Journal of Probability and Statistical Science
8(1), 45-68, Feb. 2010

Exact Inference for Testing Spatial Segregation by Nearest Neighbor Contingency Tables

Elvan Ceyhan
Koc University

ABSTRACT Nearest neighbor methods are widely used in the analysis of spatial point patterns in ecology and environmental sciences. We present exact inference on tests based on nearest neighbor contingency tables (i.e., NNCT-tests) for testing segregation and association patterns. The spatial pattern of *segregation* occurs when members of a class tend to be found near members of the same class (i.e., conspecifics), while *association* occurs when members of a class tend to be found near members of the other class or classes. The null hypothsis is randomness in the nearest neighbor structure, which may result from — among other patterns — *random labeling* (RL) or *complete spatial randomness* (CSR) of points from two or more classes (which is called *CSR independence*, henceforth). Pielou's, Dixon's, and various other NNCT-tests rely on asymptotic approximations. Exact inference has been extensively used on contingency tables in general, but not for NNCT- tests. We propose several variants of Fisher's exact test on NNCTs for testing CSR independence or RL with one- or two-sided alternatives, as well as variants of the exact version of Pearson's test on NNCTs for the two-sided alternative. We also perform a correction on odds ratio, the parameter used in exact inference for contingency tables. An extensive Monte Carlo study is provided for empirical significance level (i.e., Type I error rate) and power comparisons We demonstrate that the most conservative versions of the exact tests have the appropriate level and higher power compared to other exact and asymptotic NNCT-tests.

Keywords Association; Clustering; Complete spatial randomness; Fisher's exact test; Independence; Random labeling; Spatial point pattern.

1. Introduction

Spatial point patterns have important implications in various fields such as ecology, population biology, and epidemiology. Most of the related research has been on patterns of one

Received March 2007, revised November 2008, in final form September 2009.
Elvan Ceyhan is affiliated with the Department of Mathematics at Koc University, Istanbul, Turkey; email: elceyhan@ku.edu.tr.

 ISSN 1726-3328

type of points, which usually fall under the pattern category called *spatial aggregation* (or *clustering*) (Coomes *et al*. [12]) or *regularity*. However, it is also of practical interest to investigate the patterns of and interaction between two or more types of points (Pielou [27], Haase [27], Dixon [15], and Upton and Fingleton [30]). For convenience and generality, we call the types of points as "classes", but a *class* can be replaced by any characteristic of an individual at a particular location. For example, the pattern of spatial segregation has been investigated for *plant species* (Diggle [13]), *age classes* of plants (Hamill and Wright [20]), *fish species* (Herler and Patzner [21]), and *sexes* of dioecious plants (Nanami *et al*. [26]) and animals (Conradt [11]). Many of the epidemiological applications are for a two-class system of case and control labels (Waller and Gotway [31]).

Nearest neighbor (NN) methods are widely used in the analysis of spatial patterns and are based on some measure of dissimilarity between a (base) point and its NN (if Y is a NN of point X, then X is the base point and Y is the NN point); such as the distance between the points or the class types of the pair of NN points (see, e.g., Dixon [16]). In NNCTs, the second type of similarity is used, i.e., NNCTs are constructed using the NN frequencies of classes. Pielou [27] and Dixon [14] proposed different tests of random labeling in a spatial pattern. Pielou [27] proposed various tests based on NNCTs which are independent of quadrat size (Krebs [22]) and have been used for the two-class case only; while Dixon [14] devised overall and class-specific tests of segregation based on NNCTs for the two-class case and extended his methodology to the multi-class case (Dixon [15]). Various authors have demonstrated problems with Pielou's test (Meagher and Burdick [24] and Dixon [14]), for a general discussion and literature on the use and appropriateness of Pielou's and Dixon's tests see (Ceyhan [8]) where it is shown that Pielou's test is liberal in rejecting the randomness in the NN structure, but is appropriate for NNCTs based on a random sample of (base,NN) pairs.

Exact inference has been implemented on contingency tables (Agresti [1]), but not on NNCTs. Note that the everyday meaning and the technical meaning of the word "exact" do not coincide. In statistical methodology, "exact inference" means either finite (i.e., small as opposed to large) sample inference, or conditional (on nuisance parameters) inference (Epstein and Fienberg [17]). For contingency tables, both meanings are implied, in the sense that, finite sample distribution of the cell counts conditional on marginal totals is required.

In this article, we investigate the use of exact tests on NNCTs for spatial segregation and spatial association. We discuss the use and appropriateness of many variants of Fisher's exact test and the exact version of Pearson's test on NNCTs for testing CSR independence or RL. We also implement a correction on odds ratio for exact tests. The evaluation of these tests in terms of empirical size as well as power are performed with an extensive Monte Carlo simulation study. We demonstrate that the most conservative versions of the exact tests provide reasonable small sample alternatives to the asymptotic tests of Dixon and others (Dixon [15] and Ceyhan [7]). We describe the null and alternative patterns in Section 2, exact inference on contingency

tables in general in Section 3, NNCTs and exact inference on NNCTs in Section 4, Monte Carlo simulation analysis in Sections 5 and 6. We apply the tests to example data sets in Section 7 and provide guidelines and recommendations in Section 8.

2. Null and Alternative Patterns

For simplicity, we describe the spatial patterns between two classes; the extension to multi-class case is straightforward. Informally, we aim to test the randomness in the NN structure. Such randomness might occur if the two classes of points exhibit either of the two (random) pattern types: *complete spatial randomness independence* (CSR independence) or *random labeling* (RL).

Under CSR independence, points from each of the two classes satisfy the CSR pattern in the region of interest. On the other hand, RL is the pattern in which, given a fixed set of points in a region, class labels are assigned to these fixed points randomly so that the labels are independent of the locations. That is, CSR independence is a process defining the spatial distribution of event locations, while RL is a process, conditioned on locations, defining the distribution of labels on these locations.

Our null hypothesis of randomness in NN structure might result from either of the patterns. That is, when the points from both classes are assumed to be uniformly distributed over the region of interest, then the null hypothesis we consider is H_o : CSR *independence*. Note that this is equivalent to the case that RL is applied to a given set of points from a CSR pattern. When only the labeling of a set of fixed points (the allocation of the points could be regular, aggregated, or clustered, or of lattice type) is random, our null hypothesis is H_o : RL. Although CSR independence and RL are not same, they lead to the same null model (i.e., randomness in the NN structure) in tests using NNCT, which does not require spatially-explicit information. The distinction between CSR independence and RL is very important when defining the appropriate null model in practice. Goreaud and Pélissier [18] 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 choose the CSR independence pattern for the comparison of the empirical sizes in Section 5, because it provides a more general framework for the locations of the points, in the sense that, RL can be applied on any given set of points, but the locations of the points should be fixed for all steps of the RL process. But when we generate both classes of points under CSR independence pattern, we attain randomness in the NN structure as well as the locations. Hence, a wide range of spatial allocation of points in a study area can be analyzed by this choice.

The alternative patterns fall under two major categories called *association and segregation*. *Association* occurs if the NN of a base point is more likely to be from a class different from the

class of the base point. 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. *Segregation* occurs if the NN of a base point is more likely to be of the same class as the class of the base point; i.e., the members of the same class tend to be clumped or clustered (see, e.g., Pielou [27]). Many different forms of segregation are possible. Although it is not possible to list all segregation types, existence of it can be tested by an analysis of the NN relationships between the classes (Pielou [27]). For instance, one type of plant might not grow well (or at all) around another type of plant, and vice versa.

The term *association* is also used for the categorical interaction in contingency tables, but the association we consider is the "spatial association pattern". Segregation does not cause such an ambiguity.

3. Exact Inference on Contingency Tables

In general, contingency tables are assumed to result from one of three sampling frameworks for the cell counts: Poisson, row-wise multinomial, or overall multinomial sampling frameworks. In Poisson and row-wise multinomial frameworks, cell counts are independent, while in overall multinomial framework the cell counts have negative correlation (which vanishes in the asymptotics). Nevertheless, all frameworks yield tests that are approximately distributed as chi-square for large n (Conover [10]).

3.1 Fisher's Exact Test for Contingency Tables

For a 2×2 contingency table, let N_{ij} be the cell frequency for cell (i, j), N_i be the sum of row i, and C_j be the sum for column j. Given row and column sums (marginals), N_{11} determines the other three cell counts in the contingency table and has the *hypergeometric distribution* with non-centrality parameter θ. In general, the null case of independence in the contingency tables is equivalent to $H_o : \theta = 1$, where θ is the odds ratio (or non-centrality parameter) in contingency tables (Agresti [1]). Fisher's exact test can be two-sided or one-sided for 2×2 contingency tables, while for contingency tables of dimension other than 2×2 only the two-sided alternative is available (Agresti [2]).

There are various ways to calculate p-values for different alternatives in exact inference on contingency table (Agresti [1]). We call these as *variants of Fisher's exact test* in this article. Exact p-values tend to be more conservative than most approximate (asymptotic) ones (Agresti [1]).

3.1.1 Variants of Fisher's Exact Test for One-Sided Alternatives

For the one-sided alternatives, the probabilities of more extreme tables are summed up, including or excluding the probability of the table itself (or some middle way). Let the probability of the contingency table itself be $p_t = f(n_{11}|n_1, n_2, c_1; \theta = 1)$, where $f(\cdot)$ is the

probability density function of hypergeometric distribution. For testing the one-sided alternative $H_o : \theta = 1$ versus $H_a : \theta > 1$, we consider the following four methods in calculating the p-value: Let $p = \sum_S f(t|n_1, n_2, c_1; \theta = 1)$, then

(i) with $S = \{t : t \geq n_{11}\}$, we get the *table-inclusive version* which is denoted as $p^{>}_{\text{inc}}$,

(ii) with $S = \{t : t > n_{11}\}$, we get the *table-exclusive version*, denoted as $p^{>}_{\text{exc}}$.

(iii) Using $p = p^{>}_{\text{exc}} + p_t/2$, we get the *mid-p version*, denoted as $p^{>}_{\text{mid}}$.

(iv) We can also use *Tocher corrected version* which is denoted as $p^{>}_{\text{Toc}}$.

Tocher's modification makes Fisher's exact test less conservative, by including the probability for the current table based on a randomized test (Tocher [29]). When table-inclusive version of the p-value, $p^{>}_{\text{inc}}$, is larger, but table-exclusive version, $p^{>}_{\text{exc}}$, is less than the level of the test α, a random number, U, is generated from uniform distribution in $(0, 1)$, and if $U \leq (\alpha - p^{>}_{\text{exc}})/p_t$, $p^{>}_{\text{exc}}$ is used, otherwise $p^{>}_{\text{inc}}$ is used as the p-value. That is,

$$p^{>}_{\text{Toc}} = \begin{cases} p^{>}_{\text{exc}} & \text{if } U \leq (\alpha - p^{>}_{\text{exc}})/p_t, \\ p^{>}_{\text{inc}} & \text{otherwise.} \end{cases} \tag{1}$$

Note also that $p^{>}_{\text{exc}} = p^{>}_{\text{inc}} - p_t$ and $p^{>}_{\text{mid}} = p^{>}_{\text{inc}} - p_t/2$. Furthermore, $p^{>}_{\text{exc}} \leq p^{>}_{\text{Toc}} \leq p^{>}_{\text{inc}}$ and $p^{>}_{\text{exc}} < p^{>}_{\text{mid}} < p^{>}_{\text{inc}}$.

For testing the one-sided alternative $H_o : \theta = 1$ versus $H_a : \theta < 1$, the p-values are as above, except the inequalities are reversed and the corresponding p-values are denoted as $p^{<}_{\text{inc}}$, $p^{<}_{\text{exc}}$, $p^{<}_{\text{mid}}$, and $p^{<}_{\text{Toc}}$, respectively.

3.1.2 Variants of Fisher's Exact Test for Two-Sided Alternatives

There is additional complexity in p-values for the two-sided alternatives. A recommended method is adding up probabilities of the same size and smaller than the probability associated with the current table. Alternatively, one can double the one-sided p-value (Agresti [1]).

Type (I): For double the one-sided p-value, we propose the following four variants:

(i) twice the minimum of p_{inc} for the one-sided tests, which is table-inclusive version for this type of two-sided test, and denoted as p^{I}_{inc},

(ii) twice the minimum of p_{inc} minus twice the table probability p_t, which is table-exclusive version of this type of two-sided test, and denoted as p^{I}_{exc},

(iii) table-exclusive version of this type of two-sided test plus p_t, which is mid-p-value for this test, and denoted as p^{I}_{mid},

(iv) Tocher corrected version, p^{I}_{Toc}, is calculated as in Equation (1).

Notice that $p^{I}_{\text{inc}} = 2 \min(p^{>}_{\text{inc}}, p^{<}_{\text{inc}})$, $p^{I}_{\text{exc}} = 2 \min(p^{>}_{\text{exc}}, p^{<}_{\text{exc}}) = p^{I}_{\text{inc}} - 2\, p_t$, and $p^{I}_{\text{mid}} = p^{I}_{\text{exc}} + p_t$. Furthermore, $p^{I}_{\text{exc}} \leq p^{I}_{\text{Toc}} \leq p^{I}_{\text{inc}}$ and $p^{I}_{\text{exc}} < p^{I}_{\text{mid}} < p^{I}_{\text{inc}}$.

Type (II): For summing the p-values of more extreme —than that of the table— cases in both directions, the following variants are proposed. The p-value is $p = \sum_S f(t|n_1, n_2, c_1; \theta = 1)$ with

(i) $S = \{t : f(t|n_1, n_2, c_1; \theta = 1) \leq p_t\}$, which is called *table-inclusive version*, p^{II}_{inc},

(ii) the probability of the observed table is included twice, once for each side; that is $p = p^{II}_{\text{inc}} + p_t$, which is called *twice-table-inclusive version*, $p^{II}_{\text{t,inc}}$,

(iii) table-inclusive minus p_t, which is referred as *table-exclusive version*, p^{II}_{exc},

(iv) table-exclusive plus one-half the p_t, which is called *mid-p version*, p^{II}_{mid} and,

(v) *Tocher corrected version*, p^{II}_{Toc}, is obtained as before.

Note that $p^{II}_{\text{exc}} = p^{II}_{\text{inc}} - p_t$ and $p^{II}_{\text{mid}} = p^{II}_{\text{exc}} + p_t/2$. Furthermore, $p^{II}_{\text{exc}} \leq p^{II}_{\text{Toc}} \leq p^{II}_{\text{inc}} < p^{II}_{\text{t,inc}}$ and $p^{II}_{\text{exc}} < p^{II}_{\text{mid}} < p^{II}_{\text{inc}} < p^{II}_{\text{t,inc}}$.

3.2 Exact Version of Pearson's Test

Let μ_{ij} be the expected cell count in cell (i, j) of a contingency table. Pearson's χ^2 test is based on the test statistic

$$\mathcal{X}^2 = \sum_{j=1}^{2} \sum_{i=1}^{2} (N_{ij} - \mu_{ij})^2 / \mu_{ij},$$

which has χ^2_1 distribution in the limit provided that the contingency table is constructed from one of Poisson, row-wise, or overall multinomial sampling frameworks (Conover [10]). The exact version of Pearson's test uses the exact distribution of $\mathcal{X}^2$ rather than large sample χ^2 approximation. That is, for the two-sided alternative, we calculate the p-values as $p = \sum_S f(t|n_1, n_2, c_1; \theta = 1)$ with

(i) $S = \{t : |t - \mu_{11}| \geq |n_{11} - \mu_{11}|\}$, which is called *table-inclusive version*, p^{χ}_{inc},

(ii)-(v) the *twice-table-inclusive version*, $p^{\chi}_{\text{t,inc}}$, *table-exclusive version*, p^{χ}_{exc}, *mid-p version*, p^{χ}_{mid}, and *Tocher corrected version*, p^{χ}_{Toc}, are obtained as p^{II}_{exc}, p^{II}_{mid}, and p^{II}_{Toc}, respectively.

Observe that in all of these versions of exact tests, the most conservative ones are the table-inclusive versions, and the least conservative are the table-exclusive versions.

4. Nearest Neighbor Contingency Tables

Consider two classes labeled as $\{1, 2\}$. Let n_i be the number of points from class i for $i \in \{1, 2\}$ and $n = n_1 + n_2$. If we record the class of each point and of its nearest neighbor, the NN relationships fall into four 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 (classes 1 and 2 combined together) forms n (base,NN) pairs. Then each pair is categorized according to the base label (row categories) and NN label (column categories). Denoting N_{ij} as the frequency of cell (i, j) for $i, j \in \{1, 2\}$, we obtain 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 $j \in \{1, 2\}$.

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 two-sided

alternative is that some cell counts are different from those expected under CSR independence or RL.

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

Pielou used the usual Pearson's χ^2 test of independence for testing deviations from randomness in NN-structure (Pielou [27]), but the distribution of cell counts was not appropriate (Meagher and Burdick [24]). Dixon derived the appropriate asymptotic distribution of the cell counts using Moran join count statistics (Moran [25]) and hence the appropriate test which also has a χ^2-distribution (Dixon [14]). Moreover, Dixon's test is acceptable for moderate to large sample sizes (e.g., class sizes $n_i \geq 30$).

Ceyhan [8] compared these tests and demonstrated that Pielou's test is only appropriate for a random sample of (base,NN) pairs. Although, both tests are shown to be consistent (in the sense that, the power tends to 1 as sample sizes go to infinity under the alternatives), only Dixon's test has the appropriate nominal level, while Pielou's test is liberal in rejecting CSR independence or RL.

4.1 Exact Inference on NNCTs

Pielou's test and Dixon's tests are based on χ^2 approximation for large samples (i.e., row sums). As the sample sizes increase, the approximation for Dixon's test gets better for testing spatial segregation. When row and column sums are small, exact inference might be an alternative to the asymptotic NNCT tests.

For N_{11}, conditional on marginal sums, to follow a hypergeometric distribution, cell counts should originate from either Poisson or row-wise multinomial sampling frameworks (Conover [10]). In the overall multinomial sampling framework, N_{11} approximately follows a hypergeometric distribution. However, in the case of NNCTs, cell counts do not conform to any of these sampling frameworks (Ceyhan [8]).

In all these frameworks, the dependence between cell counts, and the dependence between rows of the NNCTs are caused by the spatial correlation of points. For spatial data, a (base, NN) pair is more likely to be a reflexive pair, rather than a non-reflexive one under CSR independence or RL. A (base,NN) pair (x, y) is *reflexive* if (y, x) is also a (base,NN) pair, that is, if x is a NN of y and y is a NN of x. Furthermore, a point can serve as a NN up to six other points in $\mathbb{R}^2$;

that is, a point can be shared as a NN by as many as six points. The spatial correlation between (base,NN) pairs and hence between cell counts in NNCTs are due to reflexivity and shared NN structure. Furthermore, these problems persist in the asymptotics also (Clark and Evans [9]). Thus, under any of Poisson, row-wise multinomial, and overall multinomial sampling frameworks, as in the case of Pielou's tests, Fisher's exact test is not appropriate for testing CSR independence or RL. The independence of rows and individual trials (base,NN pairs) would follow if NNCT is based on a random sample of (base,NN) pairs, which unfortunately is not realistic in practical situations.

While Pielou's test is liberal in rejecting CSR independence or RL with levels almost twice the nominal level (Ceyhan [8]), exact tests for contingency tables are conservative in general. Hence for exact tests, there are two competing factors: liberalness due to the inherent dependence between cell counts, and conservativeness due to the exact nature of the tests. Thus, these two factors might counterbalance each other, thereby rendering a variant of exact tests appropriate for CSR independence or RL.

Under CSR independence or RL, the odds ratio in the NNCT is given by $\theta = \frac{\mu_{11}\,\mu_{22}}{\mu_{12}\,\mu_{21}} = \frac{(n_1-1)(n_2-1)}{n_1\,n_2}$, since

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

See Dixon [14] for the derivation of the expectation in Equation (2). Observe that the expected cell counts depend only on the size of each class (i.e., row sums), but not on column sums. In a NNCT, row sums are fixed, while column sums are random quantities. On the other hand, Pielou's test and Fisher's exact test depend on both row and column sums. This is part of the problem, but the greater problem is the dependence due to reflexivity and shared NN structure. Notice also that as $n_i \to \infty$ for both $i = 1, 2$, the odds ratio $\theta \to 1$. Therefore, the null hypothesis for Fisher's exact test, namely, $H_o : \theta = 1$ is only equivalent to testing CSR independence or RL in the limit. On the other hand, for moderate to large sample sizes, under CSR independence or RL, $\theta \approx 1$. Hence, even if the distribution of the cell counts were correct, Fisher's exact test in the general form would be an approximate test for CSR independence or RL.

As N_{11} gets large, the odds ratio also gets larger than 1. Hence the alternative $H_a : \theta > 1$ is approximately (exactly) equivalent to the segregation alternative for finite samples (in the limit). The same holds for $H_a : \theta < 1$ under the association alternative. On the other hand, the two-sided alternative $H_a : \theta \neq 1$ is asymptotically equivalent to any deviation from CSR independence or RL. Since Fisher's exact test is a finite sample test, under CSR independence or RL, $\theta = \frac{(n_1-1)(n_2-1)}{n_1\,n_2} < 1$, which implies, if N_{11} had a hypergeometric distribution, it would have been a non-central hypergeometric distribution with non-centrality parameter, θ.

When only row sums in a contingency table are fixed with independent binomial samples, alternative exact tests are available. These tests are less conservative than Fisher's exact test (Berger and Boos [5]). However, the rows in a NNCT are not independent binomial samples, and Fisher's exact test is already mildly liberal in the best case, so we do not consider this approach for NNCTs (see Section 5).

We are applying exact and 'corrected' (i.e., a correction on the odds ratio is implemented by setting $\theta = \frac{(n_1-1)(n_2-1)}{n_1 n_2}$ for H_o.) exact tests for the usual test of independence in a 2×2 contingency table which has 1 df. Hence they turn out to be the exact versions of Pielou's test but not of Dixon's test. Alas, for Dixon's test, the exact version is not available, so we are left with Monte Carlo tests.

Remark 1. Consistency is an asymptotic property and exact tests are designed and used for finite samples. But as $n \to \infty$, the null hypothesis becomes $H_o : \theta = 1$ under CSR independence or RL, and Fisher's exact tests become equivalent to Pearson's χ^2 test. So Fisher's exact tests are also consistent, since Pielou's test is Pearson's χ^2 test for NNCTs, and is proved to be consistent by Ceyhan [8]. □

5. Empirical Significance Levels of the Tests under CSR Independence

We implement Monte Carlo simulations to evaluate the finite sample performance of the tests in terms of empirical size. For the null case, we simulate the CSR independence pattern only, with classes X and Y of sizes n_1 and n_2, respectively. At each of $N_{mc} = 10000$ replicates, under H_o, we generate data for some combinations of $n_1, n_2 \in \{10, 20, 30, 50, 100, 150, 200\}$ points independent identically distributed (iid) from $\mathcal{U}((0,1) \times (0,1))$, the uniform distribution on the unit square. Let $\{X_1, \ldots, X_{n_1}\}$ be the set of class X points and $\{Y_1, \ldots, Y_{n_2}\}$ be the set of class Y points.

As the tests under consideration, namely, Fisher's exact test, and exact version of Pearson's test, are not appropriate in testing the null pattern of CSR independence or RL, we do not perform power comparisons under various alternatives for all the versions right away. We first investigate which variant(s) are more appropriate (i.e., have the empirical sizes about the nominal level, α). We present the empirical significance levels for the tests under H_o : *CSR independence* with or without the correction on odds ratio against segregation or association alternatives in Table 2. The empirical sizes significantly smaller (larger) than .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 size estimates from the nominal level of .05. For these proportion tests, we also use $\alpha = .05$ to test against empirical size being equal to .05. With $N_{mc} = 10000$, empirical sizes less (greater) than .0464 (.0536) are deemed conservative (liberal) at $\alpha = .05$ level.

Table 2 The empirical significance levels without correction on odds ratio (top) and with finite sample correction (bottom) for the one-sided tests under H_o : CSR *independence* with $N_{mc} = 10000$, for some combinations of $n_1, n_2 \in \{10, 20, 30\}$ at $\alpha = .05$. $\hat{\alpha}^S_{\text{inc}}$ is the estimated empirical significance level for the table-inclusive version of the one-sided test of segregation, $\hat{\alpha}^S_{\text{exc}}$ is for the table-exclusive version, $\hat{\alpha}^S_{\text{mid}}$ is for the mid-p-value version, $\hat{\alpha}^S_{\text{Toc}}$ is for the Tocher corrected version. The notation is similar for the association alternative with S being replaced by A. ^{c}:The empirical size is significantly smaller than .05; i.e., the test is conservative. $^\ell$:The empirical size is significantly larger than .05; i.e., the test is liberal.

Empirical significance levels for the one-sided tests								
(n_1, n_2)	$\hat{\alpha}^S_{\text{inc}}$	$\hat{\alpha}^S_{\text{exc}}$	$\hat{\alpha}^S_{\text{mid}}$	$\hat{\alpha}^S_{\text{Toc}}$	$\hat{\alpha}^A_{\text{inc}}$	$\hat{\alpha}^A_{\text{exc}}$	$\hat{\alpha}^A_{\text{mid}}$	$\hat{\alpha}^A_{\text{Toc}}$
(10, 10)	$.0421^c$ $.0440^c$	$.1504^\ell$ $.1554^\ell$	$.0619^\ell$ $.0833^\ell$	$.0692^\ell$ $.0984^\ell$	$.0839^\ell$ $.0562^\ell$	$.2448^\ell$ $.1786^\ell$	$.1229^\ell$ $.0872^\ell$	$.1295^\ell$ $.1016^\ell$
(10, 20)	$.0440^c$ $.0622^\ell$	$.1347^\ell$ $.1789^\ell$	$.0670^\ell$ $.0949^\ell$	$.0736^\ell$ $.0960^\ell$	$.0836^\ell$ .0503	$.2330^\ell$ $.1768^\ell$	$.1158^\ell$ $.1012^\ell$	$.1310^\ell$ $.0997^\ell$
(10, 30)	.0499 .0532	$.1481^\ell$ $.1575^\ell$	$.0685^\ell$ $.0770^\ell$	$.0767^\ell$ $.0913^\ell$	$.0584^\ell$ $.0625^\ell$	$.2310^\ell$ $.2329^\ell$	$.1183^\ell$ $.1190^\ell$	$.1328^\ell$ $.1154^\ell$
(20, 10)	$.0437^c$ $.0575^\ell$	$.1371^\ell$ $.1803^\ell$	$.0684^\ell$ $.0933^\ell$	$.0749^\ell$ $.0922^\ell$	$.0778^\ell$ .0519	$.2314^\ell$ $.1767^\ell$	$.1100^\ell$ $.0988^\ell$	$.1291^\ell$ $.1001^\ell$
(20, 20)	$.0406^c$ $.0689^\ell$	$.1070^\ell$ $.1578^\ell$	$.0651^\ell$ $.1066^\ell$	$.0736^\ell$ $.0948^\ell$	$.0731^\ell$ $.0683^\ell$	$.1697^\ell$ $.1630^\ell$	$.1142^\ell$ $.1080^\ell$	$.1257^\ell$ $.0979^\ell$
(20, 30)	.0487 $.0637^\ell$	$.1182^\ell$ $.1482^\ell$	$.0716^\ell$ $.0956^\ell$	$.0760^\ell$ $.1016^\ell$	$.0786^\ell$ $.0671^\ell$	$.1682^\ell$ $.1515^\ell$	$.1154^\ell$ $.0931^\ell$	$.1170^\ell$ $.0944^\ell$
(30, 10)	.0486 $.0575^\ell$	$.1438^\ell$ $.1635^\ell$	$.0675^\ell$ $.0803^\ell$	$.0749^\ell$ $.0947^\ell$	$.0578^\ell$ $.0590^\ell$	$.2339^\ell$ $.2313^\ell$	$.1213^\ell$ $.1163^\ell$	$.1339^\ell$ $.1103^\ell$
(30, 20)	.0530 $.0594^\ell$	$.1287^\ell$ $.1322^\ell$	$.0771^\ell$ $.0883^\ell$	$.0830^\ell$ $.0918^\ell$	$.0799^\ell$ $.0744^\ell$	$.1767^\ell$ $.1583^\ell$	$.1204^\ell$ $.1001^\ell$	$.1174^\ell$ $.1000^\ell$
(30, 30)	.0536 $.0732^\ell$	$.1093^\ell$ $.1451^\ell$	$.0765^\ell$ $.0749^\ell$	$.0775^\ell$ $.0901^\ell$	$.0787^\ell$ $.0763^\ell$	$.1609^\ell$ $.1564^\ell$	$.1152^\ell$ $.0837^\ell$	$.1161^\ell$ $.0953^\ell$

Observe that without correction on odds ratio, all empirical sizes for segregation alternative are smaller than the corresponding ones for the association alternative for $n_i \leq 50$. The correction on odds ratio slightly increases (decreases) the empirical levels for segregation (association) alternatives. The increase in empirical levels for segregation alternatives are significant for most sample size combinations, while the decrease in empirical levels for association alternatives is significant for some of them. Without correction on odds ratio, odds ratio is less than 1, hence the test tends to reject for the segregation alternative less than that under CSR independence, and vice versa for association alternative. The order of estimated empirical sizes from smallest

to largest is table-inclusive, mid-p-value, Tocher corrected version, and table-exclusive versions. This ordering is also from most conservative to the least conservative versions for testing independence in general contingency tables.

Table 3 The empirical significance levels without correction on odds ratio (top), with finite sample correction (middle), and the mixed method for the two-sided tests under H_o with $N_{mc} = 10000$, for some combinations of $n_1, n_2 \in \{10, 20, 30\}$ at $\alpha = .05$. $\hat{\alpha}^I_{\text{inc}}$ is the estimated empirical significance level for the two-sided test of segregation with the table-inclusive version of Pearson type I test, $\hat{\alpha}^I_{\text{exc}}$ is for table-exclusive version, $\hat{\alpha}^I_{\text{mid}}$ is for mid-p-value version, $\hat{\alpha}^I_{\text{Toc}}$ is for Tocher corrected version. Superscript labeling is as in Table 2.

Empirical significance levels for the two-sided tests									
(n_1, n_2)	$\hat{\alpha}^I_{\text{inc}}$	$\hat{\alpha}^I_{\text{exc}}$	$\hat{\alpha}^I_{\text{mid}}$	$\hat{\alpha}^I_{\text{Toc}}$	$\hat{\alpha}^{II}_{\text{inc}}$	$\hat{\alpha}^{II}_{\text{exc}}$	$\hat{\alpha}^{II}_{\text{mid}}$	$\hat{\alpha}^{II}_{\text{Toc}}$	$\hat{\alpha}^{II}_{\text{t,inc}}$
(10, 10)	$.0592^{\ell}$ $.0866^{\ell}$	$.2354^{\ell}$ $.3016^{\ell}$	$.1260^{\ell}$ $.0896^{\ell}$	$.1272^{\ell}$ $.1313^{\ell}$	$.0592^{\ell}$ $.0877^{\ell}$	$.1260^{\ell}$ $.1711^{\ell}$	$.0738^{\ell}$ $.1207^{\ell}$	$.0917^{\ell}$ $.1334^{\ell}$	$.0592^{\ell}$ $.0866^{\ell}$
(10, 20)	$.0839^{\ell}$ $.0693^{\ell}$	$.2725^{\ell}$ $.2610^{\ell}$	$.1193^{\ell}$ $.0992^{\ell}$	$.1346^{\ell}$ $.1229^{\ell}$	$.1072^{\ell}$ $.0893^{\ell}$	$.2000^{\ell}$ $.1813^{\ell}$	$.1331^{\ell}$ $.1156^{\ell}$	$.1354^{\ell}$ $.1246^{\ell}$	$.0686^{\ell}$ $.0600^{\ell}$
(10, 30)	$.0574^{\ell}$ .0521	$.2737^{\ell}$ $.2850^{\ell}$	$.0986^{\ell}$ $.1137^{\ell}$	$.1290^{\ell}$ $.1266^{\ell}$	$.0929^{\ell}$ $.0882^{\ell}$	$.1787^{\ell}$ $.2071^{\ell}$	$.1071^{\ell}$ $.1037^{\ell}$	$.1252^{\ell}$ $.1288^{\ell}$	$.0421^{c}$ $.0417^{c}$
(20, 10)	$.0766^{\ell}$ $.0702^{\ell}$	$.2720^{\ell}$ $.2597^{\ell}$	$.1129^{\ell}$ $.0974^{\ell}$	$.1298^{\ell}$ $.1219^{\ell}$	$.1010^{\ell}$ $.0896^{\ell}$	$.1952^{\ell}$ $.1811^{\ell}$	$.1272^{\ell}$ $.1162^{\ell}$	$.1302^{\ell}$ $.1235^{\ell}$	$.0618^{\ell}$ $.0613^{\ell}$
(20, 20)	$.0715^{\ell}$ $.0820^{\ell}$	$.1895^{\ell}$ $.2147^{\ell}$	$.1120^{\ell}$ $.1372^{\ell}$	$.1297^{\ell}$ $.1211^{\ell}$	$.0715^{\ell}$ $.0830^{\ell}$	$.1120^{\ell}$ $.1393^{\ell}$	$.1117^{\ell}$ $.1372^{\ell}$	$.1034^{\ell}$ $.1249^{\ell}$	$.0683^{\ell}$ $.0820^{\ell}$
(20, 30)	$.0798^{\ell}$ $.0778^{\ell}$	$.1949^{\ell}$ $.2010^{\ell}$	$.1086^{\ell}$ $.1069^{\ell}$	$.1205^{\ell}$ $.1204^{\ell}$	$.1054^{\ell}$ $.1020^{\ell}$	$.1648^{\ell}$ $.1692^{\ell}$	$.1116^{\ell}$ $.1086^{\ell}$	$.1209^{\ell}$ $.1232^{\ell}$	$.0704^{\ell}$ $.0699^{\ell}$
(30, 10)	.0525 .0523	$.2760^{\ell}$ $.2868^{\ell}$	$.0932^{\ell}$ $.1146^{\ell}$	$.1237^{\ell}$ $.1221^{\ell}$	$.0878^{\ell}$ $.0894^{\ell}$	$.1785^{\ell}$ $.2020^{\ell}$	$.1017^{\ell}$ $.1039^{\ell}$	$.1216^{\ell}$ $.1266^{\ell}$	$.0362^{c}$ $.0425^{c}$
(30, 20)	$.0833^{\ell}$ $.0804^{\ell}$	$.2076^{\ell}$ $.1985^{\ell}$	$.1163^{\ell}$ $.1101^{\ell}$	$.1286^{\ell}$ $.1230^{\ell}$	$.1132^{\ell}$ $.1064^{\ell}$	$.1733^{\ell}$ $.1677^{\ell}$	$.1188^{\ell}$ $.1128^{\ell}$	$.1296^{\ell}$ $.1254^{\ell}$	$.0722^{\ell}$ $.0716^{\ell}$
(30, 30)	$.0844^{\ell}$ $.0728^{\ell}$	$.1917^{\ell}$ $.1665^{\ell}$	$.1321^{\ell}$ $.0993^{\ell}$	$.1238^{\ell}$ $.1194^{\ell}$	$.0844^{\ell}$ $.0992^{\ell}$	$.1321^{\ell}$ $.1495^{\ell}$	$.0850^{\ell}$ $.1001^{\ell}$	$.0990^{\ell}$ $.1162^{\ell}$	$.0587^{\ell}$ $.0641^{\ell}$

For the segregation alternative, the table-inclusive version has about the nominal level, while mid-p, and Tocher corrected versions are slightly liberal, and the table-exclusive version is the most liberal. For the table-inclusive version, the one without correction on odds ratio is closer to the nominal level. On the other hand, for the association alternative, table-inclusive version is mildly liberal but table-inclusive version with correction on odds ratio is closer to the desired nominal level compared to other versions.

We present the empirical significance levels for the two-sided tests in Tables 3 and 4 where we use the empirical significance levels for Dixon's test, $\hat{\alpha}_D$ as a benchmark. Observe that among type I tests, table inclusive version with correction on odds ratio has empirical size closest to the nominal level, among type II tests, twice-table-inclusive without correction on odds ratio has empirical size closest to the nominal level. Among exact versions of Pearson's test, twice-table-inclusive with correction on odds ratio is about the desired nominal level.

Table 4 The empirical significance levels without correction on odds ratio (top) and with finite sample correction (bottom) for the exact version of Pearson's test under H_o with $N_{mc} = 10000$, for some combinations of $n_1, n_2 \in \{10, 20, 30\}$ at $\alpha = .05$. $\hat{\alpha}_D$ stands for the empirical size estimate for Dixon's test, and the other size notations are as in Table 3, with I (or II) replaced by χ. Superscript labeling is as in Table 2.

Empirical significance levels for Dixon's test and the exact version of Pearson's tests						
(n_1, n_2)	$\hat{\alpha}^{\chi}_{\text{inc}}$	$\hat{\alpha}^{\chi}_{\text{exc}}$	$\hat{\alpha}^{\chi}_{\text{mid}}$	$\hat{\alpha}^{\chi}_{\text{Toc}}$	$\hat{\alpha}^{\chi}_{\text{t,inc}}$	$\hat{\alpha}_D$
(10, 10)	$.0592^{\ell}$ $.0813^{\ell}$	$.1260^{\ell}$ $.1918^{\ell}$	$.0738^{\ell}$ $.1007^{\ell}$	$.0904^{\ell}$ $.1087^{\ell}$	$.0592^{\ell}$ .0523	$.0432^{c}$ —
(10, 20)	$.1072^{\ell}$ $.0845^{\ell}$	$.1671^{\ell}$ $.1972^{\ell}$	$.1276^{\ell}$ $.0930^{\ell}$	$.1274^{\ell}$ $.1225^{\ell}$	$.0624^{\ell}$ .0530	$.0422^{c}$ —
(10, 30)	$.0710^{\ell}$ .0467	$.1488^{\ell}$ $.1866^{\ell}$	$.0826^{\ell}$ $.1188^{\ell}$	$.0922^{\ell}$ $.1088^{\ell}$	$.0321^{c}$ $.0317^{c}$	$.0424^{c}$ —
(20, 10)	$.1010^{\ell}$ $.0756^{\ell}$	$.1621^{\ell}$ $.2020^{\ell}$	$.1215^{\ell}$ $.0923^{\ell}$	$.1220^{\ell}$ $.1161^{\ell}$	$.0552^{\ell}$ $.0381^{c}$	$.0406^{c}$ —
(20, 20)	$.0715^{\ell}$ $.0683^{\ell}$	$.1120^{\ell}$ $.1622^{\ell}$	$.1117^{\ell}$ $.0982^{\ell}$	$.1024^{\ell}$ $.1063^{\ell}$	$.0683^{\ell}$ $.0551^{\ell}$	$.0432^{c}$ —
(20, 30)	$.1054^{\ell}$ $.0810^{\ell}$	$.1522^{\ell}$ $.1427^{\ell}$	$.1086^{\ell}$ $.1064^{\ell}$	$.1176^{\ell}$ $.1043^{\ell}$	$.0650^{\ell}$ .0524	.0469 —
(30, 10)	$.0643^{\ell}$ $.0743^{\ell}$	$.1506^{\ell}$ $.1936^{\ell}$	$.0748^{\ell}$ $.1041^{\ell}$	$.0889^{\ell}$ $.1139^{\ell}$	$.0259^{c}$ $.0440^{c}$	$.0389^{c}$ —
(30, 20)	$.1132^{\ell}$ $.0796^{\ell}$	$.1612^{\ell}$ $.1537^{\ell}$	$.1162^{\ell}$ $.1000^{\ell}$	$.1259^{\ell}$ $.1087^{\ell}$	$.0664^{\ell}$ $.0539^{\ell}$	.0479 —
(30, 30)	$.0844^{\ell}$ $.0789^{\ell}$	$.1321^{\ell}$ $.1586^{\ell}$	$.0850^{\ell}$ $.1020^{\ell}$	$.0958^{\ell}$ $.1061^{\ell}$	$.0587^{\ell}$ $.0616^{\ell}$	$.0453^{c}$ —

6. Empirical Power Analysis

For the right-sided test (i.e., the test relative to segregation alternative) we use the table-inclusive versions of Fisher's exact test without correction on odds ratio. For the left-sided test

(i.e., the test relative to association alternative) we use the table-inclusive versions of Fisher's exact test with correction on odds ratio. For the two-sided alternative, we consider twice-table-inclusive version of type I Fisher test with correction on odds ratio, type II Fisher test without correction on odds ratio, and twice-table-inclusive version of Pearson's exact test with correction on odds ratio.

6.1 Empirical Power Analysis under the Segregation Alternatives

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

$$H_S^I : s = 1/6, \quad H_S^{II} : s = 1/4, \text{ and } H_S^{III} : s = 1/3.$$

These alternatives are illustrated with $n_1 = 100$ X and $n_2 = 100$ Y points in Figure 1.

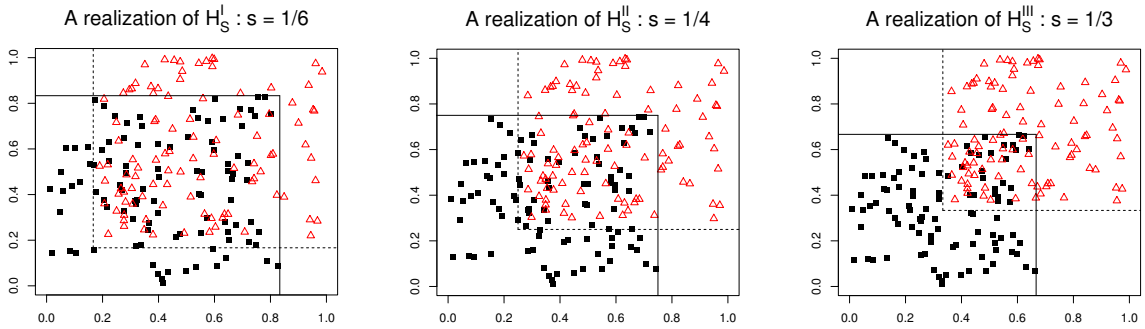

Figure 1 Three realizations for $H_S^I : s = 1/6$ (left), $H_S^{II} : s = 1/4$ (middle), and $H_S^{III} : s = 1/3$ (right) with $n_1 = 100$ X points (solid squares ■) and $n_2 = 100$ Y points (triangles △).

Observe that, from H_S^I to H_S^{III} (i.e., as s increases), the segregation between X and Y gets stronger in the sense that X and Y points tend to form one-class clumps or clusters. The power estimates are presented in Table 5. By construction these segregation alternatives are symmetric in the sense that degree of segregation of class X in (n_1,n_2) case is same as that of class Y in (n_2,n_1) case, so we only present $n_1 \leq n_2$ cases. Observe that, with the correction on odds ratio, the empirical power estimates increase (significantly for most cases). As $n = (n_1+n_2)$ gets larger, the power estimates get larger. Furthermore, as the segregation gets more severe, the power estimates get larger. The highest power estimates are usually attained by the right-sided (i.e., relative to segregation) test. The left-sided tests (i.e., tests relative to association) have virtually zero power (hence not presented). Considering the empirical significance levels and power estimates, when testing against segregation, we recommend the table-inclusive version of the right-sided Fisher's exact test without correction on odds ratio for small to moderate samples (i.e., $n_i < 30$), and Dixon's test for larger samples.

Table 5 The empirical power estimates under the segregation alternatives with $N_{mc} = 10000$ Monte Carlo replicates, for some combinations of $n_1, n_2 \in \{10, 20, 30\}$ at $\alpha = .05$. $\hat{\beta}^{S,woc}_{\text{inc}}$ stands for the power estimate of the table-inclusive version of Fisher's exact test for the right-sided alternative (without correction on the odds ratio), $\hat{\beta}^{I,wc}_{\text{inc}}$ for the table-inclusive version of type I of Fisher's exact test (with correction on the odds ratio), $\hat{\beta}^{II,woc}_{\text{t,inc}}$ for the twice-table-inclusive version of type II of Fisher's exact test (without correction on the odds ratio), $\hat{\beta}^{\chi,wc}_{\text{t,inc}}$ for the twice-table-inclusive version of Pearson's exact test (with correction on the odds ratio), and $\hat{\beta}_D$ for Dixon's test.

Empirical power estimates under H_S						
	(n_1, n_2)	$\hat{\beta}^{S,woc}_{\text{inc}}$	$\hat{\beta}^{I,wc}_{\text{inc}}$	$\hat{\beta}^{II,woc}_{\text{t,inc}}$	$\hat{\beta}^{\chi,wc}_{\text{t,inc}}$	$\hat{\beta}_D$
H^I_S	(10, 10)	.1532	.1590	.0890	.0888	.0775
	(10, 20)	.2105	.2071	.1574	.1584	.1064
	(10, 30)	.2593	.2188	.1721	.2049	.1414
	(20, 20)	.3295	.3294	.2414	.2414	.1597
	(20, 30)	.4315	.3932	.2988	.3309	.2121
	(30, 30)	.5344	.4749	.3825	.4080	.2904
H^{II}_S	(10, 10)	.4287	.4292	.2842	.2842	.2305
	(10, 20)	.5720	.5657	.4864	.4876	.3643
	(10, 30)	.6599	.6069	.5352	.5917	.4555
	(20, 20)	.7849	.7838	.6925	.6925	.5379
	(20, 30)	.8887	.8691	.7935	.8160	.6774
	(30, 30)	.9515	.9308	.8896	.9006	.8141
H^{III}_S	(10, 10)	.8197	.8197	.6890	.6890	.5817
	(10, 20)	.9286	.9272	.8844	.8846	.7837
	(10, 30)	.9633	.9513	.9245	.9466	.8787
	(20, 20)	.9923	.9923	.9808	.9808	.9375
	(20, 30)	.9984	.9978	.9946	.9954	.9819
	(30, 30)	.9997	.9994	.9989	.9991	.9969

As for the two-sided tests, the power estimates of the exact tests are significantly larger than power estimates for Dixon's test. Considering their empirical level performance, if one wants to conduct a two-sided test, we recommend twice-table-inclusive version of exact version of Pearson's test with correction on odds ratio for small sample sizes (i.e., $n_i < 30$) and for larger

samples we recommend Dixon's test.

6.2 Empirical Power Analysis under the Association Alternatives

For the association alternatives, we also consider three cases. First, we generate $X_i \overset{iid}{\sim} \mathcal{U}((0,1)\times(0,1))$ for $i = 1,2,\ldots,n_1$ and Y_j for $j = 1,2,\ldots,n_2$ as follows. For each j, we pick an i randomly, then generate $R_j \sim \mathcal{U}(0,r)$ with $r \in (0,1)$ and $T_j \sim \mathcal{U}(0,2\pi)$ and set $Y_j = X_i + R_j\,(\cos T_j, \sin T_j)'$. Appropriate choices of r will imply that classes X and Y are more associated. That is, it is more likely to have (X,Y) NNs than same-class NNs $((X,X)$ or $(Y,Y))$. The three values of r we consider constitute the three association alternatives;

$$H_A^I : r = 1/4, \quad H_A^{II} : r = 1/7, \text{ and } H_A^{III} : r = 1/10.$$

These alternatives are illustrated with $n_1 = 20$ X and $n_2 = 100$ Y points in Figure 2. Observe that, from H_A^I to H_A^{III}, the association gets more severe in the sense that X and Y points tend to occur together more and more frequently.

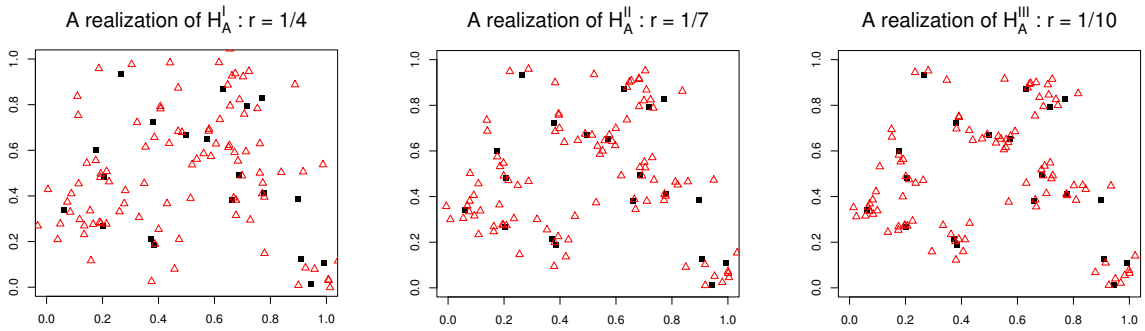

Figure 2 Three realizations for $H_A^I : r = 1/4$ (left), $H_A^{II} : r = 1/7$ (middle), and $H_A^{III} : r = 1/10$ (right) with $n_1 = 20$ X points (solid squares ■) and $n_2 = 100$ Y points (triangles △).

The power estimates are presented in Table 6. Observe that for each sample size combination, as the association gets more severe, the power estimates get larger. The highest power estimates are naturally attained by the one-sided association test among exact tests. The right-sided tests (i.e., tests relative to segregation) have virtually zero power (hence not presented). Furthermore, by construction the larger the class Y from class X, the stronger the association between them, while the larger the class X from class Y the weaker the association. Considering the empirical significance levels and power estimates, when testing against association, for small sample sizes (i.e., $n_i < 30$), we recommend the table-inclusive version of left-sided Fisher's exact test with correction on odds ratio. For larger samples, we recommend Dixon's test.

Table 6 The empirical power estimates under the association alternatives with $N_{mc} = 10000$ Monte Carlo replicates, for all combinations of $n_1, n_2 \in \{10, 20, 30\}$ at $\alpha = .05$. $\hat{\beta}_{\text{inc}}^{A,wc}$ stands

for the power estimate of the table-inclusive version of Fisher's exact test for the left-sided alternative (with correction on the odds ratio), $\hat{\beta}_{\text{inc}}^{I,wc}$, $\hat{\beta}_{\text{t,inc}}^{II,woc}$, $\hat{\beta}_{\text{t,inc}}^{\chi,wc}$, and $\hat{\beta}_D$ are as in Table 5.

Empirical power estimates under H_A						
	(n_1, n_2)	$\hat{\beta}_{\text{inc}}^{A,wc}$	$\hat{\beta}_{\text{inc}}^{I,wc}$	$\hat{\beta}_{\text{t,inc}}^{II,woc}$	$\hat{\beta}_{\text{t,inc}}^{\chi,wc}$	$\hat{\beta}_D$
H_A^I	(10, 10)	.3068	.2657	.2690	.2653	.1105
	(10, 20)	.4338	.3488	.3980	.3487	.2371
	(10, 30)	.4948	.3427	.3504	.2497	.3007
	(20, 10)	.1758	.1282	.1461	.1264	.0632
	(20, 20)	.4085	.3060	.3055	.3053	.1745
	(20, 30)	.4945	.3733	.4038	.3730	.2672
	(30, 10)	.1318	.0515	.0502	.0263	.0409
	(30, 20)	.3019	.1979	.2394	.1969	.1082
	(30, 30)	.4284	.2845	.2842	.2676	.1697
H_A^{II}	(10, 10)	.4804	.4398	.4422	.4398	.1834
	(10, 20)	.6719	.5746	.6341	.5746	.4007
	(10, 30)	.7259	.5880	.5983	.4847	.4956
	(20, 10)	.3311	.2516	.2977	.2511	.1427
	(20, 20)	.6941	.5873	.5871	.5870	.3622
	(20, 30)	.8101	.7078	.7383	.7077	.5258
	(30, 10)	.2623	.1293	.1307	.706	.1003
	(30, 20)	.5778	.4408	.4849	.4406	.2675
	(30, 30)	.7632	.6216	.6216	.6078	.4141
H_A^{II}	(10, 10)	.5476	.5116	.5138	.5116	.2222
	(10, 20)	.7680	.6806	.7351	.6806	.4853
	(10, 30)	.8228	.7006	.7107	.6037	.6003
	(20, 10)	.4262	.3365	.3913	.3363	.2153
	(20, 20)	.8199	.7382	.7381	.7381	.5063
	(20, 30)	.9169	.8525	.8765	.8525	.6869
	(30, 10)	.3527	.2091	.2112	.1305	.1783
	(30, 20)	.7378	.6169	.6526	.6169	.4268
	(30, 30)	.9025	.8081	.8081	.7966	.6157

For the two-sided tests, the power estimates of the exact tests are usually significantly larger than power estimates for Dixon's test. Considering their empirical level performance, if one wants to conduct a two-sided test, we recommend the table-inclusive version of the exact Pearson's test with correction on odds ratio for small sample sizes (i.e., $n_i < 30$); for larger sample sizes, we recommend Dixon's test.

Remark 2. **Edge Correction for the NNCT-Tests:** The CSR independence pattern assumes that the study region is unbounded for the analyzed pattern, which is not the case in practice. So it might be necessary to correct for edge effects under CSR independence (Yamada and Rogersen [32]). Two correction methods for the edge effects on exact and asymptotic NNCT-tests, namely *buffer zone correction* and *toroidal correction*, are investigated in (Ceyhan [6;8]) where it is shown that the empirical sizes of the NNCT-tests are not affected by the toroidal edge correction under CSR independence. However, toroidal correction is biased for non-CSR patterns. In particular if the pattern outside the plot (which is often unknown) is not the same as that inside it yields questionable results (Haase [19] and Yamada and Rogersen [32]). Under CSR independence, the (outer) buffer zone edge correction method does not change the sizes significantly for most sample size combinations. This is in agreement with the findings of Barot *et al*. [4] who say that NN methods only require a small buffer area around the study region. A large buffer area does not help much since once the buffer area extends past the likely NN distances (i.e., about the average NN distances), it is not adding much helpful information for NNCTs. Hence we recommend inner or outer buffer zone correction for these tests with the width of the buffer area being about the average NN distance. □

7. Examples

We illustrate the tests on three example data sets: swamp tree data (the whole region and a subset of the region) and an artificial data set. We conduct Pielou's and Dixon's test of segregation, table-inclusive version of Fisher's right-sided exact test (for segregation) without correction on odds ratio, table-inclusive version of Fisher's left-sided exact test (for association) with correction on odds ratio, table-inclusive version of type-I test with correction on odds ratio, twice-table-inclusive version of type II test without correction on odds ratio, twice-table-inclusive version of Pearson's exact test with correction on odds ratio.

7.1 Swamp Tree Data - The Whole Region

Dixon illustrates NN-methods on the tree species in a 50m by 200m rectangular plot of hardwood swamp in South Carolina, USA (Dixon [16]). The plot contains 13 different tree species, of which we only consider two, namely, bald cypress and black gum trees as they constitute the majority of the trees (there are 182 black gum trees and 91 bald cypresses). 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. The question of interest is whether these tree species exhibit segregation, association, or CSR independence. The locations of these trees in the study region are plotted in Figure 3 and the corresponding NNCT together with percentages are provided in Table 7 (top). Observe that the percentage values are suggestive of segregation for both species.

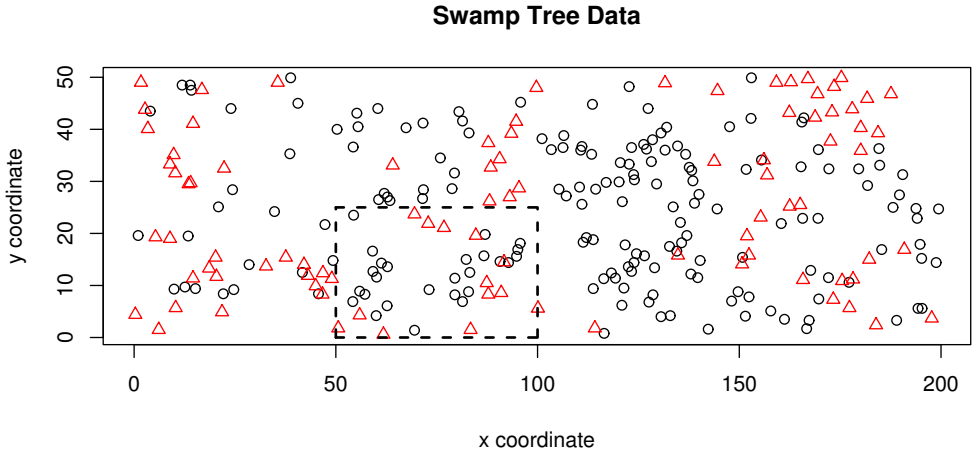

Figure 3 The scatter plots of the locations of black gum trees (circles ○) and cypress trees (triangles △). The smaller region (i.e., one-eighth of the region) is the rectangle bounded by the dashed lines.

The p-values are presented in Table 9, where we observe that all two-sided tests are significant, implying significant deviation from CSR independence; and the one-sided tests indicate that black gum trees and cypress trees are significantly segregated.

7.2 Swamp Tree Data - A Subset of the Region

Dixon's test is more reliable for the swamp tree data with the whole region, since the sample sizes are large enough for asymptotic approximation. Furthermore, the exact tests we consider become more similar to Pielou's test, hence get to be liberal in rejecting the null hypothesis for large samples. Moreover, exact tests are intended for use in small sample sizes where the asymptotic approximations fail; in particular exact tests we consider are conservative for small samples for general contingency tables, which makes them appropriate for NNCTs. Therefore, we also analyze the spatial pattern in a subset of the region (one-eighth of the region) in this data set (see the dashed box in Figure 3). This region contains 26 black gums and 12 bald cypresses, hence it is appropriate for using the exact tests, while the asymptotic approximations might fail. The corresponding NNCT and the percentages are also presented in Table 7 (bottom). Observe that the percentages for the whole region and the one-eighth of the region are very similar, suggesting the segregation of black gums and bald cypresses in this subregion also. The p-values are presented in Table 9, where it is seen that Dixon's test is not significant at .05

level, Pielou's test is significant at .01 level, and the exact tests are significant only at .05 level. Dixon's test misses the apparent segregation pattern here as it is conservative for small samples, Pielou's test suggests severe segregation, which (we think) is over-emphasized because of its liberalness. Thus, the exact tests more reliably indicate the mild segregation pattern and its significance.

Table 7 The NNCT for Swamp tree data (left) and the corresponding percentages (right) with the whole region (top) and the one-eighth of the region (bottom). B.G. = black gum trees, B.C. = bald cypress trees

Swamp Tree Data (The Whole Region)

		NN				NN		
		B.G.	B.C.	sum		B.G.	B.C.	
base	B.G.	149	33	182	base B.G.	82 %	18 %	67 %
	B.C.	43	48	91	B.C.	47 %	53 %	23 %
	sum	192	581	273		34 %	66 %	100 %

Swamp Tree Data (One-Eighth of the Region)

		NN				NN		
		B.G.	B.C.	sum		B.G.	B.C.	
base	B.G.	22	4	26	base B.G.	85 %	15 %	68 %
	B.C.	5	7	12	B.C.	42 %	58 %	32 %
	sum	27	11	38		71 %	29 %	100 %

7.3 Artificial Data

In this artificial example, a random sample of size 40 (with 20 X and 20 Y-points iid uniformly generated on the unit square). By construction, the locations of these points can be viewed a priori resulting from (possibly) different processes, so the more appropriate null hypothesis is the CSR independence pattern. The question of interest is the spatial interaction between X and Y points. We plot the locations of these points in the study region in Figure 4 and provide the corresponding NNCT together with percentages in Table 8. Observe that the percentages are slightly larger for the off-diagonal cells, which might be interpreted as presence of mild association for both classes.

Observe in Table 9 that among the two-sided tests, only Pielou's test is significant at .05 level, but the significance of Pielou's test seems to be a false alarm, as it is liberal test (Ceyhan [8]). However, notice that, p_{inc}^{A} is mildly significant which is in agreement with the NNCT-table and the figure. But the left-sided test for this particular sample size combination was

significantly liberal (see Table 2), so this result is not so reliable either. On the other hand two-sided tests suggest no significant deviation from CSR independence (as they are mildly significant only at .10 level) and this seems to be the most reliable conclusion about the pattern in question (i.e., very mild association). Figure 4 is also suggestive of such a conclusion.

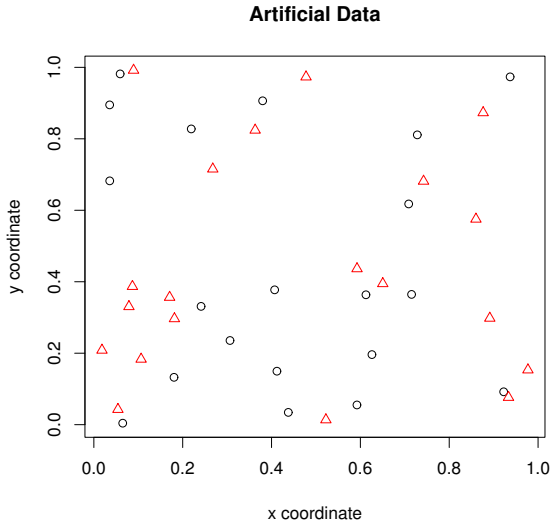

Figure 4 The scatter plots of the locations of 20 X points (circles $\circ$) and 20 Y points (triangles $\triangle$).

Table 8 The NNCT for the artificial data (left) and the corresponding percentages (right).

		NN		
		X	Y	sum
base	X	6	14	20
	Y	13	7	20
	sum	19	21	40

		NN		
		X	Y	
base	X	30 %	70 %	50 %
	Y	65 %	35 %	50 %
		47.5 %	52.5 %	100 %

8. Discussion and Conclusions

In this article we propose the use of exact tests for segregation tests based on nearest neighbor contingency tables (NNCTs). Based on our Monte Carlo simulations, we conclude that the asymptotic approximation for the NNCT-tests is appropriate for large samples. For smaller samples, Monte Carlo randomization versions of the NNCT-tests or the most conservative versions of Fisher's exact test (i.e., the ones based on table-inclusive versions in Section 3) can be used.

For the analysis of two-class spatial patterns, Ripley's K-function and related methods are extensively used in literature. The K-function techniques are based on Monte Carlo simulations

and might indicate different patterns at different distance values. Therefore, for an overall assessment of the patterns between two-classes, we first recommend the use of NNCT tests, and then Ripley's K or L function tests to get the details of the pattern(s) at different scales. Furthermore, K-function methods are most useful as pairwise comparisons. The pair correlation function $g(t)$ and Ripley's classical K- or L-functions and other variants provide information on the pattern at various scales. On the other hand NNCT-tests summarize the pattern in the data set for small scales in one compound summary statistic; more specifically, they provide information on the pattern around the average NN distance between the points.

Table 9 The test statistics (top) and p-values (bottom) for the exact and asymptotic NNCT-tests. θ is for the odds ratio which is the test statistic for variants of Fisher's exact test, C_P is for the test statistic for Pielou's test and Pearson's exact test, and C_D is for Dixon's test. p_P is the p-value for Pielou's test, and other subscripts and superscripts of the p-values are as in Tables 5 and 6.

Test statistics and p-values for the exact tests and Dixon's test							
Data	θ (i.e., odds ratio)				C_P		C_D
	$p_{\text{inc}}^{S,woc}$	$p_{\text{inc}}^{A,wc}$	$p_{\text{inc}}^{I,wc}$	$p_{\text{t,inc}}^{II,woc}$	$p_{\text{t,inc}}^{\chi,wc}$	p_P	p_D
Swamp tree data	5.0052				34.8359		23.7704
(the whole region)	$< .0001$	≈ 1.0	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$
Swamp tree data	7.1934				7.3635		5.2074
(one-eighth of the region)	.0110	.9994	.0137	.0272	.0131	.0067	.0740
Artificial data	.2402				4.9123		2.7585
	.9948	.0406	.0812	.0790	.0919	.0267	.2518

Ripley's classical K- or L-functions can be used for testing (i.e., inference) when the null pattern can be assumed to be CSR independence; that is, when the null pattern assumes first-order homogeneity for each class. When the null pattern is the RL of points from an inhomogeneous Poisson process they are not appropriate (Kulldorff [23]) Diggle's D-function is a modified version of Ripley's K-function (Diggle [13]) and adjusts for any inhomogeneity in the locations of, e.g., cases and controls. Furthermore, there are variants of $K(t)$ that explicitly correct for inhomogeneity (see Baddeley *et al.* [3]). Ripley's K-, Diggle's D- and pair correlation functions are designed to analyze univariate or bivariate spatial interaction at various scales (i.e., inter-point distances). Since the pair correlation functions are derivatives of Ripley's K-function (Stoyan and Stoyan [28]), most of the above discussion holds for them also, except $g(t)$ is reliable only for large scale interaction analysis. Hence NNCT-tests and pair cor-

relation function are not comparable but provide complimentary information about the pattern in question.

Pielou's NNCT-test is not appropriate for testing complete spatial randomness (CSR) independence or random labeling (RL). However, the finite sample nature and conservativeness of Fisher's exact test on NNCTs makes it more appropriate for testing CSR independence or RL compared to Pielou's test. We consider four main variants of the test for one-sided alternatives, and fourteen different variants (nine using Fisher's exact test, five using the exact version of Pearson's test) for the two-sided alternatives. For each variant, we also implement a correction on odds ratio which makes the parameter of testing (i.e., odds ratio) θ accurate under CSR independence or RL. This adjustment also improves the empirical sizes of some of the tests under the null case.

Out of all these variants of exact tests, the most conservative versions with the correction on odds ratio have the best performance in terms of empirical size. Furthermore, the exact tests have higher power estimates under the alternatives compared to Dixon's test. For the right-sided (left-sided) test, the table-inclusive version without (with) finite sample adjustment has the best empirical size performance. For type I two-sided tests, the table-inclusive version with correction on odds ratio has the best performance. On the other hand, for type II and Pearson's exact tests, the best performers are twice-table-inclusive versions without and with correction on odds ratio, respectively. Considering empirical size and power performance, we recommend the exact version of Pearson's test with correction on odds ratio for the two-sided tests, and then the one-sided exact tests can be performed. We recommend the exact tests when sample sizes are small (i.e., < 30); for larger samples, Dixon's test can be safely used.

Acknowledgements

I would like to thank anonymous referees, whose constructive comments and suggestions greatly improved the presentation and flow of this article.

References

[1] Agresti, A. (1992). A survey of exact inference for contingency tables, *Statistical Science*, **7**, 131-153.

[2] Agresti, A. (1996). *An Introduction to Categorical Data Analysis*, Wiley, New York.

[3] Baddeley, A. J., Møller, J., and Waagepetersen, R. (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns, *Statistica Neerlandica*, **54**(3), 329-350.

[4] Barot, S., Gignoux, J., and Menaut, J. C. (1999). Demography of a savanna palm tree: predictions from comprehensive spatial pattern analyses, *Ecology*, **80**, 1987-2005.

[5] Berger, R. L. and Boos, D. D. (1994). P values maximized over a confidence set for the

nuisance parameter, *Journal of the American Statistical Association*, **89**(427), 1012-1016.

[6] Ceyhan, E. (2007). Edge correction for exact tests on nearest neighbor contingency tables for testing spatial segregation. In *Proceedings of the Joint Statistical Meeting, Section on Statistics and the Environment, American Statistical Association*.

[7] Ceyhan, E. (2009). New tests for spatial segregation based on nearest neighbor contingency tables. Accepted for publication in *Scandinavian Journal of Statistics* with *doi:10.1001/j.1467-9469.2009.00667.x*.

[8] Ceyhan, E. (2008). On the use of nearest neighbor contingency tables for testing spatial segregation, *Environmental and Ecological Statistics. doi:10.1007/s10651-008-0104-x*.

[9] Clark, P. J. and Evans, F. C. (1955). On some aspects of spatial pattern in biological populations, *Science*, **121**, 397-398.

[10] Conover, W. J. (1999). *Practical Nonparametric Statistics, 3rd ed,* John Wiley & Sons.

[11] Conradt, L. (1998). Measuring the degree of sexual segregation in group-living animals, *The Journal of Animal Ecology*, **67**(2), 217-226.

[12] Coomes, D. A., Rees, M., and Turnbull, L. (1999). Identifying aggregation and association in fully mapped spatial data, *Ecology*, **80**(2), 554-565.

[13] Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*, Hodder Arnold Publishers, London.

[14] Dixon, P. M. (1994). Testing spatial segregation using a nearest-neighbor contingency table, *Ecology*, **75**(7), 1940-1948.

[15] Dixon, P. M. (2002). Nearest-neighbor contingency table analysis of spatial segregation for several species, *Ecoscience*, **9**(2), 142-151.

[16] Dixon, P. M. (2002). *Nearest neighbor methods*, Encyclopedia of Environmetrics, edited by Abdel H. El-Shaarawi and Walter W. Piegorsch, John Wiley & Sons Ltd., NY, **3**, 1370-1383.

[17] Epstein, L. D. and Fienberg, S. E. (1992). A survey of exact inference for contingency tables: Comment, *Statistical Science*, **7**, 160-163.

[18] Goreaud, F. and Pélissier, R. (2003). Avoiding misinterpretation of biotic interactions with the intertype K_{12}-function: population independence vs. random labelling hypotheses, *Journal of Vegetation Science*, **14**(5), 681-692.

[19] Haase, P. (1995). Spatial pattern analysis in ecology based on Ripley's K-function: Introduction and methods of edge correction, *The Journal of Vegetation Science*, **6**, 575-582.

[20] Hamill, D. M. and Wright, S. J. (1986). Testing the dispersion of juveniles relative to adults: A new analytical method, *Ecology*, **67**(2), 952-957.

[21] 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.

[22] Krebs, C. J. (1972). Ecology: the Experimental Analysis of Distribution and Abundance, Harper and Row, New York, USA.

[23] Kulldorff, M. (2006). Tests for spatial randomness adjusted for an inhomogeneity: A general framework, *Journal of the American Statistical Association*, **101**(475), 1289-1305.

[24] Meagher, T. R. and Burdick, D. S. (1980). The use of nearest neighbor frequency analysis in studies of association, *Ecology*, **61**(5), 1253-1255.

[25] Moran, P. A. P. (1948). The interpretation of statistical maps, *Journal of the Royal Statistical Society, Series B*, **10**, 243-251.

[26] 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.

[27] Pielou, E. C. (1961). Segregation and symmetry in two-species populations as studied by nearest-neighbor relationships, *Journal of Ecology*, **49**(2), 255-269.

[28] Stoyan, D. and Stoyan, H. (1996). Estimating pair correlation functions of planar cluster processes, *Biometrical Journal*, **38**(3), 259-271.

[29] Tocher, K. D. (1950). Extension of the Neyman-Pearson theory of tests to discontinuous variates, *Biometrika*, **37**, 130-144.

[30] Upton, G. J. G. and Fingleton, B. (1995). Spatial Data Analysis by Example, Volume 1: Point Pattern and Interval Data, John Wiley, Chichester.

[31] Waller, L. A. and Gotway, C. A. (2004), Applied Spatial Statistics for Public Health Data, Wiley-Interscience, NJ.

[32] 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.