Journal of Nonparametric Statistics
Vol. 22, No. 5, July 2010, 599–616

Taylor & Francis
Taylor & Francis Group

Directional clustering tests based on nearest neighbour contingency tables

Elvan Ceyhan*

Department of Mathematics, Koç University, 34450 Sarıyer, Istanbul, Turkey

(*Received 5 February 2009; final version received 3 July 2009*)

Spatial interaction between two or more classes or species has important implications in various fields, and might cause multivariate patterns such as segregation or association. Segregation occurs when members of a class or species are more likely to be found near members of the same class or conspecifics; association occurs when members of a class or species are more likely to be found near members of another class or species. The null patterns considered are random labelling and complete spatial randomness (CSR) of points from two or more classes, which is henceforth called *CSR independence*. The clustering tests based on nearest neighbour contingency tables (NNCTs) that are in use in the literature are two-sided tests. In this article, we consider the directional (i.e. one-sided) versions of the cell-specific NNCT tests and introduce new directional NNCT tests for the two-class case. We analyse the distributional properties and compare the empirical significant levels and empirical power estimates of the tests using extensive Monte Carlo simulations. We demonstrate that the new directional tests have comparable performance with the currently available NNCT tests in terms of empirical size and power. We use an ecological data set for illustrative purposes and provide guidelines for using these NNCT tests.

Keywords: association; clustering; complete spatial randomness; independence; random labelling; spatial pattern

AMS Subject Classification: 62H11; 62H17; 62H30

1. Introduction

Spatial point patterns have important implications in epidemiology, population biology, ecology, and other fields, and have been extensively studied. Most of the research on spatial patterns from the early days pertains to a pattern of one type of point with respect to the ground (e.g. density, clumpiness, etc.). These patterns for only one type of point usually fall under the pattern category called *spatial aggregation* (i.e. *clustering*) (Coomes, Rees and Turnbull 1999) or *regularity*. However, it is also of practical interest to investigate the spatial interaction of one type of point with other types (Pielou 1961). The spatial relationships among two or more types of points have interesting consequences, especially for plant species. See, for example, Pielou (1961), Pacala (1986), and Dixon (1994, 2002a,b). For convenience and generality, we refer to the different types

*Email: elceyhan@ku.edu.tr

ISSN 1048-5252 print/ISSN 1029-0311 online

DOI: 10.1080/10485250903199861
http://www.informaworld.com

of points as 'classes', but *class* can stand for any characteristic of an individual at a particular location. For example, the spatial segregation pattern has been investigated for *plant species* (Whipple 1980; Diggle 2003), *fish species* (Herler and Patzner 2005), and *sexes* of dioecious plants (Nanami, Kawaguchi and Yamakura 1999). Many of the epidemiologic applications are for a two-class system of case and control labels (Waller and Gotway 2004).

Many univariate and multivariate (i.e. one-class and multi-class) tests have been proposed for testing the segregation of two classes in statistical and other literature (Kulldorff 2006). These include comparison of Ripley's $K(t)$ functions (Diggle and Chetwynd 1991), comparison of NN distances (Diggle 2003), and nearest neighbour contingency tables (NNCTs) (Pielou 1961; Dixon 1994). Pielou (1961) proposed various tests based on NNCTs for the two-class case only and Dixon (1994) introduced an overall test of segregation and class-specific tests based on NNCTs for the two-class case and extended his tests to the multi-class case (Dixon 2002a,b). For the two-class case, Ceyhan (n.d.) discussed these tests and demonstrated that Pielou's test is liberal under complete spatial randomness (CSR) independence or random labelling (RL) and is only appropriate for a random sample of (base, NN) pairs. If v is an NN of point u, then u is called the *base point* and v is called the *NN point*. Furthermore, Ceyhan (2008a) proposed new cell-specific and overall segregation tests which are more robust to the differences in the relative abundance of classes and have better performance in terms of size and power.

In the literature, most segregation tests are two-sided tests for the two-class case or against a general alternative for the multi-class case. In particular, the NNCT tests in literature are not directional tests. In this article, we investigate the directional (i.e. one-sided) versions of the cell-specific tests of Dixon (1994, 2002a) and Ceyhan (2008a), where although it is pointed out that these tests can be used for one-sided alternatives, their performances are only assessed for the two-sided alternatives. We also introduce new directional segregation tests. We compare these NNCT tests in terms of distributional properties, empirical size, and power through extensive Monte Carlo simulations. We also compare these tests with Ripley's K or L-functions (Ripley 2004) and pair correlation function $g(t)$ (Stoyan and Stoyan 1994), which are methods for second-order analysis of point patterns. We show through simulation that the newly proposed directional tests perform similarly to, but slightly better in power than, the cell-specific tests of Ceyhan (2008a) and perform better in terms of empirical size and power than Dixon's cell-specific tests. Furthermore, we demonstrate that our tests and Ripley's L-function and related methods (i.e. second-order analysis) answer different questions about the pattern of interest.

We describe the null and alternative patterns and the construction of the NNCTs in Section 2, discuss Dixon's and Ceyhan's cell-specific segregation tests in Section 3.1, and introduce a directional version of Pielou's test of segregation in Section 3.2, and new directional tests in Section 3.3. We provide the empirical significance level analysis under CSR independence in Section 4, empirical power analysis under the alternatives in Section 5, illustrate the tests in example data sets in Section 6, and provide discussion and guidelines for using the tests in Section 7.

2. Types of spatial patterns and NNCTs

Consider a plot that contains two plant species, and the plants from each species are distributed according some point process. We are interested in the spatial interaction between these two species as to whether the plant species repel each other (i.e. are segregated), attract each other (i.e. are associated), or have no spatial interaction (e.g. exhibit CSR independence or RL). We try to answer these questions using NNCT tests and other methods currently available in the literature.

2.1. *Null and alternative patterns*

There are two benchmark hypotheses to investigate the spatial interaction between multiple classes in a multivariate process: (i) *independence*, which implies that the classes of points are generated by independent univariate processes and (ii) RL, which implies that the class labels are randomly assigned to a given set of locations in the region of interest (Diggle 2003). In this article, our null hypothesis is

$$H_o : \textit{randomness in the NN structure},$$

which might result from two random pattern types: CSR of points from two classes (called *CSR independence*, henceforth) or RL. Under CSR independence, points from each of the two classes independently satisfy CSR in the region of interest.

Although CSR independence and RL are not the same, they lead to the same null model for NNCT tests, since an NNCT does not require spatially explicit information. That is, when the points from two classes are assumed to be independently uniformly distributed over the region of interest, i.e. under the CSR independence pattern, or when only the labelling (or marking) of a set of fixed points (where the allocation of the points might be regular, aggregated, or clustered, or of lattice type) is considered, i.e. under the RL pattern, there is *randomness in the NN structure*. The distinction between CSR independence and RL is very important when defining the appropriate null model in practice. Goreaud and Pélissier (2003) state that CSR independence implies that the two classes are *a priori* the result of different processes (e.g. individuals of different species or age cohorts), whereas RL implies that some processes affect *a posteriori* the individuals of a single population (e.g. diseased *vs*. non-diseased individuals of a single species). We provide the differences in the proposed tests under these two patterns. For a more detailed discussion of CSR independence and RL, see Ceyhan (2008b).

As clustering alternatives, we consider two major types of spatial patterns: *segregation* and *association*. *Segregation* occurs if the NN of an individual (with respect to the L^2-norm) is more likely to be of the same class as the individual than to be from a different class; i.e. the members of the same class tend to be clumped or clustered (Dixon 1994; Kulldorff 2006). For instance, one type of plant might not grow well around another type of plant and vice versa. In plant biology, one class of points might represent the coordinates of trees from a species with large canopy, so that other plants (whose coordinates are the other class of points) that need light cannot grow (well or at all) around these trees. In epidemiology, one class of points might be the geographical coordinates of residences of cases and the other class of points might be the coordinates of the residences of controls. *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. For example, in plant biology, the two classes of points might represent the coordinates of mutualistic plant species, so the species depend on each other to survive. As another example, the points from one class might be the geometric coordinates of parasitic plants exploiting the other plant whose coordinates are the points of the other class.

The patterns of segregation and association result not only from multivariate interaction between the classes. It is also conceivable to have either of these patterns without any interaction between the point processes; for example, consider the case where species happen to have the same or different fine-scale habitat preferences. Each of the two patterns of segregation and association are not symmetric in the sense that, when two classes are segregated (or associated), they do not necessarily exhibit the same degree of segregation (or association). For example, when points from each of two classes labelled as X and Y are clustered at different locations, but class X is loosely clustered (i.e. its point intensity in the clusters is smaller) compared with class Y so that classes X and Y are segregated but class Y is more segregated than class X. Similarly, when class Y points are clustered around class X points but not vice versa, classes Y and X are

Table 1. NNCT for two classes.

		NN class		
		Class 1	Class 2	Sum
Base class	Class 1	N_{11}	N_{12}	n_1
	Class 2	N_{21}	N_{22}	n_2
	Sum	C_1	C_2	n

associated, but class Y is more associated with class X compared with the other way around. Although it is not possible to list all of the many different types of segregation (and association), its existence can be tested by an analysis of the NN relationships between the classes (Pielou 1961).

2.2. *Nearest neighbour contingency tables*

NNCTs are constructed using the NN frequencies of classes. Consider two classes with labels $1, 2$ which stand for classes X and Y, respectively. Let n_i be the number of points from class i for $i \in \{1, 2\}$ and n the total sample size. If the class of each point and the class of its NN were recorded, the NN relationships fall into four distinct categories: $(1, 1)$, $(1, 2)$, $(2, 1)$, $(2, 2)$ where in cell (i, j), class i is the *base class*, while class j is the class of its *NN*. That is, the n points constitute n (base,NN) pairs. Denoting N_{ij} as the frequency of cell (i, j) (i.e. the count of all (base,NN) pairs each of which has label (i, j)) for $i, j \in \{1, 2\}$ yields the NNCT in Table 1 where the column sum C_j is the number of times class j points serve as NNs for $j \in \{1, 2\}$. Under segregation, the diagonal entries N_{ii} for $i = 1, 2$, tend to be larger than expected; under association, the off-diagonals tend to be larger than expected. The general alternative is that some cell counts are different than expected under CSR independence or RL.

Pielou (1961) suggested the use of the Pearson's χ^2 test of independence for NNCTs, but her test has been shown to be inappropriate (Meagher and Burdick 1980). The main problem with her test is the violation of the independence between cell-counts (and rows and columns also) under CSR independence or RL (Dixon 1994; Ceyhan n.d.). Dixon (1994) derived the appropriate (asymptotic) sampling distribution of cell counts under RL, and Ceyhan (n.d.) demonstrated that Dixon's test is also appropriate for CSR independence; all these tests are consistent, in the sense that under any alternative (of segregation or association), the power tends to one, as sample sizes tend to infinity. While Dixon's test has the appropriate nominal size under CSR independence, Pielou's test is liberal (Ceyhan n.d.).

3. Directional segregation tests based on NNCTs

3.1. *Cell-specific tests of segregation*

Dixon (1994) proposed a series of tests for segregation based on NNCTs. In Dixon's framework, the probability of a class j point serving as an NN of a class i point depends only on the class sizes (row sums), but not the total number of times class j serves as an NN (column sums). Dixon demonstrated that under RL, one can write down the cell frequencies as Moran join count statistics (Moran 1948). He then derived the means, variances, and covariances of the cell counts (frequencies) (Dixon 1994, 2002a,b).

Under CSR independence or RL, we have

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

where n_i is the sample size for class i. Furthermore,

$$\mathbf{Var}[N_{ij}] = \begin{cases} (n+R)p_{ii} + (2n-2R+Q)p_{iii} + (n^2-3n-Q+R)p_{iiii} - (np_{ii})^2, & \text{if } i=j, \\ np_{ij} + Qp_{iij} + (n^2-3n-Q+R)p_{iijj} - (np_{ij})^2, & \text{if } i \neq j \end{cases} \tag{2}$$

with p_{xx}, p_{xxx}, and p_{xxxx} are the probabilities that a randomly picked pair, triplet, or quartet of points, respectively, are the indicated classes and are given by $p_{ii} = (n_i(n_i-1))/(n(n-1))$, $p_{ij} = (n_i n_j)/(n(n-1))$, $p_{iii} = (n_i(n_i-1)(n_i-2))/(n(n-1)(n-2))$, $p_{iij} = (n_i(n_i-1)n_j)/(n(n-1)(n-2))$, $p_{iijj} = (n_i(n_i-1)n_j(n_j-1))/(n(n-1)(n-2)(n-3))$, and $p_{iiii} = (n_i(n_i-1)(n_i-2)(n_i-3))/(n(n-1)(n-2)(n-3))$. Furthermore, Q is the number of points with shared NNs, which occur when two or more points share an NN and R is twice the number of reflexive pairs. A (base,NN) pair (u, v) is *reflexive* if (v, u) is also a (base,NN) pair. Then $Q = 2(Q_2 + 3Q_3 + 6Q_4 + 10Q_5 + 15Q_6)$ where Q_k is the number of points that serve as an NN to other points k times.

The test statistic suggested by Dixon is given by

$$Z_{ij}^{\mathrm{D}} = \frac{N_{ij} - \mathbf{E}[N_{ij}]}{\sqrt{\mathbf{Var}[N_{ij}]}}, \tag{3}$$

where $\mathbf{E}[N_{ij}]$ is given in Equation (1) and $\mathbf{Var}[N_{ij}]$ is given in Equation (2).

Ceyhan (2008a) suggested the following cell-specific test statistic:

$$T_{ij} = \begin{cases} N_{ii} - \dfrac{n_i - 1}{n-1} C_i, & \text{if } i=j, \\ N_{ij} - \dfrac{n_i}{n-1} C_j, & \text{if } i \neq j. \end{cases}$$

Then $\mathbf{E}[T_{ij}] = 0$ and the variance of T_{ij} is

$$\mathbf{Var}[T_{ij}] = \begin{cases} \mathbf{Var}[N_{ii}] + \dfrac{(n_i-1)^2}{(n-1)^2}\mathbf{Var}[C_i] - 2\dfrac{(n_i-1)}{(n-1)}\mathbf{Cov}[N_{ii}, C_i], & \text{if } i=j, \\ \mathbf{Var}[N_{ij}] + \dfrac{n_i^2}{(n-1)^2}\mathbf{Var}[C_j] - 2\dfrac{n_i}{n-1}\mathbf{Cov}[N_{ij}, C_j], & \text{if } i \neq j, \end{cases}$$

where $\mathbf{Var}[N_{ij}]$ are as in Equation (2), $\mathbf{Var}[C_j] = \sum_{i=1}^{q} \mathbf{Var}[N_{ij}] + \sum_{k\neq i}\sum_i \mathbf{Cov}[N_{ij}, N_{kj}]$ and $\mathbf{Cov}[N_{ij}, C_j] = \sum_{k=1}^{q} \mathbf{Cov}[N_{ij}, N_{kj}]$ with $\mathbf{Cov}[N_{ij}, N_{kl}]$ are as in Equations (4)–(12) of Dixon (2002a).

The proposed cell-specific test in standardised form is

$$Z_{ij}^{\mathrm{C}} = \frac{T_{ij}}{\sqrt{\mathbf{Var}[T_{ij}]}}. \tag{4}$$

THEOREM 3.1 *In a* 2×2 *NNCT,* $Z_{ij}^{\mathrm{C}} \xrightarrow{\mathcal{L}} N(0,1)$ *for all* $i, j \in \{1, 2\}$ *as* $n_1, n_2 \to \infty$ (*i.e., as* $\min(n_1, n_2) \to \infty$) *where* $\xrightarrow{\mathcal{L}}$ *stands for convergence in law.*

Proof In the two-class case, each diagonal cell count N_{ii}, for $i = 1, 2$, has an asymptotic normal distribution (Dixon 1994). Since n_1 and n_2 are fixed and $N_{12} = n_1 - N_{11}$ and $N_{21} = n_2 - N_{22}$, both N_{12} and N_{21} also converge in law to the normal distribution also. Then, we have

$$\frac{N_{ij} - \mathcal{E}[N_{ij}]}{\sqrt{\mathbf{Var}[N_{ij}]}} \xrightarrow{\mathcal{L}} N(0, 1)$$

for all i, j in the 2×2 case.

Let $\vec{C} = (N_{11}, N_{21})'$ where $'$ stands for transpose of a vector. Then for large n_1, n_2, $\vec{C} \overset{\text{approx}}{\sim} N((\mathbf{E}[N_{11}], \mathbf{E}[N_{21}])', \Sigma_{11})$, where

$$\Sigma_{11} = \begin{pmatrix} \mathbf{Var}[N_{11}] & \mathbf{Cov}[N_{11}, N_{21}] \\ \mathbf{Cov}[N_{11}, N_{21}] & \mathbf{Var}[N_{21}] \end{pmatrix}$$

Now consider

$$T_{11} = N_{11} - \frac{n_1 - 1}{n - 1}(N_{11} + N_{21}) = \frac{n_2}{n - 1}N_{11} - \frac{n_1 - 1}{n - 1}N_{21}.$$

Then $T_{11} = (n_2/(n-1), -(n_1 - 1)/(n-1))\vec{C} \overset{approx}{\sim} N(0, \mathbf{Var}[T_{11}] = (n_2/(n-1)), -((n_1 - 1)/(n-1))\Sigma_{11}(n_2/(n-1), -(n_1 - 1)/(n-1))')$ for large n_1 and n_2 (Anderson 1984; Bickel 1974). Hence, Z_{11}^{C} also converges in law to $N(0, 1)$ as $n_1, n_2 \to \infty$. The asymptotic normality of Z_{12}^{C}, Z_{21}^{C}, and Z_{22}^{C} can be shown similarly. ■

3.2. *Directional version of Pielou's test of segregation*

To detect any deviation from randomness in the NN structure, Pielou (1961) used Pearson's χ^2 test of independence, $\mathcal{X}_{\mathrm{P}}^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}(N_{ij} - \mathbf{E}[N_{ij}])^2/\mathbf{E}[N_{ij}]$, where $\mathbf{E}[N_{ij}] = (n_i\, C_j)/n$ with C_j being the observed sum for column j.

Pielou's test, when used for an NNCT based on a random sample of (base, NN) pairs, measures deviations from the independence of cell counts, but does not indicate the direction of the deviation (e.g. segregation or association). To determine the direction, one needs to check the NNCT. Pielou's test is used only for the two-sided alternatives in the two-class case in literature. We apply the conversion of Pearson's χ^2 test to a one-sided test in the usual 2×2 contingency tables to Pielou's test for NNCTs. Since $\mathcal{X}_{\mathrm{P}}^2 \overset{\text{approx}}{\sim} \chi_1^2$, for large n, we can write $\mathcal{X}_{\mathrm{P}}^2 = Z_{\mathrm{P}}^2$ where $Z_{\mathrm{P}} \overset{\text{approx}}{\sim} N(0, 1)$, where $N(0, 1)$ stands for the standard normal distribution. By some algebraic manipulations, among other possibilities, Z_{P} can be written as

$$Z_{\mathrm{P}} = \left(\frac{N_{11}}{n_1} - \frac{N_{21}}{n_2}\right)\sqrt{\frac{n_1 n_2 n}{C_1 C_2}}.$$

See Bickel and Doksum (1977) for the sketch of the derivation. For example, Z_{P} could also be written as $Z_{\mathrm{P}} = (N_{11} - (n_1\, C_1/n))[n_1 n_2 C_1 C_2/n^3]^{-1/2}$. These directional tests are not appropriate for testing CSR independence or RL, due to the inherent dependence of cell counts in NNCTs on such patterns, but are only appropriate for NNCTs based on a random sample of (base,NN) pairs.

*Remark 1 (*Empirical correction for Pielou's and related NNCT tests) Based on extensive Monte Carlo simulation results, the one-sided version of Pielou's test statistic Z_{P} can be empirically corrected so that the transformed statistic will be approximately distributed as an $N(0, 1)$ distribution (Ceyhan 2008c). Since the empirical mean for Z_{P}, although tending to

zero for large n, is large for small samples, we also consider $Z_{\rm P}^{\rm A} = Z_{\rm P}\mathbf{I}(Z_{\rm P} \le 0)$ for the association, and $Z_{\rm P}^{\rm S} = Z_{\rm P}\mathbf{I}(Z_{\rm P} \ge 0)$ for the segregation alternatives separately, where $\mathbf{I}(\cdot)$ stands for the indicator function. As we use the critical values based on a standard normal distribution, after adjusting we want $Z_{\rm P} \sim Z$, $Z_{\rm P}^{\rm A} \sim Z^-$ and $Z_{\rm P}^{\rm S} \sim Z^+$, where $Z^- = Z\,\mathbf{I}(Z \le 0)$ and $Z^+ = Z\,\mathbf{I}(Z \ge 0)$ with $Z \sim N(0,1)$. Therefore $Z_{\rm P}$ does not need adjusting for location (we take the mean of $Z_{\rm P}$ to be 0), but needs an adjustment in scale for variance. We transform the $Z_{\rm P}$ scores as $Z_{mc} := Z_{\rm P}/\delta_n$ and $Z_{\rm P}^{\rm A}$ and $Z_{\rm P}^{\rm S}$ scores by adjusting for location and scaling as $Z_{mc}^{\rm A} := (Z_{\rm P}^{A} - \gamma_a)/\delta_a$ and $Z_{mc}^{\rm S} := (Z_{\rm P}^{\rm S} - \gamma_s)/\delta_s$, where $\delta_n = 1.277$, $\delta_a = 1.307$, $\gamma_a = 0.043$ and $\delta_s = 1.275$, and $\gamma_s = -0.057$. See Ceyhan (2008c) for their derivation.

3.3. *New directional tests of segregation*

We introduce two new directional tests motivated by $Z_{\rm P}$. Let $T_n := (N_{11}/n_1) - (N_{21}/n_2)$ and $U_n := \sqrt{(n_1 n_2)/(C_1 C_2)}$, then $Z_{\rm P} = \sqrt{n} U_n T_n$. Note that $\mathbf{E}[T_n] = ((\mathbf{E}[N_{11}]/n_1) - (\mathbf{E}[N_{21}]/n_2)) = -1/(n-1)$. Using the asymptotic normality of cell counts N_{11} and N_{21}, we have

$$\left(\frac{T_n - \mathbf{E}[T_n]}{\sqrt{\mathbf{Var}[T_n]}}\right) \xrightarrow{\mathcal{L}} N(0,1),$$

where

$$\mathbf{Var}[T_n] = \frac{\mathbf{Var}[N_{11}]}{n_1^2} + \frac{\mathbf{Var}[N_{21}]}{n_2^2} - \frac{2\,\mathbf{Cov}[N_{11}, N_{21}]}{n_1 n_2}$$

with $\mathbf{Var}[N_{ij}]$ are as in Equation (2) and

$$\mathbf{Cov}[N_{11}, N_{21}] = (n - R + Q)\, p_{112} + (n^2 - 3n - Q + R) p_{1112} - n^2 p_{11} p_{12}$$

(see Dixon (2002a) for the derivation).

We propose two tests based on T_n:

$$\text{(i)}\ \ Z_{\rm I} = U_n \left(\frac{T_n - \mathbf{E}[T_n]}{\sqrt{\mathbf{Var}[T_n]}}\right), \quad \text{(ii)}\ \ Z_{\rm II} = \frac{T_n - \mathbf{E}[T_n]}{\sqrt{\mathbf{Var}[T_n]}}. \tag{5}$$

THEOREM 3.2 *In a 2×2 NNCT, $Z_{\rm I}$ and $Z_{\rm II}$ converge in law to $N(0,1)$ distribution as $n_1, n_2 \to \infty$.*

Proof In the two-class case, the asymptotic normality of $Z_{\rm II}$ can be shown as in the proof of Theorem 3.1.

Let ν_i be the proportion of points from class i in the population for $i = 1, 2$. Then $\lim_{n, n_i \to \infty} n_i/n = \nu_i$; and under CSR independence or RL, $C_i/n \xrightarrow{p} \nu_i$ as $n, n_i \to \infty$ for $i = 1, 2$. So, $U_n \xrightarrow{p} 1$ as $n, n_i \to \infty$. Then by Slutsky's theorem and asymptotic normality of $Z_{\rm II}$, we have the asymptotic normality of $Z_{\rm I}$. ■

Note that $Z_{\rm I}$ and $Z_{\rm II}$ values are positive under segregation and negative under association alternatives. Furthermore, as can be seen in the proof of Theorem 3.2, they are asymptotically equivalent. Note also that T_n and T_{ij} values are both linear combinations of cell counts in columns of the NNCT.

Remark 2 (The status of Q and R under CSR independence and RL) Q and R are fixed under RL, but random under CSR independence. The quantities given in Equations (1) and (2), and all the quantities depending on these expectations also depend on Q and R. Hence, these expressions

are appropriate under the RL model. Under the CSR independence model they are conditional – on Q and R – variances and covariances. Hence, under the CSR independence pattern, the asymptotic distributions of the tests in Equations (3)–(5) are conditional on Q and R.

The unconditional variances and covariances can be obtained by replacing Q and R with their expectations (Ceyhan 2009). Unfortunately, given the difficulty of calculating the expectation of Q under CSR independence, it is reasonable and convenient to use test statistics employing the conditional variances and covariances even when assessing their behaviour under the CSR independence model. Cox (1981) calculated analytically that $\mathbf{E}[R|N = n] = 0.6215n$ for a planar Poisson process. Alternatively, one can estimate the expected values of Q and R empirically. For example, for a homogeneous planar Poisson pattern, we have $\mathbf{E}[Q|N = n] \approx 0.6328n$ and $\mathbf{E}[R|N = n] \approx 0.6211n$ (estimated empirically by 1,000,000 Monte Carlo simulations for various values of N on unit square). Note that $\mathbf{E}[R|N = n]$ agrees with the analytical result of Cox (1981). When Q and R are replaced by $0.63\,n$ and $0.62\,n$, respectively, the so-called *QR-adjusted* tests are obtained. However, QR-adjustment does not improve on the unadjusted NNCT tests (Ceyhan 2008d).

Remark 3 (asymptotic structures for the NNCT tests) There are two major types of asymptotic structures for spatial data (Lahiri 1996). In the first, any two observations are required to be at least a fixed distance apart, hence as the number of observations increase, the region on which the process is observed eventually becomes unbounded. This type of sampling structure is called 'increasing domain asymptotics'. In the second type, the region of interest is a fixed bounded region and more and more points are observed in this region. Hence, the minimum distance between data points tends to zero as the sample size tends to infinity. This type of structure is called 'infill asymptotics', due to Cressie (1993). The sampling structure in our asymptotic sampling distribution could be either one of infill or increasing domain asymptotics, as we only consider the class sizes and the total sample size tending to infinity regardless of the size of the study region. The relation of CSR independence and RL with these asymptotic structures is discussed in more detail in Ceyhan (2009).

4. Empirical significance levels under CSR independence

We simulate the CSR independence pattern by generating n_1 points from class X and n_2 points from class Y, both of which are uniformly distributed on the unit square $(0, 1) \times (0, 1)$ for some combinations of n_1 and n_2.

We present the empirical significance levels of the tests for the right-sided alternative (i.e. with respect to the segregation alternative) in Table 2 and for the left-sided alternative (i.e. with respect to the association alternative) in Table 3. The sizes significantly smaller (larger) than 0.05 are marked with ^{c} ($^\ell$), which indicate that the corresponding test is conservative (liberal). The asymptotic normal approximation to proportions is used in determining the significance of the deviations of the empirical sizes from 0.05. For these proportion tests, we also use $\alpha = 0.05$ as the significance level. With $N_{\text{mc}} = 10{,}000$, empirical sizes less than 0.0464 are deemed conservative, and those greater than 0.0536 are deemed liberal at $\alpha = 0.05$ level.

For the segregation alternative, the directional version of Pielou's test is liberal in rejecting the null hypothesis; the empirically corrected version of Pielou's test as in Remark 1 is about the desired level for larger samples, but is conservative or liberal for smaller samples. The size performance of Dixon's cell-specific test for cell $(1, 1)$ at (n_1, n_2) is similar to that for cell $(2, 2)$ at (n_2, n_1). On the other hand, at each (n_1, n_2) the size performance of Ceyhan's cell-specific test for cell $(1, 1)$ is similar to that for cell $(2, 2)$. Dixon's cell-specific tests are usually liberal, in particular for the smaller sample for different relative abundance cases. Ceyhan's cell-specific tests are liberal when $n_i \leq 30$ for both $i = 1, 2$ or when the relative abundances are very different.

Table 2. The empirical significance levels of the tests for the segregation (right-sided) alternatives under H_o : CSR *independence* with $N_{\text{mc}} = 10{,}000$, n_1, n_2 in $\{10, 30, 50, 100\}$ at $\alpha = 0.05$.

Sizes	Empirical significance levels of the tests for the segregation (i.e. right-sided) alternatives							
(n_1, n_2)	$\hat{\alpha}^{\text{D}}_{1,1}$	$\hat{\alpha}^{\text{D}}_{2,2}$	$\hat{\alpha}^{\text{C}}_{1,1}$	$\hat{\alpha}^{\text{C}}_{2,2}$	$\hat{\alpha}^{\text{P}}_{Z}$	$\hat{\alpha}^{P,Z}_{\text{mc}}$	$\hat{\alpha}_{\text{I}}$	$\hat{\alpha}_{\text{II}}$
(10,10)	0.0515	0.0489	0.0491	0.0491	0.0844^{ℓ}	0.0422^{c}	0.0526	0.0499
(10,30)	0.0960^{ℓ}	0.0468	0.0631^{ℓ}	0.0643^{ℓ}	0.0846^{ℓ}	0.0576^{ℓ}	0.0613^{ℓ}	0.0651^{ℓ}
(10,50)	0.0936^{ℓ}	0.0435^{c}	0.0684^{ℓ}	0.0677^{ℓ}	0.0947^{ℓ}	0.0548^{ℓ}	0.0693^{ℓ}	0.0678^{ℓ}
(30,10)	0.0430^{c}	0.0900^{ℓ}	0.0571^{ℓ}	0.0567^{ℓ}	0.0760^{ℓ}	0.0511	0.0545^{ℓ}	0.0575^{ℓ}
(30,30)	0.0490	0.0530	0.0556^{ℓ}	0.0555^{ℓ}	0.0803^{ℓ}	0.0557^{ℓ}	0.0557^{ℓ}	0.0555^{ℓ}
(30,50)	0.0652^{ℓ}	0.0479	0.0482	0.0484	0.0792^{ℓ}	0.0445^{c}	0.0479	0.0480
(50,10)	0.0441^{c}	0.0915^{ℓ}	0.0655^{ℓ}	0.0665^{ℓ}	0.0955^{ℓ}	0.0531	0.0682^{ℓ}	0.0655^{ℓ}
(50,30)	0.0492	0.0664^{ℓ}	0.0515	0.0509	0.0829^{ℓ}	0.0468	0.0511	0.0511
(50,50)	0.0577^{ℓ}	0.0546^{ℓ}	0.0514	0.0509	0.0804^{ℓ}	0.0421^{c}	0.0526	0.0522
(50,100)	0.0571^{ℓ}	0.0464	0.0509	0.0508	0.0921^{ℓ}	0.0495	0.0524	0.0508
(100,50)	0.0434^{c}	0.0584^{ℓ}	0.0499	0.0500	0.0909^{ℓ}	0.0483	0.0512	0.0498
(100,100)	0.0515	0.0500	0.0485	0.0485	0.0927^{ℓ}	0.0484	0.0485	0.0485

Note: $\hat{\alpha}^{\text{D}}_{i,i}$ and $\hat{\alpha}^{\text{C}}_{i,i}$ are the empirical significance levels of Dixon's and Ceyhan's cell-specific tests for cell $(i, i), i = 1, 2$, respectively, $\hat{\alpha}^{\text{P}}_{Z}$ is for the directional version of Pielou's test Z_{P}, $\hat{\alpha}^{P,Z}_{\text{mc}}$ is for the empirically corrected version of Z_{P}, $\hat{\alpha}_{\text{I}}$ and $\hat{\alpha}_{\text{II}}$ are for the new directional tests provided in Equation (5).
^{c} The empirical size is significantly smaller than 0.05; i.e. the test is conservative.
$^{\ell}$ The empirical size is significantly larger than 0.05; i.e. the test is liberal.

Table 3. The empirical significance levels of the tests for the association (left-sided) alternatives under H_o : CSR *independence* with $N_{\text{mc}} = 10{,}000$, n_1, n_2 in $\{10, 30, 50, 100\}$ at $\alpha = 0.05$.

Sizes	Empirical significance levels of the tests for the segregation (i.e. right-sided) alternatives							
(n_1, n_2)	$\hat{\alpha}^{\text{D}}_{1,1}$	$\hat{\alpha}^{\text{D}}_{2,2}$	$\hat{\alpha}^{\text{C}}_{1,1}$	$\hat{\alpha}^{\text{C}}_{2,2}$	$\hat{\alpha}^{\text{P}}_{Z}$	$\hat{\alpha}^{P,Z}_{\text{mc}}$	$\hat{\alpha}_{\text{I}}$	$\hat{\alpha}_{\text{II}}$
(10,10)	0.0412^{c}	0.0455^{c}	0.0467	0.0454^{c}	0.1574^{ℓ}	0.0858^{ℓ}	0.0484	0.0425^{c}
(10,30)	0.0000^{c}	0.0490	0.0342^{c}	0.0362^{c}	0.1399^{ℓ}	0.0600^{ℓ}	0.0296^{c}	0.0362^{c}
(10,50)	0.0000^{c}	0.0484	0.0057^{c}	0.0087^{c}	0.0574^{ℓ}	0.0006^{c}	0.0006^{c}	0.0086^{c}
(30,10)	0.0494	0.0000^{c}	0.0333^{c}	0.0319^{c}	0.1406^{ℓ}	0.0556^{ℓ}	0.0274^{c}	0.0332^{c}
(30,30)	0.0450^{c}	0.0430^{c}	0.0504	0.0505	0.1115^{ℓ}	0.0537^{ℓ}	0.0505	0.0505
(30,50)	0.0611^{ℓ}	0.0564	0.0494	0.0494	0.1172^{ℓ}	0.0600^{ℓ}	0.0493	0.0495
(50,10)	0.0545	0.0000^{c}	0.0080^{c}	0.0058^{c}	0.0544^{ℓ}	0.0004^{c}	0.0004^{c}	0.0079^{c}
(50,30)	0.0520	0.0594^{ℓ}	0.0475	0.0479	0.1173^{ℓ}	0.0572^{ℓ}	0.0467	0.0477
(50,50)	0.0486	0.0494	0.0503	0.0500	0.1041^{ℓ}	0.0580^{ℓ}	0.0522	0.0517
(50,100)	0.0548^{ℓ}	0.0491	0.0487	0.0491	0.1090^{ℓ}	0.0534	0.0475	0.0486
(100,50)	0.0485	0.0515	0.0465	0.0464	0.1063^{ℓ}	0.0515	0.0453^{c}	0.0464
(100,100)	0.0478	0.0493	0.0475	0.0475	0.1092^{ℓ}	0.0592^{ℓ}	0.0476	0.0475

Note: The labelling and superscripting are as in Table 2.

The size performance of the new directional tests is similar to that of Ceyhan's cell-specific tests. Ceyhan's cell-specific tests and the new directional tests are more robust to differences in relative abundance of the classes.

For the association alternative, the directional version of Pielou's test is extremely liberal for all sample size combinations. The empirically corrected version is still liberal for most sample sizes and extremely conservative for $(10, 50)$ and $(50, 10)$ cases. When the relative abundances of classes are very different (see $(10, 50)$ and $(50, 10)$ cases), both tests are severely affected, but the corrected version is extremely conservative. For small samples ($n_i \leq 30$), Dixon's cell-specific tests are extremely conservative for the cell associated with the smaller sample when the relative abundances are very different. For large samples, Dixon's cell-specific tests are liberal for the cell associated with the smaller sample when the relative abundances are very different. Ceyhan's cell-specific tests are extremely conservative when both samples are small and are about

the desired level for large samples. New tests' size performance is similar to that of Ceyhan's tests. Furthermore, the effect of the differences in relative abundances is most severe on Dixon's cell-specific tests.

The empirical size performance for the two-sided alternatives is similar to the left- and right-sided alternatives (Ceyhan 2008c).

Remark 4 (edge correction for NNCT tests) Edge effects are a constant problem in the analysis of empirical (i.e. bounded) data sets, and much effort has gone into the development of edge correction methods (Yamada and Rogersen 2003). The CSR independence pattern assumes that the study region is unbounded for the analysed pattern, which is not the case in practice. So the edge (or boundary) effects might confound the test results if the null pattern is the CSR independence. Two correction methods for the edge effects on NNCT tests, namely buffer zone correction and toroidal correction, are investigated in Ceyhan (n.d., 2007, 2008e), where it is recommended that inner or outer buffer zone correction for NNCT tests could be used with the width of the buffer area being about the average NN distance. But larger buffer areas are not recommended since they are wasteful, with little additional gain. On the other hand, toroidal edge correction is recommended with points within the average NN distance in the additional copies around the study region, provided that there are no data clusters around the edges. For larger distances, the gain might not be worth the effort. We extend these recommendations for the new directional tests also.

5. Empirical power analysis

We consider three cases for each of segregation and association alternatives. Based on the empirical size estimates provided in Section 4, we omit the directional versions of Pielou's test Z_{P} and its empirically corrected version (see Remark 1) from further consideration.

5.1. *Empirical power analysis under segregation alternatives*

For the segregation alternatives, we generate $X_i \overset{\text{iid}}{\sim} \mathcal{U}((0, 1-s) \times (0, 1-s))$ and $Y_j \overset{\text{iid}}{\sim} \mathcal{U}((s, 1) \times (s, 1))$ for $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$. Note the level of segregation is determined by the magnitude of $s \in (0, 1)$. We consider the following three segregation alternatives:

$$H_{\mathrm{S}}^{\mathrm{I}} : s = \frac{1}{6}, \quad H_{\mathrm{S}}^{\mathrm{II}} : s = \frac{1}{4}, \quad \text{and} \quad H_{\mathrm{S}}^{\mathrm{III}} : s = \frac{1}{3}.$$

Observe that, from $H_{\mathrm{S}}^{\mathrm{I}}$ to $H_{\mathrm{S}}^{\mathrm{III}}$, the segregation gets stronger in the sense that X and Y points tend to form one-class clumps or clusters.

The power estimates for the right-sided versions under segregation alternatives are presented in Figure 1. We only present the power estimates for $H_{\mathrm{S}}^{\mathrm{I}}$ and $H_{\mathrm{S}}^{\mathrm{II}}$, since under $H_{\mathrm{S}}^{\mathrm{III}}$ the power estimates for all tests are very close to 1 (Ceyhan 2008c). We defer the power estimates for the two-sided alternative to the technical report (Ceyhan 2008c); and omit the power estimates of the tests for the left-sided alternative under the segregation alternatives as they are virtually zero. Observe that, for all directional tests, as $n = (n_1 + n_2)$ gets larger, the power estimates get larger under each segregation alternative; for the same $n = (n_1 + n_2)$ values, the power estimate is larger for classes with similar sample sizes; and as the segregation gets stronger, the power estimates get larger at each (n_1, n_2) combination. The power estimates for Ceyhan's cell-specific tests and the new versions of the directional tests are similar and higher than those for Dixon's cell-specific tests. Under the segregation alternative, the new directional tests and Ceyhan's cell-specific tests have

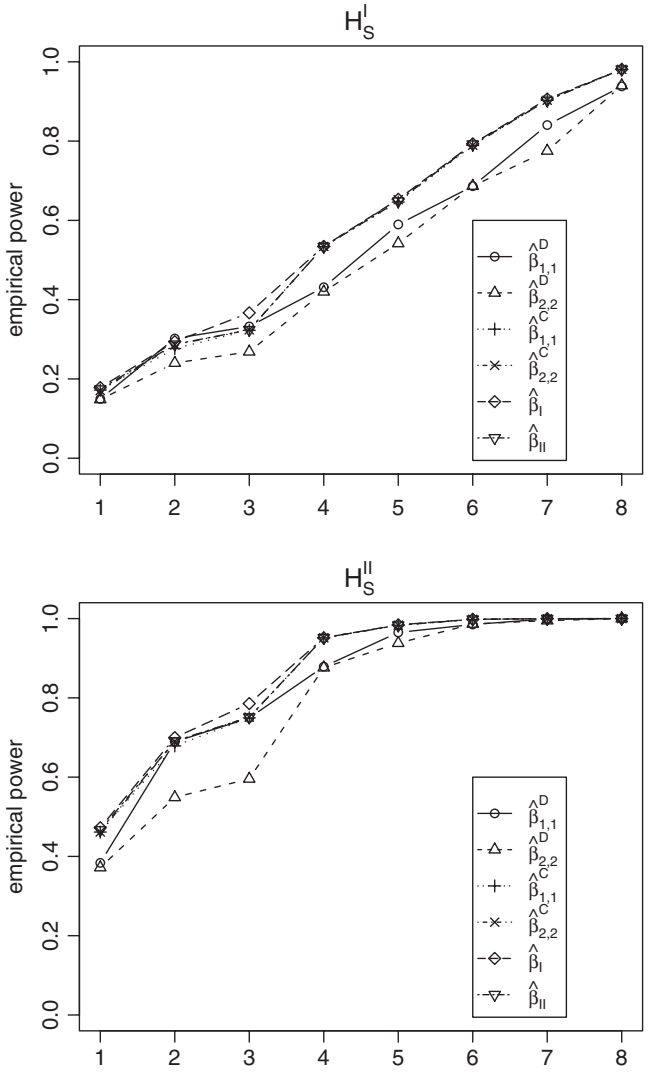

Figure 1. The empirical power estimates for Dixon's and Ceyhan's cell-specific tests for cell (i, i), $i = 1, 2$ and the new directional tests under the segregation alternatives H_S^I and H_S^{II} in the two-class case for the right-sided alternatives (which is sensitive for the segregation pattern). $\hat{\beta}^D_{i,i}$ and $\hat{\beta}^C_{i,i}$ stand for Dixon's and Ceyhan's cell-specific tests for cell (i, i) $i = 1, 2$, respectively, and $\hat{\beta}_I$ and $\hat{\beta}_{II}$ stand for the versions I and II of the new directional tests of segregation. The horizontal axis labels are $1 = (10,10)$, $2 = (10,30)$, $3 = (10,50)$, $4 = (30,30)$, $5 = (30,50)$, $6 = (50,50)$, $7 = (50,100)$, $8 = (100,100)$.

similar performance, because they are constructed similarly. That is, each of these tests involve cell and column information, whereas Dixon's cell-specific test only involves the information from the associated cell. Among the tests considered, version I of the new directional tests seems to have the highest power estimates.

5.2. *Empirical power analysis under association alternatives*

For the association alternatives, we consider three cases. First, we generate $X_i \overset{\text{iid}}{\sim} \mathcal{U}((0, 1) \times (0, 1))$ for $i = 1, 2, \ldots, n_1$. Then we generate Y_j for $j = 1, 2, \ldots, n_2$ as follows. For each j, we

pick an i randomly, then generate Y_j as $X_i + R_j\,(\cos T_j, \sin T_j)'$ where $R_j \overset{\text{iid}}{\sim} \mathcal{U}(0, r)$ with $r \in (0, 1)$ and $T_j \overset{\text{iid}}{\sim} \mathcal{U}(0, 2\pi)$. In the pattern generated, appropriate choices of r will imply association between classes X and Y. That is, it will be more likely to have (X, Y) or (Y, X) NN pairs than same-class NN pairs (i.e. (X, X) or (Y, Y)). The three values of r we consider constitute the following three association alternatives:

$$H_{\mathrm{A}}^{\mathrm{I}} : r = \frac{1}{4}, \quad H_{\mathrm{A}}^{\mathrm{II}} : r = \frac{1}{7}, \quad \text{and} \quad H_{\mathrm{A}}^{\mathrm{III}} : r = \frac{1}{10}.$$

Observe that, from $H_{\mathrm{A}}^{\mathrm{I}}$ to $H_{\mathrm{A}}^{\mathrm{III}}$, the association gets stronger in the sense that X and Y points tend to occur together more and more frequently. By construction, for similar sample sizes, the association between X and Y are at about the same degree as association between Y and X. For

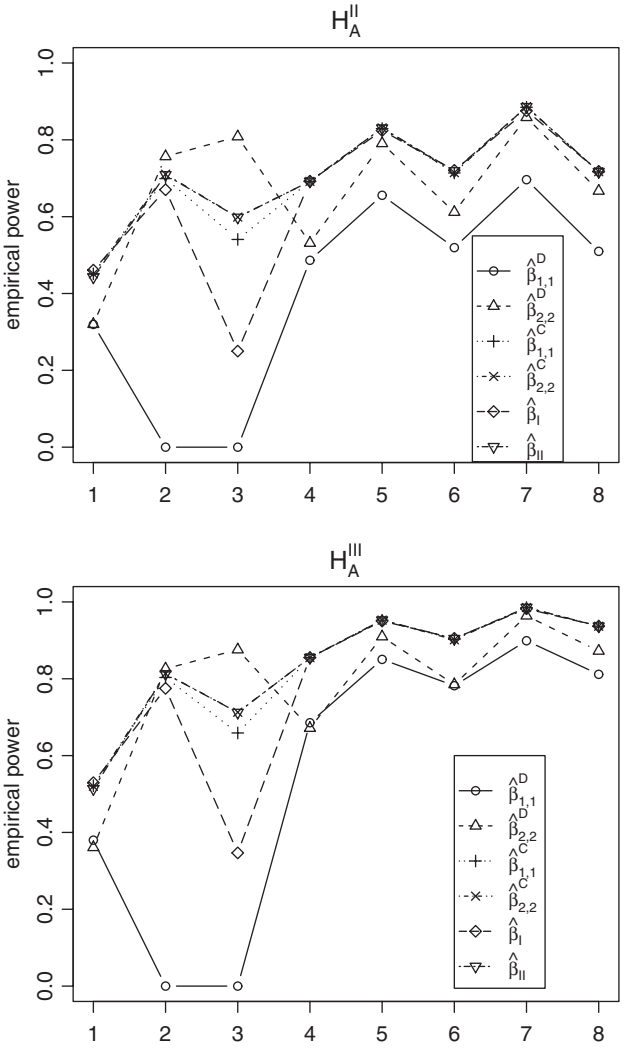

Figure 2. The empirical power estimates for Dixon's and Ceyhan's cell-specific tests for cell (i, i), $i = 1, 2$ and the new directional tests under the association alternatives $H_{\mathrm{A}}^{\mathrm{II}}$ and $H_{\mathrm{A}}^{\mathrm{III}}$ in the two-class case for the left-sided alternatives. (which is sensitive for the association pattern). The power and horizontal axis labelling is as in Figure 1.

very different sample sizes, smaller sample is associated with the larger but the abundance of the larger sample confounds its association with the smaller.

The power estimates for the left-sided versions under association alternatives are presented in Figure 2. We only present the power estimates for H_A^{II} and H_A^{III}, since under H_A^{I}, the power estimates are not so high for all tests (Ceyhan 2008c). We defer the the power estimates for the two-sided alternative to the technical report (Ceyhan 2008c); and omit the power estimates of the tests for the right-sided alternative under the association alternatives as they are virtually zero. Observe that when sample sizes are similar (see $n_1 = n_2$ cases), for all tests, as $n = (n_1 + n_2)$ gets larger, the power estimates get larger; and as the association gets stronger, the power estimates get larger.

For smaller samples, Dixon's test has the highest power. For larger samples the new versions and Ceyhan's test have similar (although less similar compared with the segregation case) power performance, but they both have higher power compared to Dixon's tests. The power performance is highly dependent on the level of relative abundances of the classes. This might be due to the fact that by construction, when class X is much larger than class Y, the two classes are not strongly associated, since NN of X points could also be from the same class with a high probability. The lack of association when class Y is larger occurs for the same reason.

6. Example data

We illustrate the tests on an ecological data set. Dixon (1994, 2002a,b) illustrates NN methods on a 50×200 m rectangular plot of hardwood swamp in South Carolina, USA. The plot contains 13 different tree species, of which we only consider two, namely: bald cypress (*Taxodium distichum*) and black gum trees (*Nyssa sylvatica*). The question of interest is whether these tree species are segregated, associated, or satisfy CSR independence. For more detail on the data, see Dixon (2002b). The locations of these trees in the study region are plotted in Figure 3 and the corresponding NNCT, together with percentages, is provided in Table 4. Observe that the percentages are suggestive of segregation for both species.

The locations of the tree species can be viewed *a priori*, resulting from different processes, so the more appropriate null hypothesis is the CSR independence pattern. Hence, our inference will be conditional (see Remark 2). We calculate $Q = 178$ and $R = 156$ for this data set. We present Dixon's and Ceyhan's cell-specific and new directional test statistics, and the associated p-values of the two-, right-, and left-sided alternatives in Table 5, where p_{asy} stands for the p-value based on the asymptotic approximation. p_{mc} is the p-value based on a 9999 Monte Carlo replication

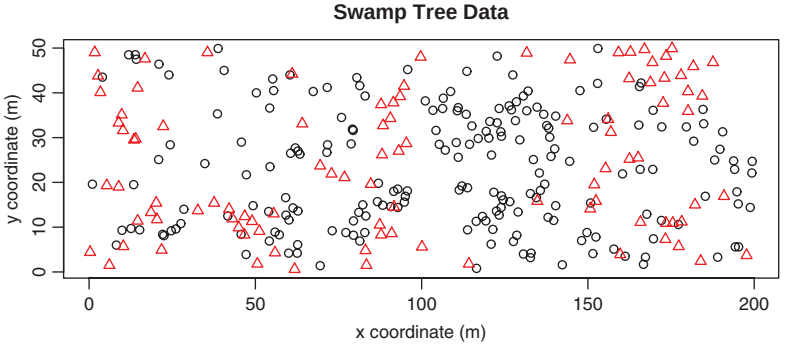

Figure 3. The scatter plots of the locations of black gum trees (circles ○) and bald cypress trees (triangles △).

Table 4. The NNCT for swamp tree data and the corresponding percentages (in parenthesis).

		NN		
		BG	BC	Sum
Base	BG	149(82%)	33(18%)	182(67%)
	BC	43(47%)	48(53%)	91(23%)
	Sum	192(34%)	81(66%)	273(100%)

Note: BG, black gum trees, BC, bald cypress trees.

Table 5. Test statistics and the associated p-values for the two-sided and directional alternatives for the swamp tree data.

Test statistics and the associated p-values for the swamp tree data						
	Z_{11}^{D}	Z_{22}^{D}	Z_{11}^{C}	Z_{22}^{C}	Z_{I}	Z_{II}
Test statistics	4.47	3.54	4.62	4.61	4.76	4.61
	Against the two-sided alternative					
p_{asy}	<0.0001	0.0004	<0.0001	<0.0001	<0.0001	<0.0001
p_{mc}	<0.0001	0.0003	<0.0001	<0.0001	<0.0001	<0.0001
p_{rand}	<0.0001	0.0004	<0.0001	<0.0001	<0.0001	<0.0001
	Against the right-sided (i.e. segregation) alternative					
p_{asy}	<0.0001	0.0002	<0.0001	<0.0001	<0.0001	<0.0001
p_{mc}	<0.0001	0.0002	<0.0001	<0.0001	<0.0001	<0.0001
p_{rand}	<0.0001	0.0003	<0.0001	<0.0001	<0.0001	<0.0001
	Against the left-sided (i.e. association) alternative					
p_{asy}	≈1.0	0.9998	≈1.0	≈1.0	≈1.0	≈1.0
p_{mc}	≈1.0	0.9998	≈1.0	≈1.0	≈1.0	≈1.0
p_{rand}	≈1.0	0.9998	≈1.0	≈1.0	≈1.0	≈1.0

Note: p_{asy}, p_{mc}, and p_{rand} stand for the p-values based on the asymptotic approximation, Monte Carlo simulation, and randomisation of the tests, respectively.

of CSR independence in the same plot. First, we calculate the test statistics for the current data. Then we generate 182 and 91 points uniformly from two classes in the same region 9999 times. At each Monte Carlo replication, we calculate the corresponding test statistics. Then we combine the test statistics from the data and the Monte Carlo simulations and determine the rank of the tests statistics, say q. Then for the segregation alternative, p_{mc} is $(10000 - q)/10000$ and for the association alternative p_{mc} is $q/10000$. p_{rand} is based on Monte Carlo randomisation of the labels on the given locations of the trees 9999 times. That is, we perform RL of two labels to the 273 tree locations in the study area so that 182 of the locations are labelled with class 1 and 91 are labelled with class 2. At each step, we also calculate the test statistics, and then combine these test statistics with those from the original data. Determining the rank of the test statistics from data yields p_{rand} as in the p_{mc} case. Notice that p_{asy}, p_{mc}, and p_{rand} are very similar for each test, so the asymptotic approximation seems to be a reliable shortcut compared with the computationally expensive Monte Carlo alternatives. All tests are significant for the two-sided alternative, implying deviation from CSR independence, and among the directional tests, right-sided tests are significant, indicating a significant segregation between black gums and bald cypresses.

The results based on NNCT tests pertain to small-scale interaction at about the average NN distances. To find out the causes of the segregation and the type and level of interaction between the tree species at different scales (i.e. distances between the trees), we also present the second-order analysis of the swamp tree data (Diggle 2003) using the functions (or some modified version of them) provided in the spatstat package in R (Baddeley and Turner 2005). We use

Ripley's univariate and bivariate L-functions, which are modified versions of his K-functions. For a rectangular region, to remove the bias in estimating $K(t)$, using t values up to 1/4 of the smaller side length of the rectangle is recommended. So we take the values $t \in [0, 12.5]$ in our analysis, since the rectangular region is 50×200 m. But Ripley's K-function is cumulative, so interpreting the spatial interaction at larger distances is problematic (Wiegand, Gunatilleke and Gunatilleke 2007). The pair correlation function $g(t)$ is better for this purpose (Stoyan and Stoyan 1994). The pair correlation function of a (univariate) stationary point process is defined as $g(t) = K'(t)/2\pi t$ where $K'(t)$ is the derivative of $K(t)$. However, if $g(t) > 0$, the pair correlation function estimates might have critical behaviour for small t since the estimator of variance and hence the bias are considerably large (Stoyan and Stoyan 1996). So pair correlation function analysis is more reliable for larger distances. We can use Ripley's L-function for distances up to the average NN distance, or use NNCT tests for about the average NN distance.

Ripley's univariate L-functions and the pair correlation functions for both species combined and each species for the swamp tree data are presented in Figure 4. The average NN distance in the swamp tree data is 3.08 ± 1.70 m (mean $\pm$ standard deviation). So for up to about 3 m, Ripley's L-function suggests that all trees combined do not significantly deviate from CSR, however, black gums seem to be significantly aggregated for distances of about [2, 3] meters and bald cypresses are significantly aggregated for distances of about 3 m. For other distances in [0, 3] meters, the pattern is not significantly different from CSR. Hence, segregation of the species detected by the NNCT tests might be due to different levels and types of aggregation of the species in the study region. For distances larger than 3 m, the pair correlation function suggests that all trees are significantly aggregated for distances about [3, 5] meters; black gums are significantly aggregated for about [3, 7] and [9, 11] meters; and bald cypresses are significantly aggregated for about [3, 9] and [10, 12.5] meters. For other distances in [3, 12.5] meters, the pattern is not significantly different from CSR.

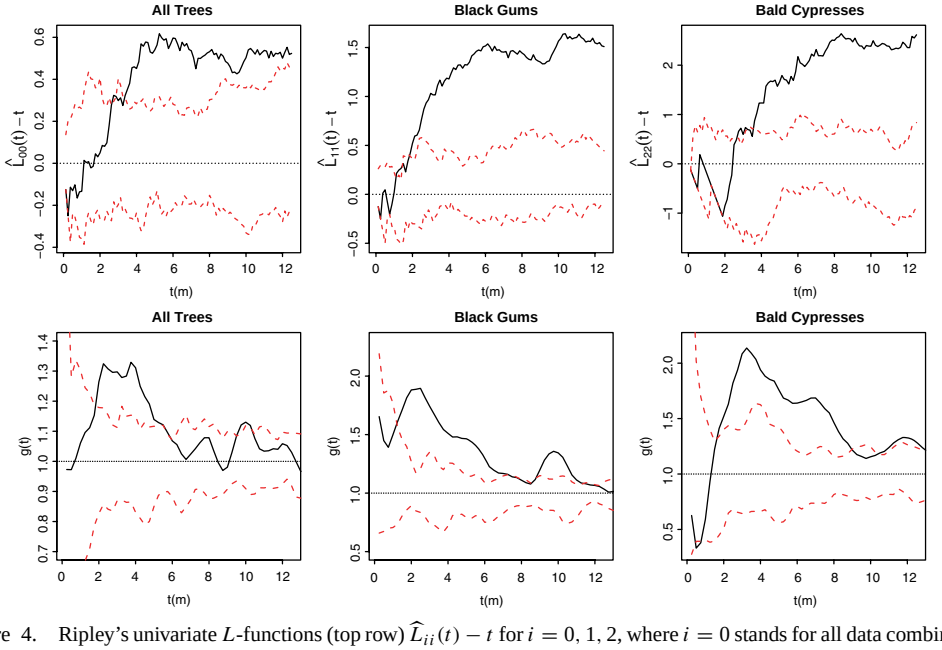

Figure 4. Ripley's univariate L-functions (top row) $\widehat{L}_{ii}(t) - t$ for $i = 0, 1, 2$, where $i = 0$ stands for all data combined, $i = 1$ for black gums, and $i = 2$ for bald cypresses; and pair correlation functions $g(t)$ for all trees combined and for each species (bottom row). Wide dashed lines are the upper and lower (pointwise) 95% confidence bounds for the functions based on Monte Carlo simulation under the CSR independence pattern.

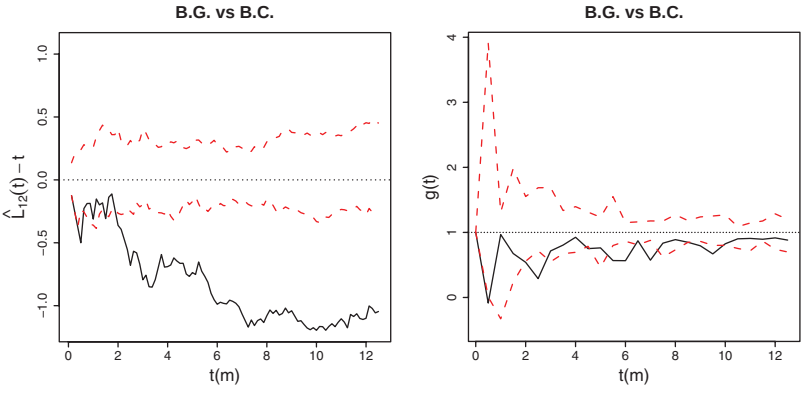

Figure 5. Ripley's bivariate L-function $\widehat{L}_{12}(t) - t$ (left) and pair correlation function $g(t)$ (right) for the swamp tree data. Wide dashed lines are the upper and lower (pointwise) 95% confidence bounds for the functions based on Monte Carlo simulations under the CSR independence pattern. BG = black gums and BC = bald cypresses.

Ripley's bivariate L-function $L_{ij}(t)$ is symmetric in i and j in theory, that is, $L_{ij}(t) = L_{ji}(t)$ for all i, j. In practice, although edge corrections will render it slightly asymmetric, i.e., $\widehat{L}_{ij}(t) \neq \widehat{L}_{ji}(t)$ for $i \neq j$. The corresponding estimates are pretty close in our example, so we only present one of them. The same definition of the pair correlation function can be applied to Ripley's bivariate K or L-functions as well.

Ripley's bivariate L-function and the bivariate pair correlation function for the species in swamp tree data are plotted in Figure 5. For 0–3 m, Ripley's bivariate L-function suggests that the tree species are significantly segregated for distances of about 0.5 and [1.8, 3] meters, and do not significantly deviate from CSR independence for other distances. For 3–12.5 m, the pair correlation function suggests that the tree species are significantly segregated for distances about [5, 7] and 9 meters, and do not significantly deviate from CSR independence for other distances.

The NNCT methods and the second-order methods can also be applied to biomedical data such as the spatial interaction of defected neurons and healthy neurons. See Ceyhan (2008c) for an example.

7. Discussion and conclusions

Pielou's and Dixon's segregation tests based on NNCTs are χ^2 tests, and are hence used for two-sided alternatives. That is, when the null patterns of CSR independence or RL are rejected, these tests do not indicate the direction of the alternative pattern, which consist of segregation or association patterns. In this article, we discuss directional (i.e. one-sided) tests of segregation based on NNCTs. We consider a directional version of Pielou's test by partitioning the χ^2 test statistic in the usual fashion (Bickel and Doksum 1977). However, the directional version of Pielou's test, just like the usual version of Pielou's test, is liberal in rejecting the null case of CSR independence or RL. We also consider the directional versions of the cell-specific tests due to Dixon (1994) and Ceyhan (2008a). Furthermore, we introduce two new directional tests of segregation.

Based on our Monte Carlo simulations, we conclude that the asymptotic approximation for the cell-specific and the directional tests is appropriate only when the corresponding cell count in the NNCT is larger than 10. When a cell count is less than 10, we recommend the Monte Carlo randomisation of these tests. Type I error rates (empirical significance levels) of Ceyhan's cell-specific and of the new directional tests are more robust to the differences in sample sizes (i.e. differences in relative abundances). The new directional tests and Ceyhan's cell-specific tests

have very similar size and power performance, due to the similarity in their construction. But both tests perform better compared with Dixon's cell-specific tests. Considering the empirical significance levels and power estimates of the tests, we recommend version II of the new tests (defined in Equation (5)).

NNCT tests summarise the pattern in the data set for small scales at about the average NN distance. On the other hand, pair correlation function $g(t)$ and Ripley's classical K or L-functions and other variants provide information on the pattern at various scales. Our example illustrates that for distances around the average NN distance, NNCT tests and Ripley's bivariate L-function yield similar results. If significant, the cell-specific test and the new tests (for the two-sided alternative) imply significant deviation from the null pattern. Furthermore, the sign of the test statistic will be suggestive of segregation (if positive) and association (if negative). But these tests are more powerful against the one-sided alternatives. For a data set for which CSR independence is the reasonable null pattern, we recommend the NNCT tests if the question of interest is the spatial interaction at small scales (i.e. about the mean NN distance). One can also perform Ripley's K or L-function and only consider distances up to around the average NN distance and compare the results with those of the NNCT analysis. If the spatial interaction at higher scales is of interest, pair correlation function is recommended (Loosmore and Ford 2006). On the other hand, if the RL pattern is the reasonable null pattern for the data, we recommend the NNCT tests if the small-scale interaction is of interest and Diggle's D-function (Diggle 2003, p. 131) if the spatial interaction at higher scales is also of interest.

In this article, NNCTs are based on NN relations using the usual Euclidean distance. But, the NN relation could also be based on dissimilarity measures between finite or infinite dimensional data points, which can be spatial or nonspatial. Such a generalisation of the NNCT tests make this methodology potentially useful in many fields. For example, in functional data analysis (Ferraty and Vieu 2006; Cuevas, Febrero and Fraiman 2006), as long as a distance or a dissimilarity measure is available, one can construct NNCTs and analyse various patterns of clustering. In this general context the NN of object x refers to the object with the minimum dissimilarity to x. The extension of the RL pattern to this general context is straightforward, but extra care should be taken for such an extension of CSR independence. In either case, the term Q, which is the number of points with shared NNs, needs to be revised as $\widetilde{Q} := 2\sum_{k=1}^{N}\binom{k}{2}Q_k$. Alternatively, rather than generalising the tests to high dimensions or non-Euclidean dissimilarity measures, one can project the points (using the original data or the inter-point distance structure) to a lower dimension (preferably to $\mathbb{R}^2$) by multidimensional scaling (Cox and Cox 2001), and apply the spatial pattern tests discussed in this article.

Acknowledgements

I would like to thank an anonymous associate editor and three referees, whose constructive comments and suggestions greatly improved the presentation and flow of the paper. Some of the Monte Carlo simulations presented in this article were executed on the Hattusas cluster of Koç University High Performance Computing Laboratory.

References

Anderson, T.W. (1984), *An Introduction to Multivariate Statistical Analysis*, New York: John Wiley & Sons Inc.

Baddeley, A.J., and Turner, R. (2005), 'spatstat: An R Package for Analyzing Spatial Point Patterns', *Journal of Statistical Software*, 12(6), 1–42.

Bickel, P.J. (1974), 'Edgeworth Expansions in Non-parametric Statistics', *Annals of Statistics*, 2, 1–20.

Bickel, P.J., and Doksum, A.K. (1977), *Mathematical Statistics, Basic Ideas and Selected Topics*, Englewood Cliffs, NJ: Prentice Hall.

Ceyhan, E. (2007), 'Edge Correction for Cell- and Class-Specific Tests of Segregation Based on Nearest Neighbor Contingency Tables', in *Proceedings of the International Conference on Environment: Survival and Sustainability*, Near East University.

Ceyhan, E. (2008a), 'Overall and Pairwise Segregation Tests Based on Nearest Neighbor Contingency Tables', *Computational Statistics & Data Analysis*, 53(8), 2786–2808.
Ceyhan, E. (2008b), 'Overall and Pairwise Segregation Tests Based on Nearest Neighbor Contingency Tables', arXiv:0805.1629v2 [stat.ME], Tech. Rep. # KU-EC-08-1.
Ceyhan, E. (2008c), 'Directional Clustering Tests Based on Nearest Neighbor Contingency Tables', arXiv:0902.0990v1 [stat.ME], Tech. Rep. # KU-EC-09-1.
Ceyhan, E. (2008d), 'QR-adjustment for Clustering Tests Based on Nearest Neighbor Contingency Tables', arXiv:0807.4231v1 [stat.ME], Tech. Rep. # KU-EC-08-5.
Ceyhan, E. (2008e), 'New Tests for Spatial Segregation Based on Nearest Neighbor Contingency Tables', arXiv:0808.1409v1 [stat.ME], Tech. Rep. # KU-EC-08-6.
Ceyhan, E. (2009), 'Class-specific Tests of Segregation Based on Nearest Neighbor Contingency Tables', *Statistica Neerlandica*, 63(2), 149–182.
Ceyhan, E. (n.d.), 'On the use of Nearest Neighbor Contingency Tables for Testing Spatial Segregation', *Environmental and Ecological Statistics*, doi:10.1007/s10651-008-0104-x.
Coomes, D.A., Rees, M., and Turnbull, L. (1999), 'Identifying Aggregation and Association in Fully Mapped Spatial Data', *Ecology*, 80(2), 554–565.
Cox, T.F. (1981), 'Reflexive Nearest Neighbours', *Biometrics*, 37(2), 367–369.
Cox, T., and Cox, M. (2001), *Multidimensional Scaling*, Boca Raton, FL: Chapman and Hall.
Cressie, N.A.C. (1993), *Statistics for Spatial Data*, New York: Wiley.
Cuevas, A., Febrero, M., and Fraiman, R. (2006), 'On the use of the Bootstrap for Estimating Functions with Functional Data', *Computational Statistics & Data Analysis*, 51(2), 1063–1074.
Diggle, P.J. (2003), *Statistical Analysis of Spatial Point Patterns*, London: Hodder Arnold Publishers.
Diggle, P.J., and Chetwynd, A.G. (1991), Second-order Analysis of Spatial Clustering for Inhomogeneous Populations', *Biometrics*, 47, 1155–1163.
Dixon, P.M. (1994), 'Testing Spatial Segregation Using a Nearest-neighbor Contingency Table', *Ecology*, 75(7), 1940–1948.
Dixon, P.M. (2002a), 'Nearest-neighbor Contingency Table Analysis of Spatial Segregation for Several Species, *Ecoscience*, 9(2), 142–151.
Dixon, P.M. (2002b), 'Nearest Neighbor Methods', in *Encyclopedia of Environmetrics* (Vol. 3), eds. Abdel H. El-Shaarawi and Walter W. Piegorsch, New York: John Wiley & Sons Ltd., pp. 1370–1383.
Ferraty, F., and Vieu, P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, Springer Series in Statistics, Berlin: Springer.
Goreaud, F., and Pélissier, R. (2003), 'Avoiding Misinterpretation of Biotic Interactions with the Intertype K12-function: Population Independence vs. Random Labelling Hypotheses', *Journal of Vegetable Science*, 14(5), 681–692.
Herler, J., and Patzner, R.A. (2005), 'Spatial Segregation of Two Common Gobius Species (Teleostei: Gobiidae) in the Northern Adriatic Sea', *Marine Ecology*, 26(2), 121–129.
Kulldorff, M. (2006), Tests for Spatial Randomness Adjusted for an Inhomogeneity: A General Framework', *Journal of the American Statistical Association*, 101(475), 1289–1305.
Lahiri, S.N. (1996), 'On Consistency of Estimators Based on Spatial Data Under Infill Asymptotics', *Sankhya: Indian Journal of Statistics*, Series A, 58(3), 403–417.
Loosmore, N., and Ford, E. (2006), 'Statistical Inference Using the G or K Point Pattern Spatial Statistics', *Ecology*, 87, 1925–1931.
Meagher, T.R., and Burdick, D.S. (1980), The use of Nearest Neighbor Frequency Analysis in Studies of Association', *Ecology*, 61(5), 1253–1255.
Moran, P.A.P. (1948), The Interpretation of Statistical Maps', *Journal of the Royal Statistical Society*, Series B, 10, 243–251.
Nanami, S.H., Kawaguchi, H., and Yamakura, T. (1999), 'Dioecy-induced Spatial Patterns of Two Codominant Tree Species, Podocarpus nagi and Neolitsea aciculata', *Journal of Ecology*, 87(4), 678–687.
Pacala, S.W. (1986), 'Neighborhood Models of Plant Population Dynamics. II. Multispecies Models of Annuals', *Theoretical Population Biology*, 29, 262–292.
Pielou, E.C. (1961), 'Segregation and Symmetry in Two-species Populations as Studied by Nearest-neighbor Relationships', *Journal of Ecology*, 49(2), 255–269.
Ripley, B.D. (2004), *Spatial Statistics*, New York: Wiley-Interscience.
Stoyan, D., and Stoyan, H. (1994), *Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics*, New York: John Wiley and Sons.
Stoyan, D., and Stoyan, H. (1996), 'Estimating Pair Correlation Functions of Planar Cluster Processes', *Biometrical Journal*, 38(3), 259–271.
Waller, L.A., and Gotway, C.A. (2004), *Applied Spatial Statistics for Public Health Data*, NJ: Wiley-Interscience.
Whipple, S.A. (1980), 'Population Dispersion Patterns of Trees in a Southern Louisiana Hardwood Forest', *Bulletin of the Torrey Botanical Club*, 107, 71–76.
Wiegand, T., Gunatilleke, S., and Gunatilleke, N. (2007), 'Species Associations in a Heterogeneous Sri Lankan Dipterocarp Forest', *The American Naturalist*, 170(4), 77–95.
Yamada, I., and Rogersen, P.A. (2003), 'An Empirical Comparison of Edge Effect Correction Methods Applied to K-function Analysis', *Geographical Analysis*, 35(2), 97–109.