 0.1, 0.2, \ldots, 1.0$ plotted as a *gray-scale image* together with the contours. The test statistics decrease in value as gray level gets darker

test statistics are significant for cell (1,1) statistic for $r_c \leq 0.3$, while cell (2,2) statistics are positively significant for $r_c \leq 0.2$. The negative test statistics are significant for cell (1,1) for $r_c \geq 0.5$ and $r_i \geq 0.4$, but the negative statistics are not significant for cell (2,2). Therefore, the simulation results suggest that when there is strong clustering, it is reflected as segregation at both cell-specific tests, and when clustering is weak, but inhibition is strong, it is only reflected at the cell corresponding to the class from the inhibition process.

Remark 4.2 Segregation or association can result even if each class is independently generated from a different pattern. The discussion and investigation provided in this section may help decide what to expect between classes if the generative pattern is known or estimated fairly well for each class. Furthermore, one can fit the best model for each class, and assess various aspects (e.g., power, distribution or estimation) for the current or new methods. If one detects segregation or association in a multi-class setting, then s/he can fit the generative pattern for each class. This will help understand the causes of the observed pattern of segregation or association. □

5 Spatial patterns between three classes

The two-class patterns of Sect. 4 are not so realistic in practice, since usually marked point patterns in real life consist of three or more classes. In the three-class case, we consider various association patterns as well as mixed patterns. Because, with three or more classes, patterns and interactions may get very complicated, in the sense that classes may exhibit different patterns or different levels of the same pattern with respect to the other classes. Hence, a complete analysis and interpretation in such multi-class cases are more challenging compared to the two-class case.

5.1 Association between three classes

In the three-class case, we parameterize the association alternatives as follows. Let $\mathcal{X}_{n_1}$ be a random sample from $\mathcal{U}((0,1) \times (0,1))$. Then generate Y_j and Z_k for $j = 1, 2, \ldots, n_2$ and $k = 1, 2, \ldots, n_3$ as follows. For each j, select an i randomly from $\{1, 2, \ldots, n_1\}$, and set $Y_j := X_i + R_j^Y (\cos T_j, \sin T_j)'$ where $R_j^Y \overset{iid}{\sim} \mathcal{U}(0, r_y)$ with $r_y \in (0,1)$ and $T_j \overset{iid}{\sim} \mathcal{U}(0, 2\pi)$. Similarly, for each k, select an i' randomly from $\{1, 2, \ldots, n_1\}$, and set $Z_k := X_{i'} + R_k^Z (\cos U_\ell, \sin U_\ell)'$ where $R_k^Z \overset{iid}{\sim} \mathcal{U}(0, r_z)$ with $r_z \in (0,1)$ and $U_k \overset{iid}{\sim} \mathcal{U}(0, 2\pi)$.

We consider the following association alternatives:

$$H_A^1 : r_y = 1/7, r_z = 1/10, \quad H_A^2 : r_y = 1/10, r_z = 1/20,$$
$$H_A^3 : r_y = 1/13, r_z = 1/30, H_A^4 : r_y = \frac{1}{2\sqrt{n_1}}, r_z = 1/10,$$
$$H_A^5 : r_y = \frac{1}{2\sqrt{n_1}}, r_z = \frac{1}{4\sqrt{n_1}}, \quad H_A^6 : r_y = \frac{1}{2\sqrt{n_1 + n_2}},$$
$$r_z = \frac{1}{2\sqrt{n_1 + n_3}}, \quad H_A^7 : r_y = \frac{1}{2\sqrt{n}}, r_z = \frac{1}{4\sqrt{n}} \tag{12}$$

As r_y or r_z or both decrease, the level of association increases. That is, the association between classes X and Y and association between classes X and Z get stronger from H_A^1 to H_A^7. By construction, classes Y and Z are associated with class X, while classes Y and Z are not associated, but perhaps mildly segregated for small r_y and r_z. Furthermore, by construction, classes X and Z are more associated compared to classes X and Y.

Similar to type C association, under these association alternatives, although R_j^Y, R_k^Z, θ_j^Y and θ_k^Z are uniformly distributed in their respective ranges, Y_j are not uniformly distributed in the circles centered at X_i with radius r_y, and Z_k are not uniformly distributed in the circles centered at $X_{i'}$ with radius r_z. In fact, letting $Y = (T_1, V_1)$ and $Z = (T_2, V_2)$, joint pdf of (T_1, V_1) is $f_{T_1,V_1}(t_1, v_1) =$

$\frac{1}{2\pi r_y\sqrt{t_1^2+v_1^2}}$ for $0<t_1^2+v_1^2\le r_y^2$, and joint pdf of (T_2,V_2) is $f_{T_2,V_2}(t_2,v_2)=\frac{1}{2\pi r_z\sqrt{t_2^2+v_2^2}}$ for $0<t_2^2+v_2^2\le r_z^2$.

Under the above association alternatives, we have various levels of association for each pair of classes.

Under $H_A^1 - H_A^3, Y$ points are associated with X points and Z points are associated with X points with the level of association increasing from H_A^1 to H_A^3. In each alternative, Z points are more strongly associated with X points compared to Y points.

Under H_A^4 and H_A^5, Y and Z points are associated with X points with the level of association depending on n_1. With $n_1 = 100$, under H_A^4, Y points are more strongly associated with X points, and under H_A^5, Z points are more strongly associated with X points.

Under H_A^6, Y points are associated with X points with the level of association depending on $n_1 + n_2$ and Z points are associated with X points with the level of association depending on $n_1 + n_3$. With $n_1 = n_2 = n_3 = 100$, both levels of association are same

Under H_A^7, Y and Z points are associated with X points with the level of association depending on n. With $n_1 = n_2 = n_3 = 100$, Z points are more strongly associated with X points.

The means and the 95 % empirical CIs around the means for the overall tests under the association alternatives are presented in Fig. 16(left). Overall test statistics are moderately significant in H_A^1 and are highly significant for all other cases with H_A^7 having the most significant values. This is in agreement with our construction that H_A^1 contains the weakest association levels, and H_A^7 contains the strongest association levels.

The means and the 95 % empirical CIs around the means for the cell-specific tests under the association alternatives are presented in Fig. 17. By symmetry in the sample sizes, cells (i, j) and (j, i) for $i \neq j$ have the same cell-specific test statistics, hence only cells (1,1), (1,2), (1,3), (2,2), (2,3), and (3,3) are presented in Fig. 17.

5.1.1 *Comparison of test statistics for each cell (i, j)*

Cell (1,1) statistics are negative under all association cases and most significant under H_A^7. This implies the most significant lack of segregation of X points occurs under H_A^7, where we have the strongest association pattern.

Cell (1,2) statistics are either mildly or highly positive and the most significant value occurs under H_A^4. This suggests that H_A^4 contains the strongest association between X and Y points. Indeed, in our construction, only in this alternative we have the association of Y points with X points is stronger than the association of Z points with X points.

Cell (1,3) statistics are mildly or highly positive with the most significant values occurring under H_A^2, H_A^3, H_A^5, and H_A^7 (highest test statistics occurs under H_A^7). Under these alternatives, in our construction, association of Z points with X points is stronger than that of Y points with X points, with r_z being at least half of r_y.

Cell (2,2) statistics are mildly or highly negative with the most significant values occurring under $H_A^4 - H_A^7$ (lowest statistics occur under H_A^4 and H_A^7) which implies severe lack of segregation of Y points under these alternatives. In our construction, association of Y points with X points is stronger under these alternatives compared to other alternatives.

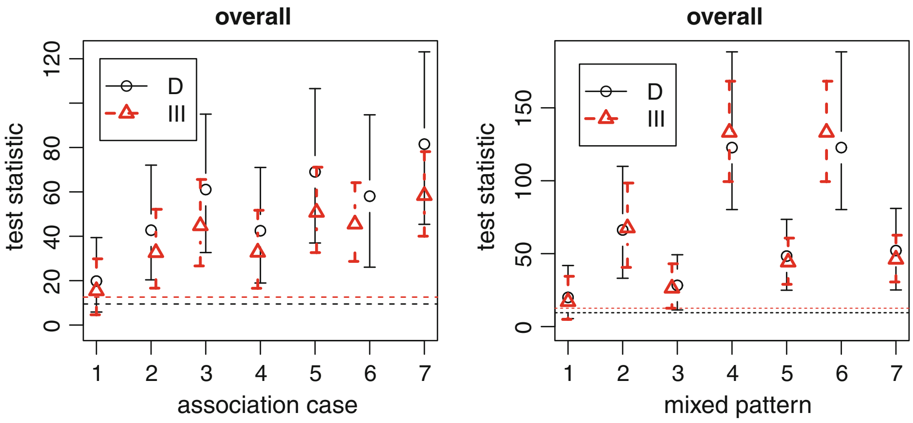

Fig. 16 The test statistics (means and 95 % empirical CIs) for the overall tests based on 10,000 Monte Carlo replications under the association alternatives (*left*) and mixed alternatives (*right*) in the three-class case with $n_1 = n_2 = n_3 = 100$. The *dashed horizontal lines* are as in Fig. 2, *legend labeling* is as in Fig. 1, and the *horizontal axis* contains the association case numbers 1–7 of Sect. 5.1 and mixed pattern numbers 1–7 of Sect. 5.2

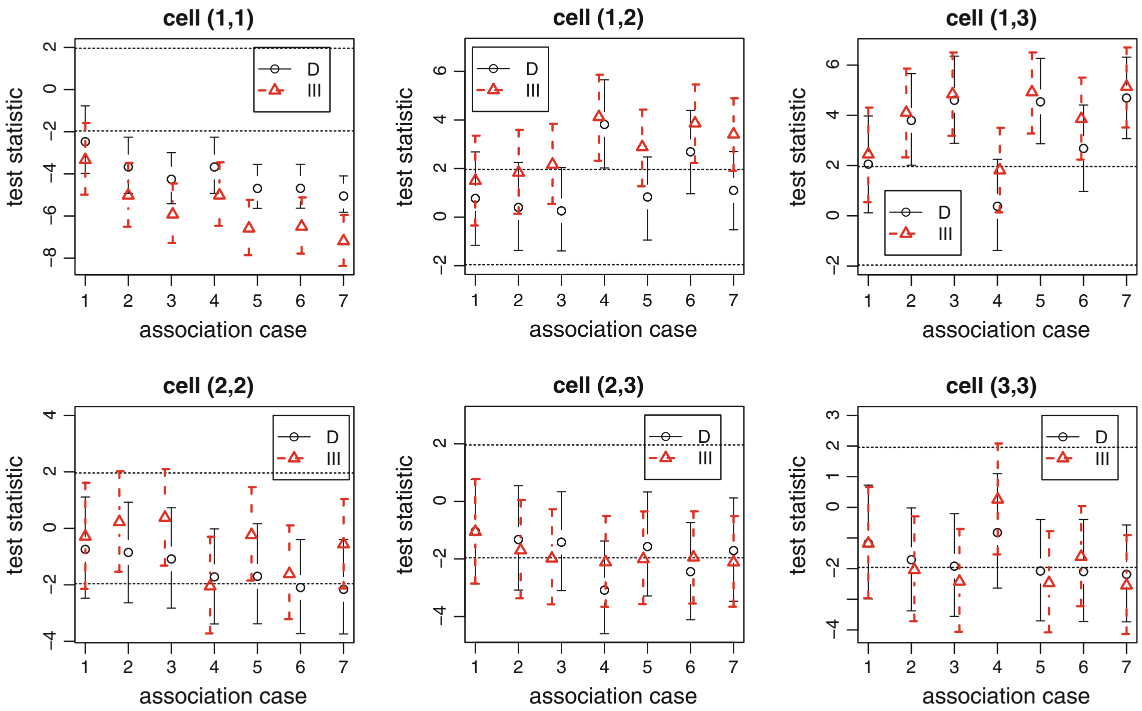

Fig. 17 The test statistics (means and 95 % empirical CIs) for the cell (1,1), (1,2), (1,3), (2,2), (2,3), (3,3) statistics based on 10,000 Monte Carlo replications under the association alternatives in the three-class case with $n_1 = n_2 = n_3 = 100$. The *dashed horizontal lines* are as in Fig. 6, *legend labeling* is as in Fig. 1, and the *horizontal axis* is association case numbers 1–7 corresponding to $H_A^1 - H_A^7$ in Sect. 5.1

Cell (2,3) statistics are mildly or highly negative with the most significant values occurring under H_A^4 which implies severe lack of association between Y and Z points. In our construction, Y and Z points are not associated.

Cell (3,3) statistics are mildly or highly negative with the most significant values occurring under H_A^2, H_A^3 and $H_A^5 - H_A^7$ (lowest statistics occurring under H_A^3, H_A^5, and H_A^7) which implies severe lack of segregation of Z points. In our construction, Z points are strongly associated with X points.

5.1.2 *Comparison of cell-specific test statistics under each association alternative*

Under H_A^1, cell (1,1) statistics are moderately negative, and cell (2,2), (2,3), and (3,3) statistics are mildly negative. On the other hand, cell (1,2) statistics are mildly positive and cell (1,3) statistics are moderately positive. All these together imply that X points exhibit moderate lack of segregation while Y and Z points have mild lack of segregation; and there is mild lack of association between Y and Z points; and there is mild association between X and Y points, and moderate association between X and Z points. In our construction, Y points are associated with X points, and so are Z points but at a higher level.

Under H_A^2, we observe the same trends as in H_A^1 but at a higher level and under H_A^3 we observe the same trends as in H_A^2 but at a higher level.

Under H_A^4, cell (1,1) statistics are highly significant in the negative direction, cell (2,2) and (2,3) statistics are moderately negative and cell (3,3) statistics are mildly negative. On the other hand, cell (1,3) statistics are mildly positive and cell (1,2) statistics are strongly positive. All these considered imply severe lack of segregation for X points, and mild lack of segregation for Y and Z points, moderate lack of association between Y and Z points, strong association between X and Y points and mild association between X and Z points. In our construction Y and Z points are associated with X points where association of Y points is stronger.

Under H_A^5 cell (1,1) statistics are highly negative, cell (2,2), (2,3), and (3,3) statistics are moderately negative. On the other hand, cell (1,3) statistics are strongly positive and cell (1,2) statistics are moderately positive. All these considered imply severe lack of segregation for X points, and moderate lack of segregation for Y and Z points, moderate lack of association between Y and Z points, strong association between X and Z points and moderate association between X and Y points. In our construction Y and Z points are associated with X points where association of Z points is stronger.

Under H_A^6, we have a similar trend as in H_A^5 for cell (1,1), (2,2), (2,3), and (3,3) statistics. On the other hand, cell (1,2) and (1,3) statistics are moderately but equally positive. All these considered imply moderate association between X and Z points and the same level of association between X and Y points. In our construction Y and Z points are equally associated with X points.

Under H_A^7, we have a similar trend as in H_A^5 with more significant test statistics.

5.2 Mixed alternatives for three classes

In the k-class case with $k \geq 3$, it is possible to have segregation between a pair of classes, or association between another pair, or the null pattern in another one. We call such patterns as "mixed" alternatives. We explore the following mixed patterns:

Case 1: Let $\mathcal{X}_{n_1}$ be a random sample of size n_1 from $\mathcal{U}((0,1)\times(0,1))$ and $Y \overset{d}{=} X$, i.e., $\mathcal{Y}_{n_2}$ is a random sample of size n_2 from $\mathcal{U}((0,1)\times(0,1))$ and $\mathcal{Y}_{n_2}$ is independent of $\mathcal{X}_{n_1}$. Then generate Z_k for $k = 1,2,\ldots,n_3$ as follows. For each k, select an i randomly, and set $Z_k := X_i + R_k^Z(\cos U_\ell, \sin U_\ell)'$ where $R_k^Z \overset{iid}{\sim} \mathcal{U}(0, r_z)$ with $r_z \in (0,1)$ and $U_k \overset{iid}{\sim} \mathcal{U}(0, 2\pi)$. For this case, we select $r_z = 1/10$. So in this case, X and Y points follow CSR independence pattern, while Z points are associated only with X points.

Case 2: Generate $X_i \overset{iid}{\sim} \mathcal{U}((0,3/4)\times(0,3/4))$ for $i = 1,2,\ldots,n_1$, $Y_j \overset{iid}{\sim} \mathcal{U}((1/4,1)\times(1/4,1))$ for $j = 1,2,\ldots,n_2$, and generate Z_k for $k = 1,2,\ldots,n_3$ as follows. For each k, select an i randomly, and set $Z_k := X_i + R_k(\cos T_k, \sin T_k)'$ where $R_k \overset{iid}{\sim} \mathcal{U}(0, r_z)$ with $r_z = 1/10$. In this case, X and Y points are segregated, and X and Z are associated, so it is expected that Y and Z are (indirectly) segregated but at a lesser extent.

Case 3: Generate X_i and Y_j as in Case 2. Combine $\mathcal{X}_{n_1}$ and $\mathcal{Y}_{n_2}$ to form a sample of size $n_1 + n_2$, and relabel as $\mathcal{W}_{n_1+n_2} = \{W_1, W_2, \ldots, W_{n_1+n_2}\}$. Then generate Z_k for $k = 1,2,\ldots,n_3$ as follows. For each k, select an i' randomly, and set $Z_k := W_{i'} + R_k(\cos U_k, \sin U_k)'$ where $R_k \overset{iid}{\sim} \mathcal{U}(0, r_z)$ with $r_z = 1/10$ and $U_k \overset{iid}{\sim} \mathcal{U}(0, 2\pi)$. In this case, X and Y points are segregated as in Case 2, and Z points are associated with both X and Y points.

Case 4: Generate $X_i \overset{iid}{\sim} \mathcal{U}((0,2/3)\times(0,2/3))$ for $i = 1,2,\ldots,n_1$, $Y_j \overset{iid}{\sim} \mathcal{U}((1/3,1)\times(1/3,1))$ for $j = 1,2,\ldots,n_2$, and generate Z_k as in Case 2 with $r_z = 1/20$. In this case, X and Y points are segregated (at a higher level than Cases 2 and 3), and X and Z are associated, so it is expected that Y and Z are (indirectly) segregated as well but at a lesser extent.

Case 5: Generate X_i and Y_j as in Case 4 and generate Z_k as in Case 3 with $r_z = 1/20$. In this case, X and Y points are segregated as in Case 4, and Z points are associated with both X and Y points (at a higher level than Cases 1–3).

Case 6: Generate X_i and Y_j as in Case 4 and generate Z_k as in Case 2 with $r_z = 1/(2\sqrt{n_1})$. In this case, X and Y points are segregated as in Case 4, and Z points are associated only with X points and so it is expected that Y and Z points are segregated but at a lesser extent.

Case 7: Generate X_i and Y_j as in Case 4 and generate Z_k as in Case 3 with $r_z = 1/(2\sqrt{n_1 + n_2})$. In this case, X and Y points are segregated as in Case 4, and Z points are associated with both X and Y points at a higher level than Case 5.

The means and the 95 % empirical CIs around the means for the overall tests under the mixed alternatives are presented in Fig. 16(right). Observe that the overall tests are highly significant for all cases except Case 1, in which case overall tests are moderately significant. This is in agreement with the fact that in Case 1, X and Y points follow CSR independence, and Z points are associated with X points only with $r_z = 1/10$, so, Case 1 contains the weakest deviation from the null pattern. But in all other cases, X and Y points are segregated, and Z points are associated with either X points or with both of X and Y points combined. Also overall tests are mostly significant in cases 4 and 6, which include the highest levels of deviation from the null pattern.

The means and the 95 % empirical CIs around the means for the cell-specific tests under the mixed alternatives are presented in Fig. 18. Notice that by symmetry, cells (i, j) and (j, i) for $i \neq j$ have the same test statistic values, hence only cell (1,1), (1,2), (1,3), (2,2), (2,3), and (3,3) statistics are presented in Fig. 18.

5.2.1 *Comparison of test statistics for each cell (i, j)*

Cell (1,1) tests are most significant for Cases 3, 5, and 7 all of which are in the positive direction (with Case 5 being slightly higher). This is in agreement with our setup that in these cases, we have the strongest

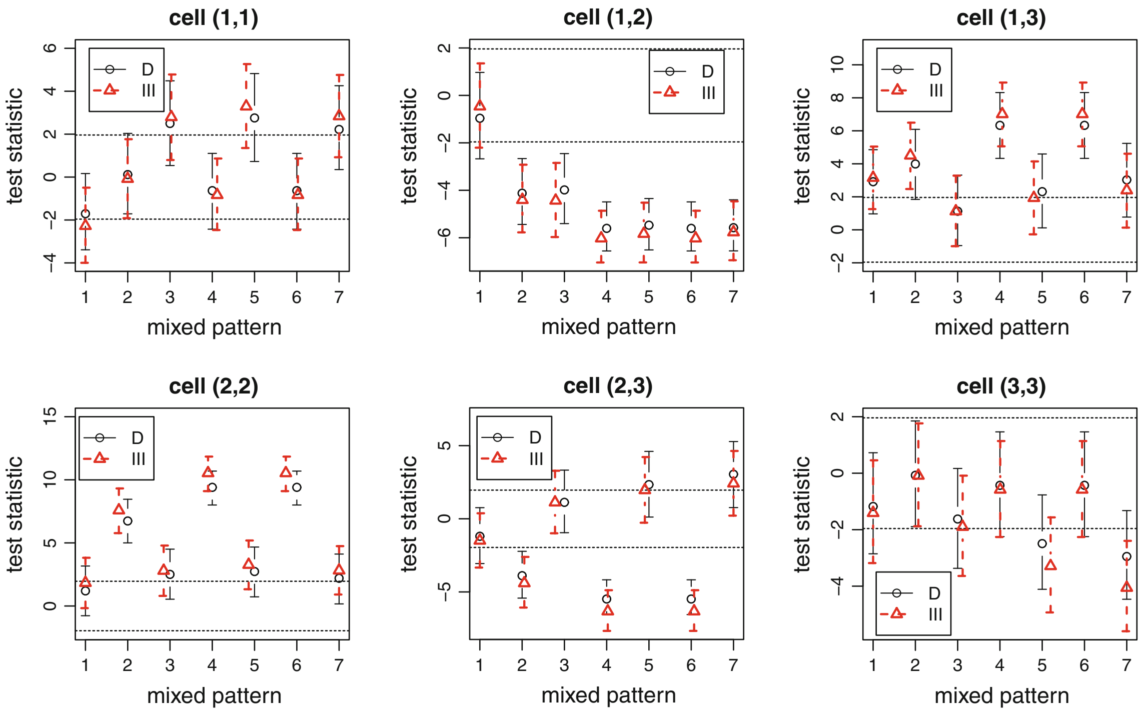

Fig. 18 The test statistics (means and 95 % empirical CIs) for the cell (1,1), (1,2), (1,3), (2,2), (2,3), (3,3) statistics based on 10,000 Monte Carlo replications under the mixed alternatives in the three-class case with $n_1 = n_2 = n_3 = 100$. The *legend labeling* is as in Fig. 1 and the *horizontal axis* is mixed pattern numbers 1–7 of Sect. 5.2

segregation between X and Y and association of Z with these points is weaker compared to the other cases, where Z is associated only with X points, which reduces the observed N_{11} value.

Cell (1,2) tests are most significant for Cases 4–7 all of which are in the negative direction suggesting either strongest lack of association or strongest segregation between classes X and Y; and in these cases X and Y points are strongly segregated compared to other cases.

Cell (1,3) tests are most significant for Cases 4 and 6 both of which are in the positive direction suggesting strongest association between classes Z and X; and in these cases Z points are more strongly associated only with X points with $r_z = 1/20$.

Cell (2,2) tests are most significant for Cases 4 and 6 both of which are in the positive direction suggesting strongest segregation of Y points from other classes. In these cases X and Y points are more strongly segregated and Z points are associated only with X points with $r_z = 1/20$. So Y points are directly segregated from X points and indirectly segregated from Z points. Therefore, segregation between X and Y is reflected in cell (2,2) but not in cell (1,1).

Cell (2,3) tests are most significant for Cases 4 and 6 in the negative direction suggesting strongest segregation between Y and Z; and in these cases Y points are more strongly segregated from X points and Z points are associated with X points but not with Y points, and hence Y and Z points are indirectly segregated. Moreover, cell (2,3) statistics are most significant for Cases 5 and 7 in the positive direction suggesting strongest association between Y and Z; and in these cases Z points are strongly associated with Y points (in addition to X points).

Cell (3,3) tests are most significant for Cases 5 and 7 in the negative direction (with Case 7 being more negative) suggesting strongest lack of segregation for class Z; and in these cases Z points are more strongly associated with both X and Y points together.

5.2.2 *Comparison of cell-specific test statistics under each mixed alternative*

Under mixed case 1, cell (1,3) statistic is the most significant one, and this being in the positive direction verifies the association between X and Z points. Cell (1,1) and (3,3) statistics are slightly negative suggesting slight lack of segregation for these classes, which might result from the association between X and Z points. Cell (1,2) and (2,3) are almost within the null region (i.e., within (−1.96,1.96)), but closer to the negative end, suggesting a very mild lack of association between classes X and Y and between classes Y and Z.

Under mixed case 2, cell-specific test statistics are within the null region for cells (1,1) and (3,3), significantly negative for cells (1,2) and (2,3), and significantly positive for cells (2,2) and (1,3). All these taken together suggest significant lack of association or significant segregation between X and Y points and between Y and Z points and significant association between X and Z points and significant segregation of Y points from others; which seem to result from the construction that we have segregation between X and Y and since Z points are associated with X points, they are also segregated from Y points.

Under mixed case 3, cell (1,1) statistic is highly significant in the positive direction, cell (2,2) statistic is moderately significant in the positive direction and cell (3,3) statistics is mildly significant in the negative direction, which implies significant segregation of X points and moderately significant segregation of Y points and lack of segregation for Z points. Also cell (1,2) statistic is highly significant in the negative direction implying lack of association between X and Y points, and cell (1,3) and (2,3) statistics are mildly significant in the positive direction implying mild association of class Z with classes X and Y. All these are in agreement with the construction that we have segregation between X and Y, and Z points are mildly associated with both X and Y points.

Under mixed case 4, cell (1,3) and (2,2) statistics are highly significant in the positive direction, cell (1,2) and (2,3) statistics are highly significant in the negative direction, and cell (1,1) and (3,3) statistics are within the null region. All these considered together imply that X and Z points are associated, and there is lack of association between X and Y points and between Y and Z points, and Y points are segregated from X and Z points. All these are in agreement with the construction that we have segregation between X and Y, and Z points are associated only with X points.

Under mixed case 5, cell (1,1), (2,2), (1,3), and (2,3) statistics are moderately significant in the positive direction, cell (3,3) statistic is moderately significant in the negative direction, and cell (1,2) statistic is highly significant in the negative direction. All these considered together imply that there is moderate segregation and severe lack of association between X and Z points, and moderate association between X and Z and between Y and Z points, and mild lack of segregation of Y points from other classes. All these are in agreement with the construction that we have segregation between X and Y, and Z points are associated with both X and Y points.

Under mixed case 6, cell (2,2) and (1,3) statistics are highly significant in the positive direction, cell (1,1) and (3,3) statistics are within the null region, and cell (1,2) and (2,3) statistics are highly significant in the negative direction. All these considered together imply that there is severe segregation of Y points, and severe association of X and Z points, and severe lack of association between X and Y and between Y and Z points. All these are in agreement with the construction that we have segregation between X and Y, and Z points are associated only with X points.

Under mixed case 7, cell (1,1), (2,2), (1,3), and (2,3) statistics are moderately significant in the positive direction, cell (1,2) statistic is highly significant in the negative direction, and cell (3,3) statistic is moderately significant in the negative direction. All these considered together imply that there is moderate segregation of X and Y points, and moderate association between X and Z points and between Y and Z points, and severe lack of association between X and Y points. All these are in agreement with the construction that we have segregation between X and Y, and Z points are associated with both X and Y points.

6 Example data sets

We apply the methodology on part of the Lansing Woods data and on bramble canes data for illustrative purposes. Both data sets are also available in the spatstat package in R (Baddeley and Turner 2005).

6.1 Lansing Woods data

This data set contains the locations of 2,251 trees in feet (ft) together with their species classification in a 924 × 924 ft (19.6 acre) plot in Lansing Woods, Clinton County, MI, USA (Gerrard 1969). The species of the trees are hickory, maple, red oak, white oak, black oak and miscellaneous trees. In our analysis, for brevity, we only consider black oaks, hickories, and maples which constitute a total of 1,352 trees. That is, we analyze the spatial patterns for the above three species as if only they exist in the area, so we ignore the possible effects of other species on the spatial interaction between these species. Thus, we first perform a 3 × 3 NNCT-analysis on this data set. See Fig. 19 for the location of the trees in this plot and Table 2 for the associated 3 × 3 NNCT together with cell percentages based on class sizes, and marginal percentages based on the grand total, n. For example, when black oak is the base species and hickory is the NN species, we observe that the cell count is 68 which is 50 % of the 135 black oaks (i.e., 50 % of the NNs of black oaks are from hickories and hickories are 52 % of all trees). Figure 19 and percentages in Table 2 suggest that tree species are segregated from each other since the observed percentages of

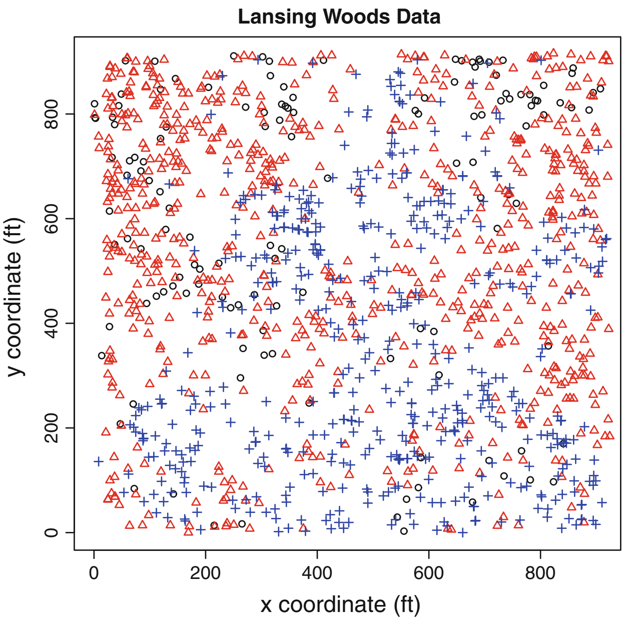

Fig. 19 The scatter plot of the locations of black oaks (*circles*), hickories (*triangles*), and maples (*pluses*) in the Lansing Woods, Clinton County, MI, USA

Table 2 The NNCT for Lansing Woods data and the corresponding percentages (in parentheses), where the cell percentages are with respect to the size of the base species, and marginal percentages are with respect to the total size

	NN			Sum
	BO	H	M	
Base				
BO	36 (27 %)	68 (50 %)	31 (23 %)	135 (10 %)
H	79 (11 %)	502 (71 %)	122 (17 %)	703 (52 %)
M	25 (5 %)	130 (25 %)	359 (70 %)	98 (38 %)
Sum	140 (10 %)	700 (52 %)	512 (38 %)	1,352 (100 %)

BO black oaks, *H* hickories, *M* maples

species in the diagonal cells are much larger than the row percentages (or species relative frequencies).

For a NNCT analysis, the null pattern depends on the particular ecological setting (Goreaud and Pélissier 2003). More specifically, under RL, the individuals of a single population are affected by some processes a posteriori (e.g., diseased vs. non-diseased individuals of a single species). On the other hand, under CSR independence, different processes generate the two classes a priori (e.g., individuals of different species or age cohorts). So for the Lansing Woods data, the more appropriate null hypothesis is the CSR independence pattern, since different processes seem to be generating the locations of the tree species a priori. For this data set, $Q = 528$ and $R = 938$ and our inference is conditional on these values. Dixon's and type III overall and cell-specific tests and the associated p-values are presented (in parentheses) in Table 3, where the first p-value in the parenthesis is based on the asymptotic approximation and the second is based on Monte Carlo randomization of the labels on the given locations of the trees 10,000 times. Notice that p-values are all significant. Moreover, both p-values in Table 3 are similar for each cell-specific test.

The overall segregation tests are both highly significant, implying a significant deviation from the CSR independence pattern for some of the tree species. As post-hoc tests, we resort to the cell-specific tests to determine which species exhibit segregation or association. In the NNCT, the diagonal cell statistics are all positive and significant at 0.05 level, hence each species seems to exhibit segregation from the others. Furthermore, the off-diagonal cell statistics are all negative and significant, implying lack of association between any pair of species, which is also reflected in the strong segregation of each species.

To understand the underlying patterns which cause or account for the segregation in the current data set, we perform point pattern fitting for locations of each species. In particular, the locations of the trees in Fig. 19 suggest that the species might be from clustering processes. Along this line, we fit Thomas and Matérn cluster processes to each species (see Baddeley and Turner 2005 for more detail). The model fitting is performed using the method of minimum contrasts (see Diggle 2003). We only present the fit for TCP, since it is a slightly better fit for the data compared to the MCP fit. In particular, the K-function for TCP with parameters $\theta = (\kappa, \mu, \sigma)$ is

$$K_\theta(r) = \pi r^2 + \frac{1}{\kappa}\left(1 - \exp\left(\frac{-r^2}{4\sigma^2}\right)\right).$$

The model is fit by determining the values of θ that minimizes $\int_a^b |\widehat{K}^q(r) - K_\theta^q(r)|^p dr$ where a and b are ranges of r values considered and p, q are indices (we use $p = 2$ and $q = 1/4$, which are the defaults in spatstat package (Waagepetersen 2007)). We choose stationary point processes in our model fitting, if needed non-stationary versions are also available (Baddeley and Turner 2005). The estimated parameter values for the TCP models for each tree species are presented in Table 4. The fitted K-function for the TCP process and the observed K-function (with Ripley's isotropic correction for edge effects) are plotted in Fig. 20. In the same figure, plotted also are theoretical K-function under HPP together with a 95 % empirical confidence interval around it. Observe that the fitted and observed K-functions are almost a perfect match. Furthermore, the observed K-function is significantly above the theoretical K-function for all distances considered (i.e., for $r \in (0, 235)$ ft). Thus, there is significant clustering for each

Table 3 Test statistics and *p*-values for the overall and cell-specific tests and the corresponding *p*-values (in parentheses)

C_D			C_{III}
Overall tests			
249.86 (<0.0001, <0.0001)			249.67 (<0.0001, <0.0001)
	BO	H	M
Dixon's cell-specific tests			
BO	5.23 (<0.0001, <0.0001)	−0.30 (0.7675, 0.7810)	−3.65 (0.0003, 0.0003)
H	1.11 (0.2689, 0.2669)	9.67 (<0.0001, <0.0001)	−11.05 (<0.0001, <0.0001)
M	−4.11 (<0.0001, <0.0001)	−11.41 (<0.0001, <0.0001)	13.30 (<0.0001, <0.0001)
Type III cell-specific tests			
BO	5.26 (<0.0001, <0.0001)	−0.25 (0.7998, 0.7965)	−3.77 (0.0002, 0.0002)
H	1.08 (0.2814, 0.2761)	11.84 (<0.0001, <0.0001)	−13.63 (<0.0001, <0.0001)
M	−5.16 (<0.0001, <0.0001)	−12.91 (<0.0001, <0.0001)	14.98 (<0.0001, <0.0001)

The first *p*-value is based on the asymptotic approximation and the second on the randomization of the tests. C_D stands for Dixon's overall test and C_{III} for type III overall test

BO black oaks, *H* hickories, *M* maples

Table 4 The estimates of the parameters (κ, μ, σ), for the TCP fitting for Lansing Woods data (*top*) and bramble Canes data (*bottom*)

Species	$\widehat{\kappa}$ (per acre)	$\widehat{\mu}$	$\widehat{\sigma}$
Lansing Woods data			
Black oaks	0.75	9.14	51.83
Hickories	1.65	21.71	66.93
Maples	1.11	23.64	62.40
Age*	$\widehat{\kappa}$ (per m^2)	$\widehat{\mu}$	$\widehat{\sigma}$
Bramble canes data			
New	13.79	2.89	0.0348
One	9.19	4.66	0.0762
Two	7.31	1.20	0.0426

* "New" stands for "newly emergent", "one" stands for "one year old", and "two" stands for "two years old" brambles

species. Hence we can assume that the current allocation of the trees for each species is a realization of the respective TCP. The significant segregation between the species seems to be accounted for by the Thomas type clustering with different parameters and different parent points.

6.2 Bramble canes data

This data set contains the locations of 823 bramble canes in meter (m) together with their ages in a 9 m^2 plot in a field (Hutchings 1979). These canes were further classified according to their ages as "newly emergent, 1 or 2 years old". The data were also analyzed by Diggle (2003) and van Lieshout and Baddeley (1999). It was found in each of the analyses that newly emergent canes exhibit spatial clustering, which is attributed to "vigorous vegetative reproduction" by Hutchings (1979).

In our analysis, first we perform a 3 × 3 NNCT-analysis. See Fig. 21 for the locations of the canes in the study plot and Table 5 for the associated NNCT together with cell and marginal percentages as obtained in Sect. 6.1. Figure 19 and percentages in Table 2 suggest that newly emergent canes are neither segregated from nor associated with the other groups (as the percentage of newly emergent NNs of itself is exactly the same as the relative frequency of the newly emergent canes in the data), 1 year old canes seem to be associated with newly emergent ones, and 2 year old canes seem to be segregated.

As in the Lansing Woods data, the more appropriate null hypothesis is the CSR independence pattern for the bramble canes data as well. Hence our inference is conditional on $Q = 550$ and $R = 558$ (which are computed for this data). Dixon's and type III overall and cell-specific tests and the associated *p*-values are presented in Table 6. The *p*-values for the cell-specific tests are provided for the two-sided alternatives, hence we need to divide them by 2 to get the correct *p*-value for the appropriate one-sided alternative. Notice that although the overall tests agree, the cell-specific tests tend to give different results (not conflicting in direction, but different in significance). The overall tests are both highly significant, suggesting a significant deviation from the CSR independence pattern for at least one group of bramble canes. Considering the post-hoc cell-specific tests, along the diagonal we observe that only the cell for (TY, TY) is significant in the positive direction implying segregation of 2 year old canes from the other groups. On the other hand, the only significant off-diagonal cells are in row 1 and column 1 in the NNCT. Thus, we conclude that 1 year old canes are significantly associated with newly emergent ones, and newly emergent ones are

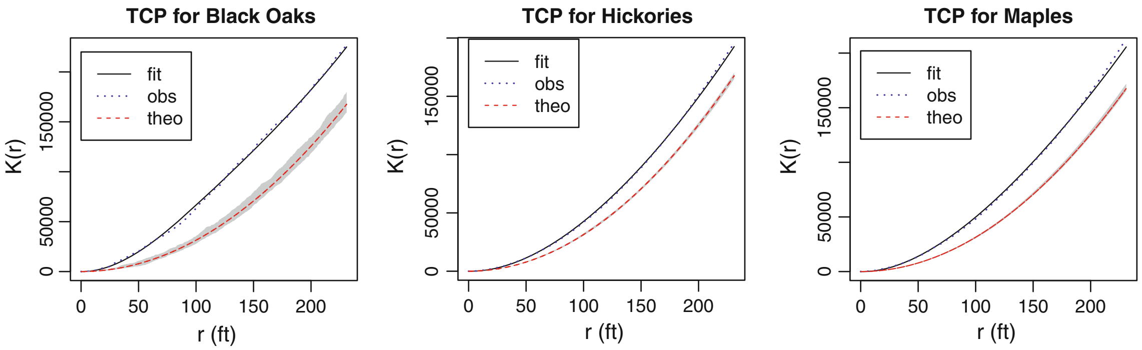

Fig. 20 The fitted *K*-function for the TCP process (fit), the observed *K*-function with Ripley's isotropic correction for edge effects (obs), and theoretical *K*-function under HPP (theo) together with a 95 % empirical confidence interval around it for each species in the Lansing Woods data

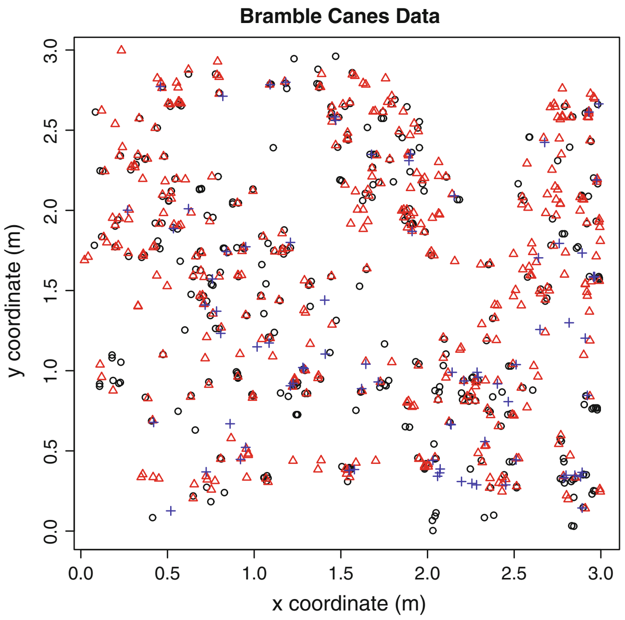

Fig. 21 The scatter plot of the locations of bramble canes where newly emergent ones are plotted as *circles*, 1 year olds as *triangles*, and 2 year olds as *pluses* in a field

Table 5 The NNCT for bramble canes data and the corresponding percentages (in parentheses) computed as in Table 2

	NN			Sum
	NE	OY	TY	
Base				
NE	158 (44 %)	186 (52 %)	15 (4 %)	359 (44 %)
OY	187 (49 %)	163 (42 %)	35 (9 %)	385 (47 %)
TY	26 (33 %)	39 (49 %)	14 (18 %)	79 (10 %)
Sum	371 (45 %)	388 (47 %)	64 (8 %)	823 (100 %)

NE newly emergent, *OY* one year old, *TY* two years old canes

moderately associated with 1 year old canes. On the other hand, (TY, NE) and (NE, TY) cells are significant in the negative direction which is caused by the significant segregation of 2 year old canes from newly emergent ones. Two year old canes and 1 year old canes do not exhibit significant deviation from CSR independence.

To discover the underlying patterns which account for the segregation/association among the bramble canes, we fit point pattern models to the locations of each group. The locations of the canes in Fig. 21 suggest that there is clustering for each group in the data. We first fit TCP to each group as described in Sect. 6.1. The estimated parameter values for the TCP models for each cane group are presented in Table 4. The fitted *K*-function for the TCP process, the *K*-function for the HPP (together with 95 % simulation envelopes), and the observed *K*-function (with Ripley's isotropic correction for edge effects) are plotted in Fig. 22(top). The observed *K*-function is significantly above the theoretical *K*-function for all distances considered (i.e., for $r \in (0, 0.75)$ m), which suggests significant clustering for each group. Observe also that the fitted and observed *K*-functions seem to be different especially for large *r* values suggesting that the pattern may not be from a stationary Thomas clustering type with homogeneous parents. The model fit with various inhomogeneous parents is also investigated, but the fit does not seem to improve (results not presented). However, when we compute the simulation envelopes based on the fitted TCP model we see that the observed *K*-function curve lies within the 95 % simulation envelope (see Fig. 22, bottom). Thus the models we fit are not that different from the reality after all.

One might also attempt fitting various inhomogeneous Poisson patterns by playing with the trend function. For example, for the 2 year old bramble canes, the simulation envelope for the fitted curve is very wide, and the plot is suggestive of an inhomogeneous Poisson process. Let $\lambda(x, y)$ is the intensity of the Poisson process as a function

Table 6 Test statistics and p-values for the overall and cell-specific tests and the corresponding p-values (in parentheses)

C_D			C_{III}
Overall tests			
17.78 (0.0068, 0.0503)			14.25 (0.0065, 0.0080)
	NE	OY	TY
Dixon's cell-specific tests			
NE	0.00 (1.000, 1.000)	1.90 (0.0576, 0.0509)	−3.59 (0.0003, <0.0001)
OY	1.68 (0.0929, 0.0879)	−1.50 (0.1341, 0.1611)	−0.18 (0.8534, 0.8173)
TY	−2.03 (0.0429, 0.0488)	0.44 (0.6625, 0.7058)	2.16 (0.0305, 0.0291)
Type III cell-specific tests			
NE	−0.38 (0.7037, 0.7060)	1.95 (0.0507, 0.0509)	−3.06 (0.0022, 0.0021)
OY	1.55 (0.1209, 0.1235)	−1.98 (0.0478, 0.0478)	1.14 (0.2564, 0.2538)
TY	−2.27 (0.0234, 0.0239)	0.34 (0.7329, 0.7358)	2.59 (0.0096, 0.0104)

The p-values and test statistics abbreviations are as in Table 3

NE newly emergent, *OY* one year old, *TY* two years old bramble canes

of the coordinates. We fit a linear trend for the log intensity of the Poisson process, and obtain the estimated model as $\lambda(x, y) = \exp(1.87 + 0.48x - 0.37y)$. We check the model fit by the usual χ^2 goodness-of-fit test with a 3×3 grid, and obtain $\chi^2 = 8.94$, $df = 6$, and $p = 0.3529$, suggesting the acceptability of our inhomogeneous Poisson fit. This fit suggests that the intensity of 2 years old canes increase as one moves toward the lower end corner of the study area which explains the segregation of 2 year old canes.

Remark 6.1 In both of the above examples, the more appropriate null hypothesis turned out to be the CSR independence pattern. If it were RL, a different strategy should be followed. For example, under RL, the locations of bramble canes would have been given in advance, and we would have assigned group labels to the points independently and randomly with probabilities proportional to the group sizes. In such a case, the alternative would be a non-RL pattern and the current data set would be a realization of that pattern. Then the model fitting would be concerned with assigning each point with a group label according to this non-RL pattern. In particular, one can perform the following steps: Pick a point, say p_1, randomly

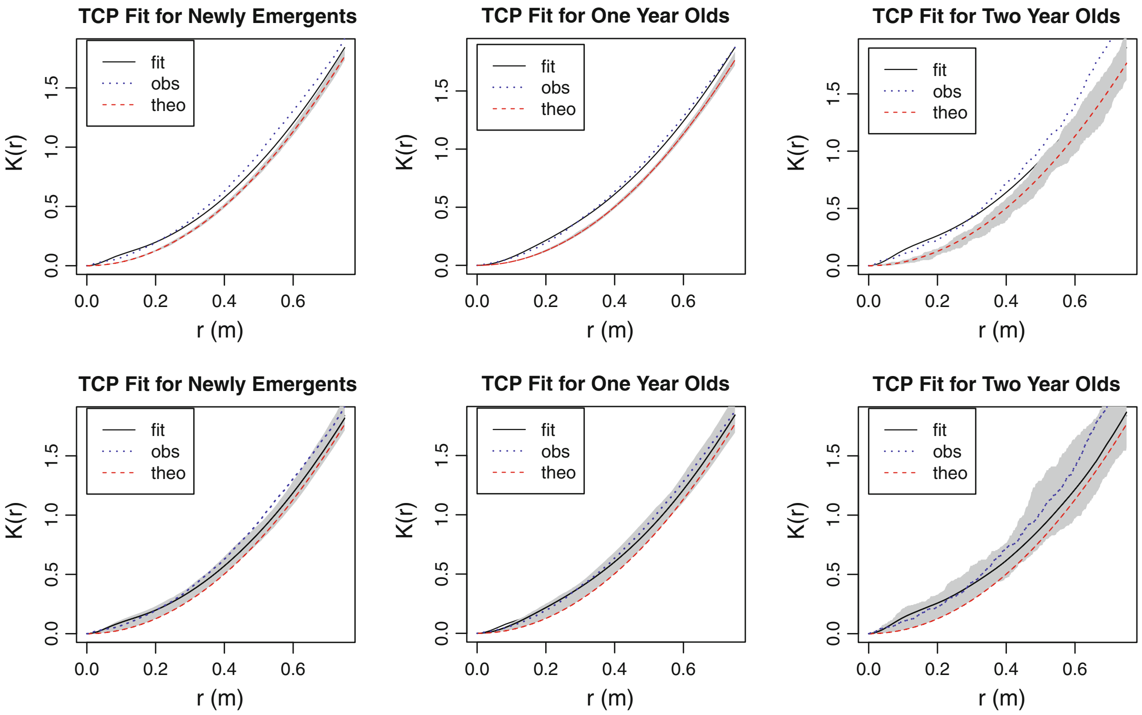

Fig. 22 Plotted at top row are the fitted K-function for the TCP process (fit), the observed K-function with Ripley's isotropic correction for edge effects (obs), and theoretical K-function under HPP (theo) together with a 95 % empirical confidence interval around it for each group in the bramble cane data. Plotted in the bottom row are the same K-functions with a 95 % simulation envelope around the fitted K-function

and assign it with a group label randomly according to the relative frequencies of the groups. Then start labeling the NN of p_1, say p_2, according to the observed percentages in the NNCT in Table 5, and label the NN of p_2, say p_3, in the same manner, and keep going (i.e., assign a label to the NN of p_k, say p_{k+1}, in the same manner). In case of a reflexive NN or if the NN is one of the already labeled points, then start over to execute the same steps among the unlabeled points. The labeling algorithm terminates until each and every point is labeled. With such an algorithmic labeling process, we would obtain non-RL patterns that are of the same type as the observed data. □

7 Discussion and conclusions

In this article, we provide characterizations and various parameterizations of the segregation and association patterns. We first list some appealing properties a segregation or an association pattern should enjoy, and discuss various parameterizations of them in light of these properties. Segregation is relatively easy to parameterize and simulate, but association is not. We propose three types of association patterns, and investigate their properties. Based on these properties and extensive Monte Carlo simulations, these alternatives behave as expected when the alternative parameter depends on the estimated intensity in the study region under the null case. In particular, in these alternatives the association parameter could be $1/(k\hat{\rho})$ with the appropriate choice of k (usually $k \geq 2$ would work) where $\widehat{\rho}$ is the estimated intensity of one or all classes in question under H_o (see Sect. 3.3.1)

In our evaluations of the alternatives, we employ tests based on NNCTs (i.e., NNCT-tests), since in the multi-class case, they provide an omnibus test for any deviation from the null case and then provide various post-hoc tests after the omnibus test is significant (which is analogous to ANOVA F-test and post-hoc t-tests in a multi-group comparison setting). Among the NNCT-tests, type III tests tend to perform better under various alternatives.

We also consider various spatial point processes with respect to homogeneous Poisson process (HPP) or with respect to clustering or regularity patterns to see which patterns cause segregation or association in a two-class setting. In particular, we observed that if one class is from HPP and the other from a regularity pattern, as the level of regularity increases, the level of association between the classes increases; also if one class is from HPP and the other from a clustering pattern, as the level clustering increases, the level segregation between the classes increases as well. Furthermore, if the number of clusters increase, the level of segregation tends to decrease, provided the number of points per cluster is inversely proportional to the number of clusters within the same support. If both classes are from an inhibition/regularity process, then as the level of regularity increases, the level of association between the classes tends to increase. Also if both classes are from a Matérn clustering process with different parents, then as the level of clustering increases, the level of segregation increases as well. On the other hand, if they have the same parent points, then the pattern does not significantly deviate from the null pattern. We also consider the multi-class patterns, with the three-class case as our example. In these cases, we investigate various types of association and mixed patterns, and observe that NNCT-tests provide a good summary of the patterns and describe all types of interaction between each pair of classes when all statistics are simultaneously considered for all cells of the NNCT.

Segregation and association may result when the classes are from dependent or independent patterns which may be regular, CSR, or clustering patterns. The proposed methodology provides some tools and guidelines for understanding which underlying generative pattern might be causing or explaining the observed segregation or association. Along this line, one can take two opposing approaches in practice:

(I) For a given data,

(a) first fit the best model for the generative pattern for each class, and
(b) hypothesize which multi-class pattern might occur based on the fitted patterns and
(c) perform an inferential analysis (e.g., apply NNCT-tests) to formally identify the patterns of segregation/association or lack of them by also attaching significance to the results.

(II) Alternatively, for a given data,

(a) one can start with I(c), i.e., discovering and testing the current multi-class patterns using, say, NNCT-tests, then
(b) s/he can hypothesize (based on the test results and the scatter plots of the classes in the study region) about the type of the generative pattern, and
(c) finally, s/he can fit the best pattern to each class formally and tries interpreting and/or understanding the underlying generative patterns behind the observed segregation or association.

Either approach has some practical appeal and the approach to be taken depends on the particular data set and the research objectives.

Acknowledgments I would like to thank an anonymous associate editor and two referees, whose constructive comments and suggestions greatly improved the presentation and flow of the paper. Most of the Monte Carlo simulations presented in this article were executed at Koç University High Performance Computing Laboratory. This research was supported by the research agency TUBITAK via Project # 111T767 and the European Commission under the Marie Curie International Outgoing Fellowship Programme via Project # 329370 titled PRinHDD.

References

Baddeley AJ, Turner R (2005) spatstat: an R package for analyzing spatial point patterns. J Stat Softw 12(6):1–42

Berman M (1986) Testing for spatial association between a point process and another stochastic process. J R Stat Soc C 35(1):54–62

Ceyhan E (2008) Overall and pairwise segregation tests based on nearest neighbor contingency tables. Comput Stat Data Anal 53(8):2786–2808

Comas C, Palahí M, Pukkala T, Mateu J (2009) Characterising forest spatial structure through inhomogeneous second order characteristics. Stoch Environ Res Risk Assess 23:387–397

Diggle PJ (2003) Statistical analysis of spatial point patterns, 2nd edn. Hodder Arnold Publishers, London

Diggle PJ, Cox TF (1983) Some distance-based tests of independence for sparsely-sampled multivariate spatial point patterns. Int Stat Rev 51:11–23

Dixon PM (1994) Testing spatial segregation using a nearest-neighbor contingency table. Ecology 75(7):1940–1948

Dixon PM (2002a) Nearest-neighbor contingency table analysis of spatial segregation for several species. Ecoscience 9(2):142–151

Dixon PM (2002b) Nearest neighbor methods. In: El-Shaarawi AH, Piegorsch WW (eds) Encyclopedia of environmetrics, vol 3. Wiley, New York, pp 1370–1383

Foxall R, Baddeley AJ (2002) Nonparametric measures of association between a spatial point process and a random set, with geological applications. Appl Stat 51(2):165–182

Gerrard DJ (1969) Competition quotient: a new measure of the competition affecting individual forest trees. Research Bulletin, Agricultural Experiment Station, Michigan State University, 20

Goodall DW (1965) Plot-less tests of interspecific association. J Ecol 53:197–210

Goreaud F, Pélissier R (2003) Avoiding misinterpretation of biotic interactions with the intertype K_{12}-function: population independence vs. random labelling hypotheses. J Veg Sci 14(5): 681–692

Harkness R, Isham V (1983) A bivariate spatial point pattern of ants' nests. Appl Stat 32:293–303

Hutchings MJ (1979) Standing crop and pattern in pure stands of *Mercurialis perennis* and *Rubus fruticosus* in mixed deciduous woodland. Oikos 31:351–357

Illian J, Burslem D (2007) Contributions of spatial point process modelling to biodiversity theory. Coexistence 148(148):9–29

Kulldorff M (2006) Tests for spatial randomness adjusted for an inhomogeneity: a general framework. J Am Stat Assoc 101(475): 1289–1305

Meliker J, Jacquez GM (2007) Space–time clustering of case–control data with residential histories: insights into empirical induction periods, age-specific susceptibility, and calendar year-specific effects. Stoch Environ Res Risk Assess 21(5):625–634

Pielou EC (1961) Segregation and symmetry in two-species populations as studied by nearest-neighbor relationships. J Ecol 49(2): 255–269

Ripley BD (2004) Spatial statistics, 2nd edn. Wiley-Interscience, New York

Searle SR (2006) Matrix algebra useful for statistics. Wiley-Interscience, New York

Stoyan D (1984) On correlations of marked point processes. Math Nach 116:197–207

Uria-Diez J, Ibáñez R, Mateu J (2013) Importance of habitat heterogeneity and biotic processes in the spatial distribution of a riparian herb (*Carex remota* L.): a point process approach. Stoch Environ Res Risk Assess 27:59–76

van Lieshout MNM, Baddeley AJ (1999) Indices of dependence between types in multivariate point patterns. Scand J Stat 26:511–532

Waagepetersen RP (2007) An estimating function approach to inference for inhomogeneous Neyman–Scott processes. Biometrics 63(1):252–258

Waller LA, Zhu L, Gotway CA, Gorman DM, Gruenewald PJ (2007) Quantifying geographic variations in associations between alcohol distribution and violence: a comparison of geographically weighted regression and spatially varying coefficient models. Stoch Environ Res Risk Assess 21:573–588