Statistics
in Medicine

Research Article

Received 5 February 2013, Accepted 2 November 2013 Published online 4 December 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6053

Segregation indices for disease clustering

Elvan Ceyhan*†

Spatial clustering has important implications in various fields. In particular, disease clustering is of major public concern in epidemiology. In this article, we propose the use of two distance-based segregation indices to test the significance of disease clustering among subjects whose locations are from a homogeneous or an inhomogeneous population. We derive the asymptotic distributions of the segregation indices and compare them with other distance-based disease clustering tests in terms of empirical size and power by extensive Monte Carlo simulations. The null pattern we consider is the random labeling (RL) of cases and controls to the given locations. Along this line, we investigate the sensitivity of the size of these tests to the underlying background pattern (e.g., clustered or homogenous) on which the RL is applied, the level of clustering and number of clusters, or to differences in relative abundances of the classes. We demonstrate that differences in relative abundances have the highest influence on the empirical sizes of the tests. We also propose various non-RL patterns as alternatives to the RL pattern and assess the empirical power performances of the tests under these alternatives. We observe that the empirical size of one of the indices is more robust to the differences in relative abundances, and this index performs comparable with the best performers in literature in terms of power. We illustrate the methods on two real-life examples from epidemiology. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: cell-specific tests; Cuzick–Edwards's tests; empirical power; empirical size; nearest neighbor contingency table; overall test; random labeling; spatial clustering

1. Introduction

Recently, spatial clustering has become a topic of extensive study in many fields such as geography, ecology, astronomy, and epidemiology. Many books such as [1] and [2] have discussed the relevant methodology; even special issues of journals are devoted to this topic; see, for example, [3] and [4]. In particular, the significance of disease clustering in human or other populations has received considerable attention ([5, 6] and [7]). Roughly speaking, a *disease cluster* is a region or neighborhood where the number of cases substantially exceeds the expected number of cases at a specific time or for a specific period [8]. There are many tests available for testing the significance of disease clustering. Among them are tests for deviation from homogeneity like the usual Pearson's chi-square statistic for quadrat data or Potthoff–Whittinghill's test [9]. Besag and Newell (1991) grouped clustering methods for detection of disease clustering into two categories as *general* or *focused* [10]. In the former, presence of any cluster over the entire region is of interest, whereas in the latter, presence of a cluster in the vicinity of a given point is investigated. In this article, we are concerned with the general type of clustering. Cuzick–Edwards's k-nearest neighbor (NN) test [11] is an example of a method to test general clustering on the basis of individual point data and has been frequently employed in epidemiology so that it is suggested in the appendix of guidelines for disease clustering [12]. See [13] for a recent review on existing disease clustering methods, their advantages, and disadvantages.

Regional count method is the procedure in which a square grid is overlaid over the region of interest, and the number of events in each quadrat is counted. With this method, both general and focused clustering can be tested. Assuming the points are from a homogeneous Poisson process (HPP), which is the null pattern, the quadrat counts would be distributed as Poisson variates, and their departure from the null case can be tested using an index of dispersion (like ratio of variance to mean), or χ^2 test for heterogeneity of the cell counts. This method has various shortcomings, especially for disease clustering. For example, the quadrats would not be square cells, but administrative units determined

Department of Mathematics, Koç University, Sarıyer, 34450, Istanbul, Turkey
**Correspondence to: Elvan Ceyhan, Department of Mathematics, Koç University, Sarıyer, 34450, Istanbul, Turkey.*
†*E-mail: elceyhan@ku.edu.tr*

by geographical limitations or human intervention. This problem can somewhat easily be overcome by extending the quadrat method to other shapes or administrative units by simply employing the chi-square goodness-of-fit test using the observed and expected numbers in each region. Other main problems are the arbitrariness in the choice of grid size and obtaining correct expected quadrat counts on a sufficiently fine grid structure [11].

Statistical methodology based on NN (or distance-based) methods include at least six different groups [14]. Each of these methods assumes as a premise that similarity or dissimilarity between a point and its NN provides useful information for statistical inference. The most straightforward dissimilarity measure is the distance between a point and its NN, whereas other methods could be based on classifying the types of points and their NNs. For example, in literature, there are spatial clustering tests based on nearest neighbor contingency tables (NNCTs) due to [15] and [16] in a two-class setting, and due to [17] in a multi-class setting. Ceyhan [18] also proposed various new segregation tests on the basis of NNCTs. These tests comprise an overall test, a compound measure of deviation from the null pattern, and cell-specific tests for pairwise comparisons after a significant overall test. Pielou [15] introduced also a ‘coefficient of segregation’ in a two-class setting, and Dixon [17] proposed ‘segregation indices’ by in a multi-class setting. However, these indices were merely introduced in passing and not studied in detail (e.g., their asymptotic distributions were not derived), nor they were applied for inferential purposes. In this article, we study their distributional properties and also propose their use for testing spatial clustering (especially of cases compared with controls).

Several indices measure spatial autocorrelation in given data, which could suggest clustering of a disease, for example, Moran’s I statistic [19] and Geary’s c statistic [20]. Furthermore, there are methods, which provide a general clustering measure for the entire study area, such as Whittemore’s statistic [21]. However, Tango [22] had showed this statistic to be inadequate, who also proposed a corrected version of it. As the general clustering methods fail to identify localized clusters, the so-called scan statistics are developed. In these methods, a rectangular or circular window scans the region to detect any anomaly in disease occurrence or intensity. Examples of scan methods are Openshaw’s geographical analysis machine [23], Besag and Newell’s method to detect clusters of size k, which comprise regions containing exactly k observed cases [10], and Kulldorff & Nagarwalla’s scan statistic [24]. In literature, despite the lack of a comprehensive comparison of many available geographical disease clustering tests, Kulldorff *et al.* [25] performed an empirical comparison using spatial scan statistic, the maximized excess events test, and the nonparametric M statistic. They showed all tests to have good power for detection of disease clusters and the first having good performance in locating disease hot spots.

In literature, clustering not only in space but also in time is of interest, especially with applications in climatology or ecology [26]. This type of clustering, called *spatio-temporal clustering*, is also of great import in disease clustering research. Tango suggested an index for disease clustering in time [27], and this index is assessed in detail for performance to detect disease clustering in time and space by [28]. Several other indices were also proposed in literature to capture spatial patterns and their evolution in time. See [29] for an example in marine biology, which measures spatial patterns of fish populations and [30] for an example in landscape ecology, where a new contagion index was proposed that also corrects for an existing index.

In this article, we propose the use of Pielou’s coefficient of segregation and Dixon’s segregation indices in detecting disease clustering against the RL of cases and controls to a set of given spatial locations. Dixon’s segregation indices are not bounded for all possible realizations of NNCTs; hence, we also suggest corrected versions of Dixon’s segregation indices, which are bounded for all cell counts (zero or positive) in an NNCT. We derive their asymptotic distributions (more specifically asymptotic normality) and compare these tests with various existing tests, namely, Cuzick–Edwards’s k-NN and combined tests, Dixon’s and type III cell-specific and overall tests in terms of empirical size and power. For the RL, we investigate the effect of the clustering of the background points (on which RL is performed), including the level of clustering and number of clusters, and the effect of differences in relative abundances on the empirical sizes of the tests. We also propose various non-RL patterns as alternatives and investigate the power performance of the tests under these alternatives via extensive Monte Carlo simulations. To the best of our knowledge, we investigated for the first time in literature the influence of the background pattern on the size performance, and we newly introduced the non-RL patterns used in this article.

We present the null and alternative patterns and construction of NNCTs in Section 2. We provide the two segregation indices for spatial and disease clustering in Section 3, where their asymptotic normality are derived. We discussed other NN-based spatial clustering tests that are used for comparative purposes in Section 4. We provide an extensive empirical size comparison of the tests under RL of points from various patterns of complete spatial randomness (CSR) or clustering in Section 5, propose four types of non-RL patterns as alternatives and provide empirical power comparison of the tests under these alternatives in Section 6, and illustrate the methodology on two real life data sets from epidemiology in Section 7. We provide discussion and conclusions in Section 8.

2. Preliminaries

2.1. *Null and alternative patterns*

In a case-control setting, we consider the null pattern

$$H_o : RL$$

which is the pattern where the class labels (i.e., case and control labels) are randomly assigned to a given set of locations or points. In the two-class setting, deviations from the null hypothesis are toward two major directions, namely, *segregation* and *association*. *Segregation* is the pattern in which NN of an individual is more likely than expected to be of the same class as the individual than to be from a different class. That is, the probability that this individual having a same-class NN is higher than the relative frequency of this class (see, e.g., [15]). On the other hand, *association* is the pattern in which NN of an individual is more likely to be from another class than expected compared with being of the same class as the individual. That is, the probability that this individual having a NN from another class is higher than the relative frequency of the other class.

In a case-control setting, segregation of cases from controls would be equivalent to clustering of cases relative to the controls. In other words, segregation of cases would imply a larger level of clustering of cases compared with the level of clustering of the healthy controls in the society. Furthermore, if, for some reason, controls are segregated, then this would also imply an (indirect) clustering of cases, but in this case, the underlying dynamics behind the disease clustering would be different. The association of the cases and controls would mean significant lack of disease clustering; moreover, it would mean clustering of points from both classes (i.e., attraction of controls by cases or vice versa). This may not be practical either, hence is not pursued in detail in the rest of the article. However, the association pattern could still be relevant to disease clustering in epidemiology in other settings. For example, one class could be the ‘sources’ of a contaminant or some other pollutant or disease-causing agent, and the other class could be the ‘cases’. The accumulation of cases around the sources more often than expected would mean clustering of a disease around these sources, which is a form of association between the classes. But we will not pursue this type of association in this article either.

2.2. *Construction of nearest neighbor contingency tables*

The segregation indices and most of the tests we consider for comparative purposes in this article are in some way related to NNCTs. We provide a brief description of NNCTs in the succeeding text. In a sample of size n, there are n NN pairs, and each NN pair consists of the point labeled as ‘base’ point and its ‘NN’ point. According to the labels of the base and NN points, NN pairs can be classified into various categories, and NNCTs are constructed using these categories. For m classes, we will have a $m \times m$ NNCT whose rows represent class labels of base points and columns represent class labels of the corresponding NN points. In the NNCT, the count in cell (or entry) (i, j) is N_{ij}, which is the number of times the NN of a (base) class i point being from class j. See also Table I (left) where C_j is the sum of column j; that is, number of times class j points serve as NNs for $j \in \{1, 2, \ldots, m\}$, and n_i is the sum of row i; that is, number of times class i points serve as base class or size of class i for $i = 1, 2, \ldots, m$. In what follows, we adopt the convention that lower case letters represent fixed quantities, whereas upper case letters represent random variables. Notice that in a NNCT analysis, row sums are assumed to be fixed (i.e., class sizes are given), whereas column sums are random variables depending on the NN relationship between the classes.

In a case-control setting, we have two classes (i.e., $m = 2$), and we reserve class label 1 for cases and class label 2 for controls. Hence, the case-control setting yields a 2×2 NNCT (Table I (right)).

Table I. The nearest neighbor contingency tables for m classes (left) and for two classes in a case-control setting (right).

		NN			
		class 1	$\ldots$	class m	total
base	class 1	N_{11}	$\ldots$	N_{1m}	n_1
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	class m	N_{m1}	$\ldots$	N_{mm}	n_m
	total	C_1	$\ldots$	C_m	n

		NN		
		case	control	total
base	case	N_{11}	N_{12}	n_1
	control	N_{21}	N_{22}	n_2
	total	C_1	C_2	n

NN, nearest neighbor.

3. Segregation indices for spatial and disease clustering

3.1. Pielou's coefficient of segregation

In a two-class setting (i.e., for $m = 2$), we define Pielou's coefficient of segregation as

$$S_P = 1 - \frac{N_{12} + N_{21}}{\mathbf{E}[N_{12}] + \mathbf{E}[N_{21}]} \tag{1}$$

where $\mathbf{E}[N_{ij}]$ is the expected value of N_{ij} [15]. Notice that the numerator in the second part of S_P is

$$N_{12} + N_{21} = \sum_{i=1}^{n} \mathbf{I}(\text{point } i \text{ is from class 1 with an NN from class 2}$$
$$\text{or point } i \text{ is from class 2 with an NN from class 1})$$

where $\mathbf{I}()$ stands for the indicator function. In general, a $m \times m$ contingency table may result from two multinomial frameworks: row-wise and overall multinomial frameworks.

3.1.1. The row-wise multinomial framework. In this framework, the rows of a contingency table result from independent multinomial distributions. In particular, in the two-class case, each row has a binomial distribution independent of the other rows (so this framework is also referred to as the *binomial framework*).

Let π_{ij} be the probability of a point from class j serving as NN to a point from class i for $i, j \in \{1, 2\}$. In this framework, we assume that $N_i = n_i$ are given and $N_{ij} \sim \text{BIN}(n_i, \pi_{ij})$, the binomial distribution with n_i independent trials and probability of success being π_{ij}. Hence, in the two-class case, we assume (N_{11}, N_{12}) and (N_{21}, N_{22}) to be independent and so are the individual trials (which are base-NN pairs). Hence this framework would be appropriate for an NNCT analysis, provided that we have an independent set of (base-NN) pairs; that is, each (base-NN) pair is independent of other pairs. In what follows, when we say data are from sparse sampling, we also assume that (base-NN) pairs constitute an (almost) independent sample. Thus, with sparse sampling under the null hypothesis of 'random assignment of case and control labels to any given point being proportional to the class sizes', we would have $N_{ij} \sim \text{BIN}(n_i, \nu_j)$, for $i, j = 1, 2$ where ν_j is the population proportion of class j points. Thus

$$S_P = 1 - \frac{N_{12} + N_{21}}{\mathbf{E}[N_{12}] + \mathbf{E}[N_{21}]} = 1 - \frac{N_{12} + N_{21}}{n_1\nu_2 + n_2\nu_1}$$

Clearly $\mathbf{E}[S_P] = 0$ and under H_o, $\mathbf{Var}[N_{12}] = n_1\nu_1\nu_2$ and $\mathbf{Var}[N_{21}] = n_2\nu_1\nu_2$, and N_{12} and N_{21} are independent. Hence, $\mathbf{Var}[S_P] = \frac{n\nu_1\nu_2}{(n_1\nu_2+n_2\nu_1)^2}$. However, in practice, ν_i would not be known and, hence, need to be estimated. Given a sample of size n_i from class i, we estimate ν_i as $\widehat{\nu}_i = n_i/n$ for $i = 1, 2$. Then for large n_i, $\mathbf{E}[N_{12}] \approx n_1 n_2/n$ and $\mathbf{E}[N_{21}] \approx n_2 n_1/n$, so $S_P \approx 1 - \frac{N_{12}+N_{21}}{(2n_1n_2/n)}$. Furthermore, we have $\mathbf{Var}[S_P] \approx \frac{n}{4n_1n_2}$. Then for large n_i, $i = 1, 2$, under sparse sampling, $S_P/\sqrt{\mathbf{Var}[S_P]}$ approximately has $N(0, 1)$ distribution where $N(\mu, \sigma)$ stands for normal distribution with mean μ and standard deviation σ.

3.1.2. The overall multinomial framework. In this setting, we assume the cell counts (or the entries) to arise from multinomial trials. That is, in the two-class case,

$$\mathbf{N} = (N_{11}, N_{12}, N_{21}, N_{22}) \sim \mathscr{M}\left(n, \nu_1\kappa_1, \nu_1\,\kappa_2, \nu_2\,\kappa_1, \nu_1\,\kappa_2\right)$$

where $\nu_1 + \nu_2 = 1$ and $\kappa_1 + \kappa_2 = 1$. As in the row-wise multinomial framework, this framework would be appropriate for an NNCT-analysis, provided we have sparsely sampled data. In this framework, $N_{ij} \sim \text{BIN}\left(n, \nu_i\kappa_j\right)$ for $i, j = 1, 2$. Using $\kappa_2 = 1 - \kappa_1$ and $\nu_2 = 1 - \nu_1$, we have $\mathbf{E}[N_{12}] = n\nu_1(1-\kappa_1)$ and $\mathbf{E}[N_{21}] = n(1-\nu_1)\kappa_1$, which yields

$$S_P = 1 - \frac{N_{12} + N_{21}}{n\left(\nu_1 + \kappa_1\right) - 2n\nu_1\kappa_1}$$

Furthermore, $\mathbf{Var}[N_{12}] = n\nu_1(1-\kappa_1)(1-\nu_1+\nu_1\kappa_1)$, $\mathbf{Var}[N_{21}] = n(1-\nu_1)\kappa_1(1-\kappa_1+\nu_1\kappa_1)$, and $\mathbf{Cov}\left[N_{12}, N_{21}\right] = -n\nu_1\kappa_2\nu_2\kappa_1 = -n\nu_1(1-\kappa_1)(1-\nu_1)\kappa_1$. So $\mathbf{Var}\left[N_{12}+N_{21}\right] = n\left(1-\nu_1-\kappa_1+2\nu_1\kappa_1\right)\left(\nu_1+\kappa_1-2\nu_1\kappa_1\right)$, which implies $\mathbf{Var}[S_P] = \frac{1-\nu_1-\kappa_1+2\nu_1\kappa_1}{n(\nu_1+\kappa_1-2\nu_1\kappa_1)}$. Under H_o, we have $\nu_1 = \kappa_1$, and so

$$\mathbf{Var}[S_P] = \frac{1}{2n}\left(\frac{\nu_2}{\nu_1} + \frac{\nu_1}{\nu_2}\right)$$

Hence, for large n_i, $i = 1, 2$, $\mathbf{Var}[S_P] \approx \frac{1}{2n}\left(\frac{n_2}{n_1} + \frac{n_1}{n_2}\right)$ and under sparse sampling $S_P/\sqrt{\mathbf{Var}[S_P]}$ approximately has $N(0,1)$ distribution.

Remark
Both of the aforementioned multinomial frameworks require that we have an independent sample of n (base-NN) pairs from the appropriate multinomial distribution. However, for completely mapped data, this assumption does not hold because of the inherent spatial dependence. For example, a base point would be more likely to be an NN of its own NN compared with being an NN of an arbitrarily selected point. However, if we have data obtained by sparse sampling, this dependence would be nonexistent or negligible, then this framework would be appropriate for the corresponding NNCT. But, in a case-control setting, sparse sampling may not be a feasible procedure, especially when the disease in question is rare. Hence, sparse sampling in general is not advisable for detection of disease clustering. On the other hand, these frameworks would work when there is a substantial amount of data from both classes in the region of interest, and sparse sampling would be a feasible practice to capture the actual interaction between the classes.

3.1.3. Pielou's coefficient of segregation under random labeling. Under RL of n_1 cases and n_2 controls to $n = n_1 + n_2$ given locations, we have $\mathbf{E}[N_{12}] = \mathbf{E}[N_{21}] = \frac{n_1 n_2}{n-1}$. So under H_o,

$$S_P = 1 - \frac{N_{12} + N_{21}}{n_1 n_2/(n-1)}$$

and $\mathbf{E}[S_P] = 0$. Furthermore,

$$\mathbf{Var}[N_{ij}] = n\,p_{ij} + Q\,p_{iij} + \left(n^2 - 3n - Q + R\right)p_{iijj} - (np_{ij})^2$$

and

$$\mathbf{Cov}\left[N_{ij}, N_{ji}\right] = Rp_{ij} + (n-R)\left(p_{iij} + p_{ijj}\right) + \left(n^2 - 3n - Q + R\right)p_{iijj} - n^2 p_{ij}\,p_{ji}$$

where $p_{ij} = \frac{n_i n_j}{n(n-1)}$, $p_{iij} = \frac{n_i(n_i-1)n_j}{n(n-1)(n-2)}$, $p_{ijj} = \frac{n_i n_j(n_j-1)}{n(n-1)(n-2)}$ and $p_{iijj} = \frac{n_i(n_i-1)n_j(n_j-1)}{n(n-1)(n-2)}$, for $(i,j) = (1,2)$ and $(i,j) = (2,1)$, R is twice the number of reflexive pairs and Q is the number of points with shared NNs, which occurs when two or more points share an NN. Then $Q = 2\left(Q_2 + 3\,Q_3 + 6\,Q_4 + 10\,Q_5 + 15\,Q_6\right)$, where Q_k is the number of points that serve as a NN to other points k times. Then $\mathbf{Var}\left[N_{12}+N_{21}\right] = \mathbf{Var}[N_{12}] + \mathbf{Var}[N_{21}] + 2\,\mathbf{Cov}\left[N_{12}, N_{21}\right]$, and for large n_i, $Z_P = S_P/\sqrt{\mathbf{Var}[S_P]}$ has approximately $N(0,1)$ distribution.

Suppose we know the population proportion, ν_i, for class $i,\ i=1,2$. Then, for large n_i, we would have $\mathbf{E}[N_{12}]=\mathbf{E}[N_{21}]\approx n\nu_1\nu_2$ and $S_P\approx 1-\frac{N_{12}+N_{21}}{n\nu_1\nu_2}$. Furthermore, $p_{ij}=\nu_i\nu_j,\ p_{iij}=\nu_i^2\nu_j,\ p_{ijj}=\nu_i\nu_j^2$, and $p_{iijj}=\nu_i^2\nu_j^2$; hence,

$$\mathbf{Var}[N_{ij}]\approx n\,\nu_i\nu_j+Q\,\nu_i^2\nu_j+(-3n-Q+R)\nu_i^2\nu_j^2 \tag{2}$$

and

$$\mathbf{Cov}\left[N_{ij},N_{ji}\right]\approx R\nu_i\nu_j+(n-R)\left(\nu_i^2\nu_j+\nu_i\nu_j^2\right)+(-3n-Q+R)\nu_i^2\nu_j^2$$

for $(i,j)=(1,2)$ and $(i,j)=(2,1)$.

3.2. Dixon's segregation indices

In a multi-class setting, Dixon [17] proposed the following indices, which are similar to the log odds-ratios in an NNCT:

$$S_{ij}^D=\begin{cases}\log\left(\frac{N_{ii}/(n_i-N_{ii})}{(n_i-1)/(n-n_i)}\right) & \text{if } i=j\\ \log\left(\frac{N_{ij}/(n_i-N_{ij})}{n_j/(n-n_j-1)}\right) & \text{if } i\neq j\end{cases} \tag{3}$$

Let

$$Z_{ij}^S=\begin{cases}\dfrac{S_{ii}^D}{\sqrt{\mathbf{Var}[N_{ii}]}\left(\frac{(n-1)^2}{n_i(n-n_i)(n_i-1)}\right)} & \text{if } i=j\\ \dfrac{S_{ij}^D}{\sqrt{\mathbf{Var}[N_{ij}]}\left(\frac{(n-1)^2}{n_in_j(n-n_j-1)}\right)} & \text{if } i\neq j\end{cases} \tag{4}$$

Under RL, as n_1 and n_2 go to infinity, $\left(N_{ij}-\mathbf{E}[N_{ij}]\right)/\sqrt{\mathbf{Var}[N_{ij}]}$ converges in law to $N(0,1)$ distribution. Then by an appropriate application of central limit theorem and the delta method, Z_{ij}^S approximately has $N(0,1)$ distribution for large n_i,n_j for all i,j.

For a derivation of this asymptotic result and the asymptotic approximations of Dixon's segregation indices when the population proportion, ν_i, of class $i,\ i=1,2$ are known; see the technical report [31].

3.2.1. A correction for Dixon's segregation indices.

Dixon's segregation indices may be unbounded in either direction depending on the cell counts in the NNCT. Let $0<n_i<n$ for all i. Then if $N_{ii}=0$, we obtain $S_{ii}^D=-\infty$ provided $n_i>1$; and if $N_{ij}=0$, we obtain $S_{ij}^D=-\infty$ provided $n_j<n-1$. Also, if $N_{ii}=n_i$, we obtain $S_{ii}^D=\infty$; and if $N_{ij}=n_i$, we obtain $S_{ij}^D=\infty$ provided $n_j<n-1$. To make the segregation indices bounded for all possible cell counts, we suggest the following corrected versions:

$$S_{ij}^{D,c}=\begin{cases}\log\left(\frac{(N_{ii}+1)/(n_i-N_{ii}+1)}{(n(n_i-1)+(n-1))/(n_i(n-n_i)+(n-1))}\right) & \text{if } i=j\\ \log\left(\frac{(N_{ij}+1)/(n_i-N_{ij}+1)}{\left(n_in_j+n-1\right)/\left(n_i(n-n_j-1)+(n-1)\right)}\right) & \text{if } i\neq j\end{cases} \tag{5}$$

where denominators are chosen in this way so that they have simpler asymptotic approximations.

Let

$$Z_{ij}^{S,c}=\begin{cases}\dfrac{S_{ii}^{D,c}}{\sqrt{\mathbf{Var}[N_{ii}]}\left(\frac{(n_i+2)(n-1)^2}{(n_i(n-n_i)+(n-1))(n_i(n-n_i)+(n-1))}\right)} & \text{if } i=j\\ \dfrac{S_{ij}^{D,c}}{\sqrt{\mathbf{Var}[N_{ij}]}\left(\frac{(n_i+2)(n-1)^2}{\left(n_in_j+n-1\right)\left(n_i(n-n_j+1)+(n-1)\right)}\right)} & \text{if } i\neq j\end{cases} \tag{6}$$

Again, by central limit theorem and the delta method, $Z_{ij}^{S,c}$ approximately has $N(0,1)$ distribution for large n_i,n_j for all i,j.

For the derivation of asymptotic distribution of these corrected versions, see the technical report [31].

4. Other nearest neighbor tests for spatial clustering

Although there are many tests available for spatial clustering of points from one class or multiple classes in the literature ([2] and [32]), one-class tests are not comparable with the segregation indices nor very useful in disease clustering. Some of the tests like Moran's I and Whittemore's tests are shown to perform poorly in detection of some kind of clustering [33], and most of the tests require Monte Carlo simulation or randomization methods to attach significance to their results. Hence, we only consider cell-specific and overall NNCT tests due to [16] and [18] and Cuzick–Edwards's k-NN tests and their combined versions [11], and compare the segregation indices with these tests in an extensive Monte Carlo simulation study in terms of size and power performance.

4.1. *Cell-specific and overall segregation tests based on nearest neighbor contingency tables*

Dixon's cell-specific and overall tests [16] and type III cell-specific and overall tests [18] are based on NNCTs. Ceyhan discussed these tests in detail in [34]; here, we only provide a brief description for completeness. For cell (i,j), Dixon [16] and Ceyhan [18] suggested

$$Z^D_{ij} = \frac{N_{ij} - \mathbf{E}[N_{ij}]}{\sqrt{\mathbf{Var}[N_{ij}]}} \quad \text{and} \quad Z^{III}_{ij} = \frac{T^{III}_{ij}}{\sqrt{\mathbf{Var}\left[T^{III}_{ij}\right]}} \tag{7}$$

as the cell-specific tests, respectively. Under RL, the expected cell counts are $\mathbf{E}[N_{ij}] = n_i(n_i-1)/(n-1)\mathbf{I}(i=j) + n_i n_j/(n-1)\mathbf{I}(i \neq j)$, and Ceyhan gave the variance $\mathbf{Var}[N_{ij}]$ in [18]. Furthermore, $T^{III}_{ij} = \left(N_{ii} - \frac{(n_i-1)}{(n-1)}C_i\right)\mathbf{I}(i=j) + \left(N_{ij} - \frac{n_i}{(n-1)}C_j\right)\mathbf{I}(i \neq j)$. Ceyhan presented the explicit forms of expectation and variance of T^{III}_{ij} in [34]. In the multi-class case with m classes, combining the m^2 cell-specific tests, Dixon [17] and [18] suggested the following quadratic forms:

$$C_D = (\mathbf{N} - \mathbf{E}[\mathbf{N}])' \, \Sigma^-_D \, (\mathbf{N} - \mathbf{E}[\mathbf{N}]) \quad \text{and} \quad C_{III} = \left(\mathbf{T^{III}}\right)' \Sigma^-_{III} \left(\mathbf{T^{III}}\right) \tag{8}$$

as overall tests, respectively. Here, $\mathbf{N}$ is the $m^2 \times 1$ vector of m rows of the NNCT concatenated row-wise, $\mathbf{E}[\mathbf{N}]$ is the vector of $\mathbf{E}[N_{ij}]$, Σ_D is the $m^2 \times m^2$ variance-covariance matrix for the cell count vector $\mathbf{N}$ with diagonal entries being equal to $\mathbf{Var}[N_{ij}]$ and off-diagonal entries being $\mathbf{Cov}\left[N_{ij}, N_{kl}\right]$ for $(i,j) \neq (k,l)$. Dixon provided the explicit forms of the variance and covariance terms in [17]. Also, Σ^-_D is a generalized inverse of Σ_D [35], and $'$ stands for the transpose of a vector or matrix. Similarly, $\mathbf{T^{III}}$ is the vector of m^2 T^{III}_{ij} values, that is,

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

and $\mathbf{E}\left[\mathbf{T^{III}}\right]$ is the vector of $\mathbf{E}\left[T^{III}_{ij}\right]$ values. Note that $\mathbf{E}\left[\mathbf{T^{III}}\right] = \mathbf{0}$ where $\mathbf{0}$ is the vector of m^2 zeros and Σ_{III} is the $m^2 \times m^2$ variance-covariance matrix of $\mathbf{T^{III}}$. Under RL, Ceyhan provided the explicit forms of the variance-covariance matrix in [34]. Then under RL, C_D approximately has a $\chi^2_{m(m-1)}$ distribution, and C_{III} approximately has a $\chi^2_{(m-1)^2}$ distribution for large n_i.

4.2. *Cuzick–Edwards's k nearest neighbor and combined tests*

For disease clustering, Cuzick and Edwards [11] suggested a k NN test on the basis of number of cases among k NNs of the case points. Let z_i be the i^{th} point and d^k_i be the number cases among k NNs of z_i. Then Cuzick–Edwards's k NN test is $T_k = \sum_{i=1}^n \delta_i d^k_i$, where

$$\delta_i = \begin{cases} 1 & \text{if } z_i \text{ is a case} \\ 0 & \text{if } z_i \text{ is a control} \end{cases} \tag{9}$$

Because the correct choice of k is not known in practice, [11] also suggested combining various T_k tests. Let $S = \{k_1, k_2, \ldots, k_m\}$ be a set of indices for k, and assume T_k with $k \in S$ being a mixture of shifts all in the same direction under an alternative. Assuming further that T_k has multivariate normal distribution, the combined test statistic is given by

$$T_S = \mathbf{1}'\Sigma^{-1/2}\mathbf{T} \tag{10}$$

where $\mathbf{T} = \left(T_{k_1}, T_{k_2}, \ldots, T_{k_m}\right)'$ (i.e., T_S is the test obtained by combining T_k tests whose indices are in S), $\mathbf{1}' = (1, 1, \ldots, 1)$, $\Sigma = \mathbf{Cov}[\mathbf{T}]$ is the variance-covariance matrix of $\mathbf{T}$. Under RL of n_1 cases and n_2 controls to the given locations in the study region, T_k approximately has N $(\mathbf{E}[T_k], \mathbf{Var}[T_k]/n_1)$ distribution for large n_1; similarly, T_S approximately has N $(\mathbf{E}[T_S], \mathbf{Var}[T_S])$ distribution for large n_1. Cuzick and Edwards provided the expected values $\mathbf{E}[T_k]$ and $\mathbf{E}[T_S]$ and variances $\mathbf{Var}[T_k]$ and $\mathbf{Var}[T_S]$ in [11].

Notice that T_1 is identical to the count for cell $(1, 1)$ in the NNCT of Table I (right). Hence, the corresponding tests $(T_1 - \mathbf{E}[T_1]) / \sqrt{\mathbf{Var}[T_1]}$ and Z_{11}^D are identical. Hence, we only consider T_2 and T_S with $S = \{1, 2\}$ for Cuzick–Edwards's tests in our comparisons.

Remark

Note that under H_o, expected values of S_P, S_{ii}^D, Z_{ii}^D, $T_k - \mathbf{E}[T_k]$, and $T_S - \mathbf{E}[T_S]$ are all zero. They tend to be positive under segregation and negative under association. On the other hand, under segregation, the diagonal cell counts, N_{ii}, would be larger, whereas under association, the off-diagonal cell counts, N_{ij}, with $i \neq j$, would be larger than expected. Hence, S_{ij}^D and Z_{ij}^D for $i \neq j$ tend to be negative under segregation and positive under association. Hence, all these tests can be employed to test spatial clustering in either direction against H_o in a two-class setting. In a case-control setting, segregation of cases from the controls would be our primary interest.

Remark

With $m = 2$ classes (or in a case-control setting), S_P, S_{ii}^D, Z_{ii}^D, C_D, T_1, and C_{III} can detect the spatial interaction at small scales (at around the average NN distance), whereas T_k with $k > 1$ can detect at larger scales (at around k-th NN distance), and so can T_S with S having indices other than 1 (at around ℓ-th NN distance with $\ell = \min k_i$ to $\ell = \max k_i$ for $k_i \in S$). Hence, S_P, S_{ii}^D, Z_{ii}^D, C_D, T_1, and C_{III} can be used to test the same type of interaction at the same smaller scales, whereas Cuzick–Edwards's tests, T_2 and T_S, can be used to do the same at larger scales.

5. Empirical size analysis of the tests

Let $\mathcal{Z}_n = \{Z_1, Z_2, \ldots, Z_n\}$ be the given set of locations for n points (called the *background pattern*). We consider RL of cases and controls to points in $\mathcal{Z}_n$ generated from various homogeneous or clustered patterns. The particular realization of the background pattern might influence the distribution of the tests. For example, although the expected values of Dixon's and type III cell-specific tests depend only on the class sizes, the corresponding variances and covariances depend on Q and R, which depend on the relative allocation of the points in the background pattern. To remove the effect of one particular realization of the Z_i points on the tests, we consider 100 different realizations of $\mathcal{Z}_n$ on which RL will be applied. For each background realization, we label n_1 of the points as class X (for cases) and the remaining $n_2 = n - n_1$ points as class Y (for controls).

Types of the background patterns:

Case 1: We generate $\mathcal{Z}_n$ points independently uniformly in the unit square $(0, 1) \times (0, 1)$, that is, $Z_i \overset{iid}{\sim} \mathcal{U}((0, 1) \times (0, 1))$ for $i = 1, 2, \ldots, n$. We consider (i) $n_1 = n_2 = 10, 20, \ldots, 100$ to determine the effect of increasing but equal sample sizes; (ii) $n_1 = 30$ and $n_2 = 30, 40, \ldots, 120$ to determine the effect of the differences in the sample sizes with number of cases fixed and number of controls increasing; and (iii) $n_2 = 30$ and $n_1 = 30, 40, \ldots, 120$ to determine the effect of the differences in the sample sizes with number of controls fixed and number of cases increasing. We perform the aforementioned RL scheme 1000 times for each (n_1, n_2) combination at each background realization.

Case 2: We generate $Z_i \overset{iid}{\sim} \mathcal{U}\left(S_\delta^I\right)$ for $i = 1, 2, \ldots, n$ where $S_\delta^I = ((0, 1) \times (0, 1)) \cup ((\delta, 1 + \delta) \times (\delta, 1 + \delta))$. We consider $\delta = 0.2, 0.4., \ldots, 2.0$, so that as δ increases, the level of clustering of background points increases. We perform the aforementioned RL 1000 times for each δ at each background realization with $n_1 = n_2 = 100$.

Case 3: We generate $Z_i \overset{iid}{\sim} \mathcal{U}\left(S_\delta^{II}\right)$ for $i = 1, 2, \ldots, n$ where $S_\delta^{II} = ((0, 1) \times (0, 1)) \cup ((1 + \delta, 2 + \delta) \times (0, 1))$. We consider $\delta = 0.0, 0.2, 0.4., \ldots, 1.4$, so that as δ increases, the level of clustering of background patterns increases. We perform the aforementioned RL 1000 times for each δ at each background realization with $n_1 = n_2 = 100$.

Case 4: We generate $Z_i \overset{iid}{\sim} \mathcal{U}(S_{\delta,k})$ for $i = 1, 2, \ldots, n$ where $S_{\delta,k} = ((0,1) \times (0,1)) \cup ((1+\delta, 2+\delta) \times (0,1)) \ldots \cup ((k-1)(1+\delta), k + (k-1)\delta) \times (0,1))$, which yields k squares along the x-axis for the support of Z_i, with successive squares being δ units apart. We consider $\delta = 0.5$, so that each square is clearly separated, and $k = 1, 2, \ldots, 10$ so that the sensitivity of the empirical sizes of the tests to the number distinct clusters could be assessed. We perform the aforementioned RL 1000 times for each k at each background realization with $n_1 = n_2 = 100$.

Case 5: In this case, we generate Z_i points from Matérn's cluster process in the unit square, denoted MatClust(κ, r, μ) [36]. First, we generate 'parent' points from a Poisson process with intensity κ, and then each parent is replaced by N points independently uniformly generated inside the circle centered at the parent point with radius r, where $N \sim \text{Poisson}(\mu)$. For each background realization, we generate one realization of $\mathcal{Z}_n$ from MatClust(κ, r, μ), and let n be the number of points in this realization. Then, we label $n_1 = \lfloor n/2 \rfloor$ of these points as cases, and $n_2 = n - n_1$ as controls, where $\lfloor x \rfloor$ stands for the floor of x. Here, we take $\kappa = 1, 2, \ldots, 10$, $\mu = \lfloor 200/\kappa \rfloor$, and $r = 0.1$ in our simulations. That is, we take $(\kappa, \mu) \in \{(1, 200), (2, 100), (3, 66) \ldots, (10, 20)\}$, so that on the average, we would have about 200 Z points of which 100 are X and 100 are Y points.

We plot sample realizations from these background cases in Figure 1. At each Monte Carlo replication of RL in each of the aforementioned cases, we compute the following test statistics: Pielou's coefficient of segregation, S_P, Dixon's segregation indices, S_{ij}^D, for $i, j = 1, 2$, and the corrected versions, $S_{ij}^{D,c}$, for $i, j = 1, 2$, Dixon's cell-specific tests, Z_{ij}^D, for $i, j = 1, 2$ type III cell-specific tests, Z_{ij}^{III}, for $i, j = 1, 2$, Dixon's overall test, C_D, type III overall test, C_{III}, Cuzick–Edwards's k NN tests, T_k, for $k = 1, 2$, and combined test, T_S, for $S = \{1, 2\}$ (which is denoted as $T_{1,2}$ in short). However, the case-control setting corresponds to a two-class case. Hence, in our further analysis, we only consider and present S_{ii}^D for $i = 1, 2$ among Dixon's segregation indices, because $S_{11}^D = -S_{12}^D$ and $S_{22}^D = -S_{21}^D$; Z_{ii}^D for $i = 1, 2$ among Dixon's cell-specific tests, because $Z_{11}^D = -Z_{12}^D$ and $Z_{22}^D = -Z_{21}^D$; Z_{ii}^{III} for $i = 1, 2$ among type III cell-specific tests, because $Z_{11}^{III} = -Z_{21}^{III}$ and $Z_{22}^{III} = -Z_{12}^{III}$. Furthermore, among Cuzick–Edwards's k NN tests, we only consider and present T_2, and combined test for $S = \{1, 2\}$,

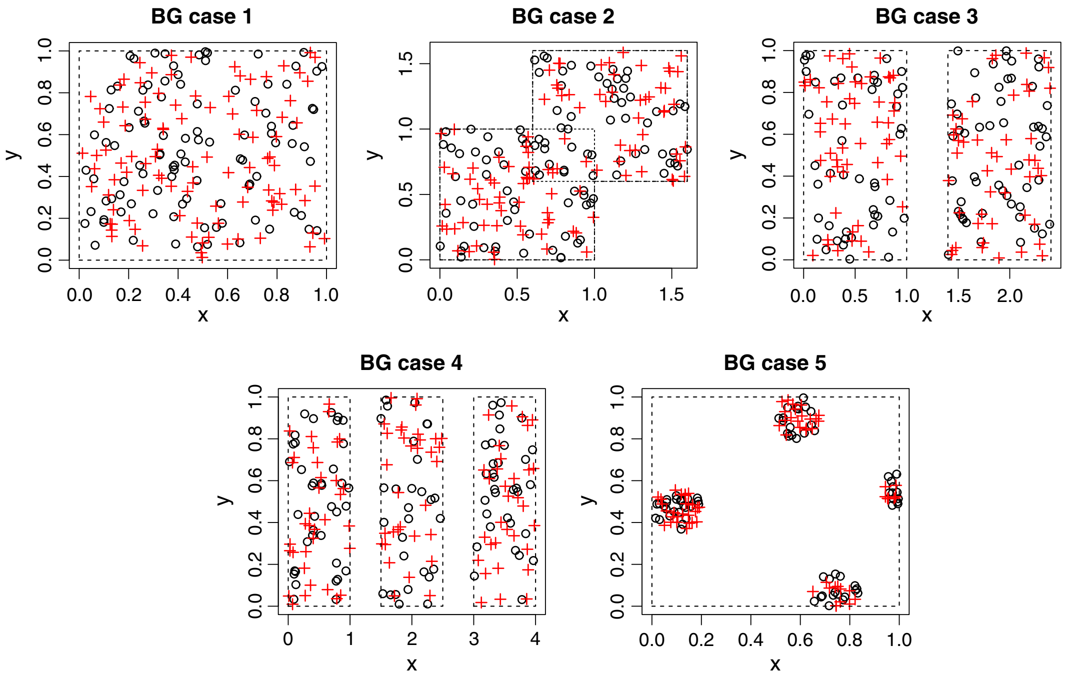


Figure 1. Sample plots of the realizations of the background (BG) pattern cases 1-5 with $n = 200$. The random labeling of $n_1 = 100$ cases and $n_2 = 100$ is applied on each background realization. The cases are denoted with pluses (+) and controls with circles (∘). We take $\delta = 0.6$ in case 2, $\delta = 0.4$ in case 3, $k = 3$ in case 4, and $(\kappa, \mu) = (5, 40)$ in case 5.

because we have $(T_1 - \mathbf{E}[T_1]) / \sqrt{\mathbf{Var}[T_1]} = Z^D_{11}$ in the two-class case. Also, S^D_{ij} for $i, j = 1, 2$ and the corrected versions, $S^{D,c}_{ij}$, for $i, j = 1, 2$ provide very similar empirical size estimates, hence we presented only the former. In our empirical size analysis (and also in the power analysis in Section 6), we use standardized forms of Pielou's coefficient of segregation and Cuzick–Edwards's tests. That is, we use $Z_P = S_P / \mathbf{Var}[S_P]$, and $(T_k - \mathbf{E}[T_k]) / \sqrt{\mathbf{Var}[T_k]}$ for $k = 1, 2$, and $(T_S - \mathbf{E}[T_S]) / \sqrt{\mathbf{Var}[T_S]}$ for $S = \{1, 2\}$.

In case 1, we have the background pattern from an HPP; that is, each realization of $\mathcal{Z}_n$ is from the CSR pattern. In this case, we investigate the effect of equal but increasing sample sizes, and differences in the relative abundances (in both directions, with fixed number of cases and increasing number of controls and vise versa). In case 2, we consider an increasing level of clustering along the diagonal $y = x$ with increasing δ, the two clusters intersect for $\delta < 1$ and the clusters become disjoint for $\delta > 1$. In case 3, we already have two disjoint clusters along the x-axis, and the level of clustering increases with increasing δ. Hence in cases 2 and 3, the effect of clustering level on the empirical sizes is assessed. In case 4, we already have k disjoint clusters with $\delta = 0.5$ and assess the effect of number of clusters on the empirical sizes. In case 5, we have clusters where the size and location of the clusters are random according to a Matérn clustering process. In this case, we assess the effect of such clustering on the empirical sizes.

In Figures 2 and 3, we present the empirical size estimates for the right-sided alternative (i.e., toward segregation) only. We deferred the sizes for left-sided alternative (i.e., toward association) to the technical report [31], because association is not the relevant direction for the disease clustering considered. The empirical size estimates are computed as follows. For each Monte Carlo replication, we computed test statistics, and the size is estimated on the basis of the asymptotic critical values. For (standardized versions of) Pielou's coefficient of segregation, Dixon's segregation indices, Dixon's cell-specific tests, type III cell-specific tests, and Cuzick–Edwards's k NN and combined tests, we use the critical value $z_{.95} = 1.96$ for the right-sided (clustering or segregation) alternative (and $z_{.05} = -1.96$ would have been used for the left-sided (association) alternative). For example, the empirical size of S_P is calculated for

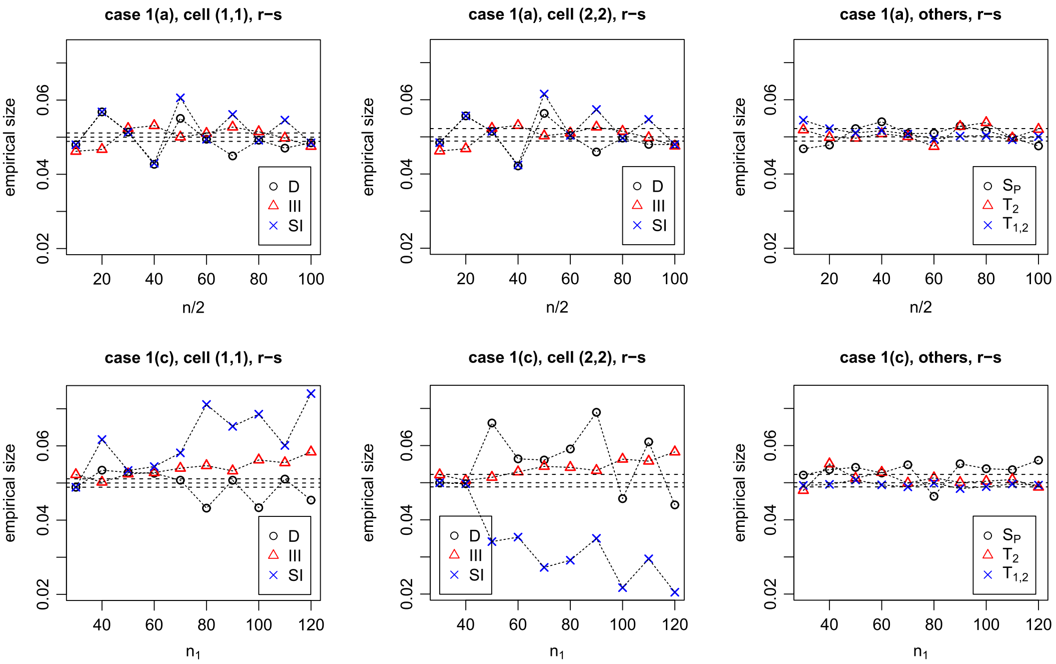


Figure 2. The empirical size estimates for the tests under the random labeling of points from background cases 1(a) and 1(c) for the right-sided alternative. In case 1(a) (top row), we take $n_1 = n_2 = n/2 = 20, 30, \ldots, 100$, and in case 1(c) (bottom row), we take $n_1 = 30, 40, \ldots, 100$ and $n_2 = 30$. In the plot titles, r-s stands for 'right-sided', and in the legends, D stands for Dixon's cell-specific tests, III for type III cell-specific tests, SI for Dixon's segregation indices, S_P for Pielou's coefficient of segregation, T_2 for Cuzick–Edwards's 2 NN test, and $T_{1,2}$ for Cuzick–Edwards's combined test, T_S, for $S = \{1, 2\}$. The dashed horizontal lines are at 0.04887 and 0.05113, the lower and upper bounds for significant deviation from 0.05. Also, empirical size estimates for each test are joined by straight lines for better visualization.

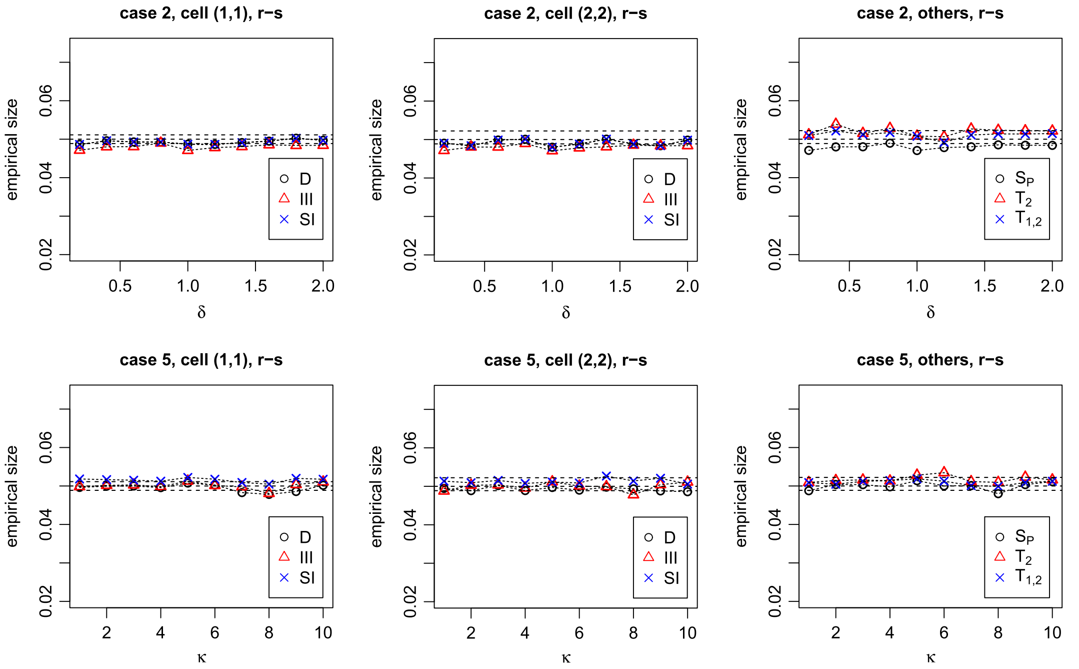


Figure 3. The empirical size estimates of the tests for the right-sided alternative under the random labeling case 2 (top row) with $n_1 = n_2 = 100$, and we take $\delta = 0.2, 0.4, \ldots, 1.4$, and under the random labeling case 5 (bottom row) with n_1 and n_2 being about half the number of generated points from the Matérn cluster process, and we use $\kappa = 1, 2, \ldots, 5$. The dashed horizontal lines and legend labeling are as in Figure 2.

the right-sided alternative as $\frac{1}{N_{mc}} \sum_{i=1}^{N_{mc}} \mathbf{I}(Z_{P,i} > 1.96)$ where we have 1000 Monte Carlo replications for each of background realizations, and because there are 100 different realizations, we would have $N_{mc} = 100000$ and $Z_{P,i}$ is the standardized version of Pielou's coefficient of segregation for the sample in i^{th} replication. On the other hand, for Dixon's overall test, we use 95th percentile of χ^2_1 distribution, which is $\chi^2_{1,.95} = 3.84$, and for type III overall test, we use $\chi^2_{2,.95} = 5.99$.

We present the empirical significance levels under cases 1(a) and 1(c) for the right-sided alternative in Figure 2. In case 1(a), we have equal but increasing sample sizes (i.e., $n_1 = n_2 = n/2 = 10, 20, \ldots, 100$), and as expected, the size performance gets better (i.e., empirical sizes tend to approach to the nominal size of 0.05) as n increases. Furthermore, all the tests have empirical size estimates around the null region (i.e., around the band between 0.04887 and 0.05113). These bounds for the null region are estimated as follows. With $N_{mc} = 100000$, an empirical size estimate larger than 0.05113 is deemed liberal, whereas an estimate smaller than 0.04887 is deemed conservative at 0.05 level (based on binomial critical values with $n = 100000$ trials and probability of success 0.05).

Among the cell-related tests (i.e., cell-specific tests and Dixon's segregation indices), size estimates of type III test are closer to the nominal level of 0.05. When all the tests considered type III tests, Pielou's test and Cuzick–Edwards's tests have less fluctuation around 0.05, and $T_{1,2}$ is closest to the nominal level and has the least fluctuation. For the left-sided alternative (i.e., toward association), Dixon's cell-specific tests fluctuate more around 0.05, compared with other tests, and Dixon's segregation indices are extremely liberal. Among cell-related tests, type III has the best size performance. Furthermore, T_2 is mostly conservative, and S_P fluctuates around the null region but is close to it. All tests considered, $T_{1,2}$ is closest to the nominal level and has the least fluctuation, then comes type III cell-specific test and S_P. It should be noted at this point that it is not quite fair to compare T_2 and $T_{1,2}$ with the tests related to NNCTs. T_2 tests the spatial interaction at around second NN distance, and $T_{1,2}$ tests the spatial interaction around the first and second NN distances (and is expected to perform better because it uses more information), whereas tests based on NNCTs test the interaction around the first NN distance. Thus, the tests based on NNCTs and T_1 test the interaction at the same scale; however, T_1 is equivalent to Dixon's cell (1, 1) test.

In case 1(c), we have $n_2 = 30$ and $n_1 = 30, 40, \ldots, 120$; that is, the difference in relative abundance increases as n_1 increases, and in this case, with increasing n_1, the disease incidence rate is increasing. Hence, in this case, we investigate the effect of increasing incidence rate (starting from 50% and increasing to 80%) on the empirical sizes. For the right-sided alternatives, among cell $(2,2)$ statistics, Dixon's test fluctuates between liberalness and conservativeness, Dixon's segregation index tends to be conservative (with level of conservativeness increasing with n_1), and type III statistic is closest to the null region, but its size estimate seems to increase with n_1. Among cell $(1,1)$ statistics, Dixon's segregation indices tend to be liberal (with level of liberalness increasing with n_1), and type III cell-specific test has the same performance as in cell $(2,2)$, and Dixon's cell-specific test is closest to the nominal level. S_P, T_2 and $T_{1,2}$ are much closer to the null region, with $T_{1,2}$ being closest. All tests considered, S_P, T_2, and $T_{1,2}$ have best performance, with $T_{1,2}$ having slightly better performance. For the left-sided alternatives, among cell $(2,2)$ statistics, Dixon's test fluctuates between conservativeness and the desired level (and tends to get more conservative with increasing n_1), Dixon's segregation index tends to be extremely liberal (although fluctuating, the level of liberalness tends to increase with n_1), and type III cell-specific test is close to the null region, but its size estimate seems to decrease to become conservative with n_1. Among cell $(1,1)$ statistics, Dixon's segregation indices are starting liberal and getting conservative eventually with increasing n_1, and type III has the same performance as in cell $(1,1)$. S_P, T_1, and $T_{1,2}$ are closest to the null region (and slightly conservative for some values of n_1). All tests considered, S_P, T_2, and $T_{1,2}$ have best performance, with $T_{1,2}$ having slightly better performance. Hence, the differences in the relative abundances increasing in favor of cases (i.e., increasing incidence rate of the disease) confounds most test statistics. Among the tests considered, $T_{1,2}$ seems to be the most robust to such differences in sample sizes, whereas cell-related tests are severely confounded by such differences. Among the tests for small scale interaction, S_P is most robust to differences in relative abundances. The better performance of Cuzick–Edwards's tests in this case is no coincidence, because these tests are designed to detect the clustering of cases (i.e., class 1 points), and the number of class 1 points increases in this case.

In case 1(b), we have $n_1 = 30$ and $n_2 = 30, 40, \ldots, 120$; that is, the difference in relative abundance increases as n_2 increases, and in this case, with increasing n_2, the disease incidence rate is decreasing. Hence, in this case, we investigate the effect of decreasing incidence rate (starting from 50% and decreasing to 20%) on the empirical sizes. The trends in S_P and type III tests are as in case 1(c); with the roles of classes switched, the tests yield the same results for a given data. Furthermore, Dixon's cell (i,i) statistics and segregation indices behave similar to those for cell (j,j) of case 1(c) for $i \neq j$ switching also n_1 with n_2. Hence, case 1(b) empirical sizes are not presented. However, in this case, the performance of T_2 and $T_{1,2}$ deteriorate, and they tend to become more liberal as n_2 increases, and S_P has the best performance. Hence, the differences in the relative abundances increasing in favor of controls (i.e., decreasing incidence rate of the disease) confound most test statistics. Among the tests considered, S_P seems to be the most robust statistics to such differences in sample sizes (i.e., for decreasing incidence rates) in both directions, whereas T_2 and $T_{1,2}$ are more robust to differences in favor of cases (i.e., for increasing incidence rates).

We presented the empirical size estimates under cases 2 and 5 for the right-sided alternative in Figure 3. The size estimates under cases 3 and 4 are similar to case 2, hence are omitted. In case 2, we have equal sample sizes with $n_1 = n_2 = 100$, but with increasing δ, the level of clustering of the two clusters in the background pattern increases (in fact, with $\delta > 1$, the clusters get disjoint). For the right-sided alternative, all tests are almost within the null region with Dixon's cell $(1,1)$ statistics closest to the nominal level. Furthermore, Dixon's cell-specific tests and segregation indices exhibit almost identical size performance; T_2 and $T_{1,2}$ tend to be slightly liberal, whereas others tend to be slightly conservative. For the left-sided alternative, all tests except Dixon's segregation indices are almost within the null region, but T_2, $T_{1,2}$, S_P, and type III tests tend to slightly conservative, whereas Dixon's cell-specific tests are slightly liberal and Dixon's segregation indices are severely liberal with size estimates being about 0.06. Hence, with sample sizes being equal and large, most tests are unaffected seriously with increasing level of clustering in the background realizations, and Dixon's segregation indices are most severely confounded by increasing δ; T_2 and $T_{1,2}$ have better size performance for both alternatives (with $T_{1,2}$ being the best). There is no clear (increasing or decreasing) trend in the size estimates of the tests with increasing δ. In case 4, we also observe that with sample sizes being equal and large, the sizes of the tests are not affected by the increasing number of clusters in the background realizations.

We presented the empirical size estimates under case 5 for the right-sided alternative in Figure 3. In this case, we have sample sizes $n_1 = n_2 = 100$ on the average, and random number of clusters κ (with increasing κ, the number of clusters tend to increase), and the locations of the clusters are also random. For the right-sided alternative, all tests are almost within the null region, but Dixon's segregation indices for cell $(1,1)$ and T_2 tend to be slightly liberal for some of the κ values, whereas other tests are around 0.05. Type III tests, Dixon's cell-specific tests, and S_P seem to have the best performance. For the left-sided alternative, all tests except Dixon's segregation indices are almost within the null region, but Dixon's segregation indices tend to be liberal, and T_2 and $T_{1,2}$ are slightly conservative. Type III tests, S_P, and $T_{1,2}$ seem to have the best performance. Hence, with randomly occurring and randomly increasing number of clusters, most tests are not affected seriously. Dixon's segregation indices have the worst size performance under RL of this type of background clustering.

We presented the empirical size estimates for the overall NNCT tests under cases 1–5 in Figure 4. Dixon's overall test is severely liberal in cases 2–5, and conservative for small samples and liberal for large samples in case 1(a), and conservative or within the null region in cases 1(b) and (c). On the other hand, type III overall test is slightly conservative or within the null region for all cases and has better performance compared with Dixon's overall test. Furthermore, there is no clear trend in the size estimates as the equal sample sizes increase, or level and number of clusters increase. On the other hand, as the discrepancy between the sample sizes (i.e., differences in relative abundances) in cases 1(b) and (c) increases, the size estimates of the overall tests tend to decrease eventually.

6. Empirical power analysis of the tests under non-random labeling alternatives

We propose various non-RL alternatives where case and control labels are assigned (with a pattern deviating from RL pattern) to the points generated from various homogeneous or clustering processes.

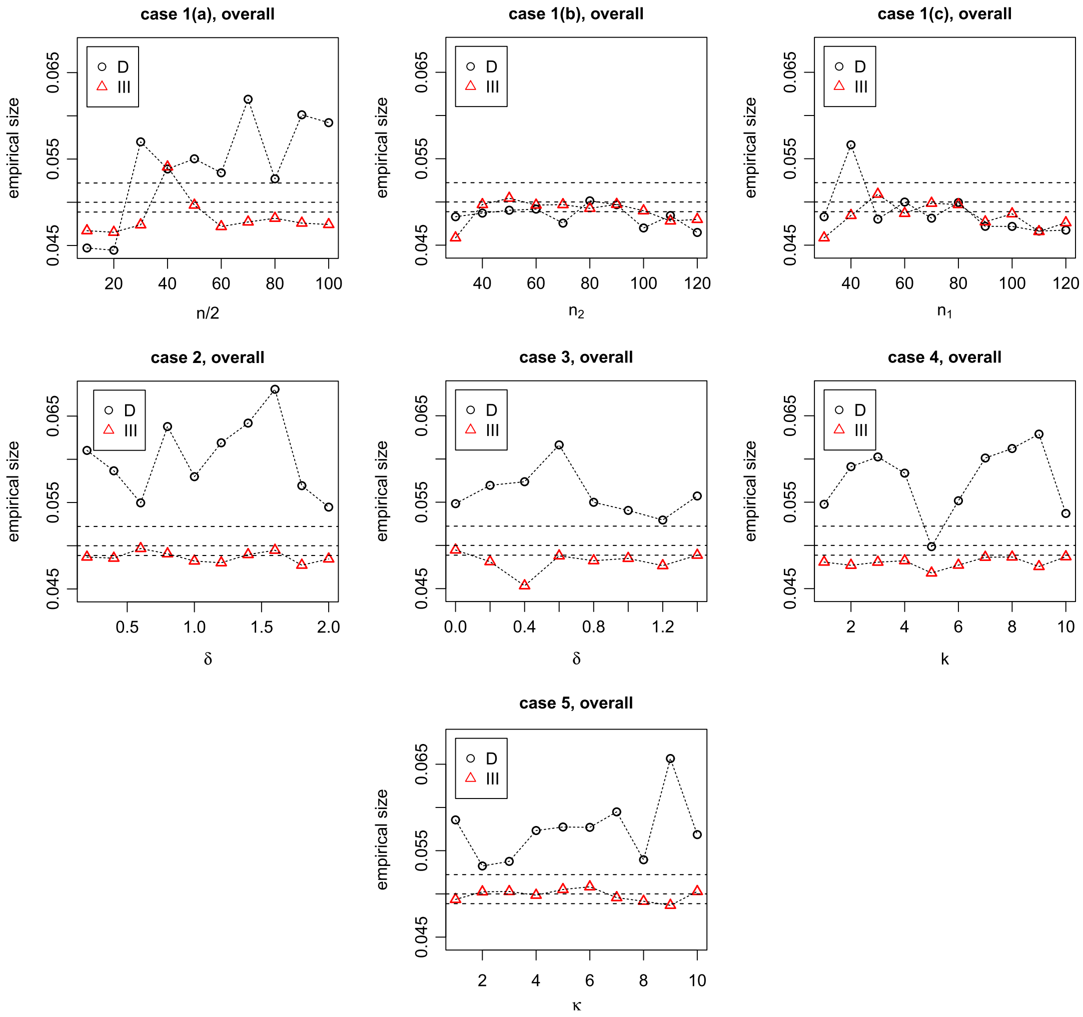


Figure 4. The empirical size estimates of the overall nearest neighboring contingency table tests under the random labeling of points from background cases 1(a)–(c) (top row in that order from left to right), and cases 2–5 (starting at second row and ordered from left to right). The dashed horizontal lines are as in Figure 2, and in the legends, D stands for Dixon's overall test, and III stands for type III overall test.

In all these alternatives, we generated the background points in $\mathcal{Z}_n$ independently uniformly in the unit square $(0,1) \times (0,1)$, that is, $Z_i \overset{iid}{\sim} \mathcal{U}((0,1) \times (0,1))$ for $i = 1, 2, \ldots, n$. To remove the effect of one particular realization of the points on the test, we consider 100 different realizations. We only use realizations from HPP pattern for the background, because the level and number of clusters seem not to affect the size performance of the tests. Hence, in the non-RL alternatives, we only consider various non-RL schemes on the points from HPP.

Types of the non-RL patterns:
Case 1: Select a Z_i randomly from $\mathcal{Z}_n$, label it as a case. Find its k NNs and label them as cases with probabilities proportional to $\frac{n_1}{n} + \rho\left(1 - \frac{n_1}{n}\right), \frac{n_1}{n} + \frac{\rho}{2}\left(1 - \frac{n_1}{n}\right), \ldots, \frac{n_1}{n} + \frac{\rho}{k}\left(1 - \frac{n_1}{n}\right)$ until the number of cases first exceeds n_1. We use (a) $\rho = -0.2, 0.0, 0.2, 0.4, 0.6, 0.8$ and $k = 1$, which only assigns the first NN and (b) $\rho = 0.0, 0.2, 0.4, 0.6, 0.8$ and $k = 3$, which assigns the first three NNs according to the aforementioned probabilities.
Case 2: In this case, we have an initial proportion, π_i, and an ultimate proportion, π_u, with $\pi_u > \pi_i$. First, label the initial proportion, π_i, of points as cases randomly and pick a case among them randomly. Then find the k NNs of this case and label them as cases with probabilities proportional to $\rho, \rho/2, \ldots, \rho/k$. Select a point randomly among these k NNs, find its k NNs, and assign them as cases with the aforementioned probabilities until we have the proportion of cases first exceeding π_u. We use $\pi_i = 0.3$, $\pi_u = 0.5$, and $\rho = 0.2, 0.4, 0.6, 0.8$ and consider (a) $k = 1$, which only assigns the first NN, and (b) $k = 3$, which assigns the first three NNs according to the aforementioned probabilities.
Case 3: Pick a Z_i randomly from $\mathcal{Z}_n$, mark it as a case and label others as a case with probabilities inversely proportional to their distances to Z_i. More specifically, we use probabilities proportional to $\frac{\rho}{k_d}\left(1 - \frac{d_{ji}}{d_{\max}}\right)^{k_p}$ where d_{ji} is the distance from Z_j to Z_i for $j \neq i$, $d_{\max}$ is the maximum of d_{ji} values, $k_p > 0$ and $k_d \geq 1$. We stop when we first exceed n_1 cases. In our simulations, we employ the usual Euclidean distance and use (a) $\rho = 0.2, 0.4, \ldots., 1.0$, $k_d = 1$, and $k_p = 3$; (b) $\rho = 0.8$, $k_d = 3, 6, \ldots, 15$, and $k_p = 3$; and (c) $\rho = 0.8$, $k_d = 1$, and $k_p = 1, 2, \ldots, 5$.
Case 4: Pick k_0 points $z_1^{'}, z_2^{'}, \ldots, z_{k_0}^{'}$ from $\mathcal{Z}_n$ randomly as sources. Let φ_G be the PDF of $BVN(\mu, \sigma_1 = \sigma_2 = \sigma, \rho = 0)$ where $BVN(\mu, \sigma_1, \sigma_2, \rho)$ stands for the bivariate normal distribution with mean vector $\mu = (\mu_1, \mu_2)$, standard deviations of univariate components are σ_1 and σ_2, and the correlation between the components is ρ. Then for each $j = 1, 2, \ldots, k_0$, compute $\varphi_{G,j}(z_i)$ for all $i = 1, 2, \ldots, n$ where $\varphi_{G,j}$ is the PDF of $BVN(\mu = z_j^{'}, \sigma_1 = \sigma_2 = \sigma, \rho = 0)$ and add these PDF values. That is, find $p_G(z_i) = \sum_{j=1}^{k_0} \varphi_{G,j}(z_i)$ for each $i = 1, 2, \ldots, n$. Then label the points as cases with probabilities proportional to the value of the PDF sums at these points. More specifically, we use probabilities $\frac{1}{p_{\max}}(p_G(z_1), p_G(z_2), \ldots, p_G(z_n))$ where $p_{\max} = \max_{i=1}^{n} p_G(z_i)$. We stop when we first exceed n_1 cases. We use (a) $k_0 = 3$ and $\sigma_1 = \sigma_2 = \sigma = 0.1, 0.2, \ldots, 0.8$; and (b) $k_0 = 1, 2, \ldots, 8$ and $\sigma_1 = \sigma_2 = \sigma = 0.4$.

We plot sample realizations from these non-RL cases in Figure 5. We simulate 1000 Monte Carlo replications for each parameterization in each case at each background realization. For example, with a particular background realization, in case 3(a), we simulate 1000 replications for $k_p = 3$, and $k_d = 1$ and each of $\rho = 0.2, 0.4, \ldots., 1.0$ with $n_1 = n_2 = 100$.

We computed the empirical power estimates similar to the empirical sizes. Furthermore, S_{ij}^D for $i, j = 1, 2$, and the corrected versions, $S_{ij}^{D,c}$ for $i, j = 1, 2$, provide very similar empirical power estimates; hence only the former are presented.

We plotted the power estimates based on z-scores in Figures 6–9. In all these cases, Dixon's cell-specific test and segregation index for cell (i, i) provide very similar power estimates. Furthermore, we only consider right-sided alternatives, because by design, the non-RL alternatives are for segregation (or clustering) of class 1, and the power estimates for the left-sided alternatives are virtually zero.

We presented the empirical power estimates under cases 1(a) and (b) in Figure 6. Notice that as ρ increases, the power estimates tend to increase as well. That is, when the probability of assigning the same label (i.e., class 1 label) to NNs increases, the level of segregation; hence, the power of the tests increases. Furthermore, the power estimates are higher for $k = 1$ (case 1(a)) compared with $k = 3$ (case 1(b)) for each test. Hence, in this type of non-RL with ρ, n_1, n_2 being fixed, as the number of NNs to be labeled increases, the power estimate tends to decrease, that is, the level of segregation decreases. In

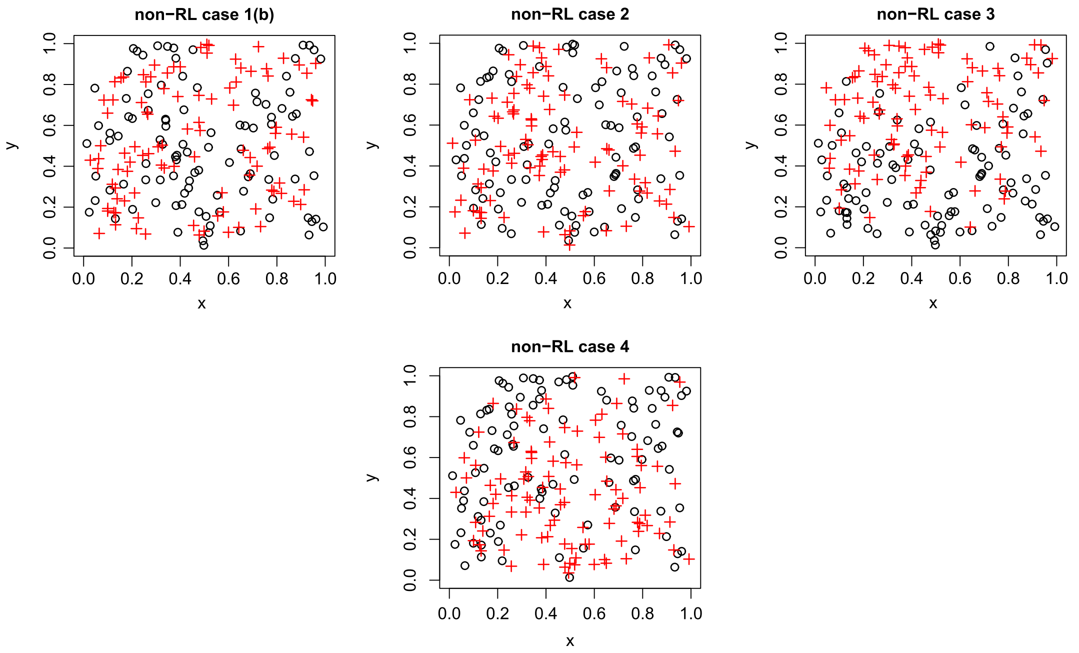


Figure 5. Sample plots of the realizations of non-random labeling cases 1–4 with $n_1 = n_2 = 100$ with the same background pattern from homogeneous Poisson process. The cases are denoted with pluses (+) and controls with circles (∘). We take $\rho = 0.8$ and $k = 3$ in case 1; $\rho = 0.8$ and $k = 1$ in case 2; $\rho = 0.4$, $k_d = 1$, and $k_p = 3$ in case 3; and $k_0 = 3$ and $\sigma = 0.4$ in case 4.

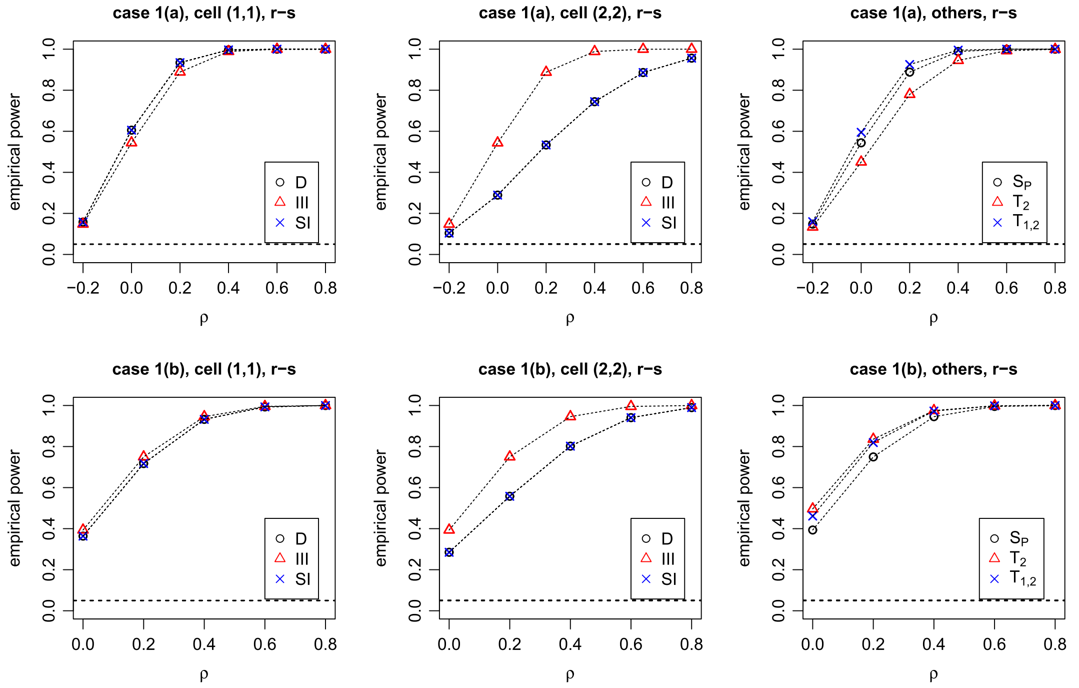


Figure 6. Empirical power estimates under the non-random labeling cases 1(a)–(b). In case 1(a) (top row), we take $\rho = -0.2, 0.0, 0.2, \ldots, 0.8$ and $k = 1$, and in case 1(b) (bottom row), we take $\rho = 0.0, 0.2, \ldots, 0.8$ and $k = 3$. The dashed horizontal lines are at 0.05 and 1.0, and legend labeling is as in Figure 2.

case 1(a), among cell (1, 1) statistics, Dixon's cell-specific test and segregation index have slightly higher power compared with type III cell-specific test; among cell (2, 2) statistics, type III statistics have much higher power than Dixon's tests; and among other test statistics, Pielou's coefficient of segregation and $T_{1,2}$ have higher power (with the latter having highest power). In case 1(b), among cell-specific tests,

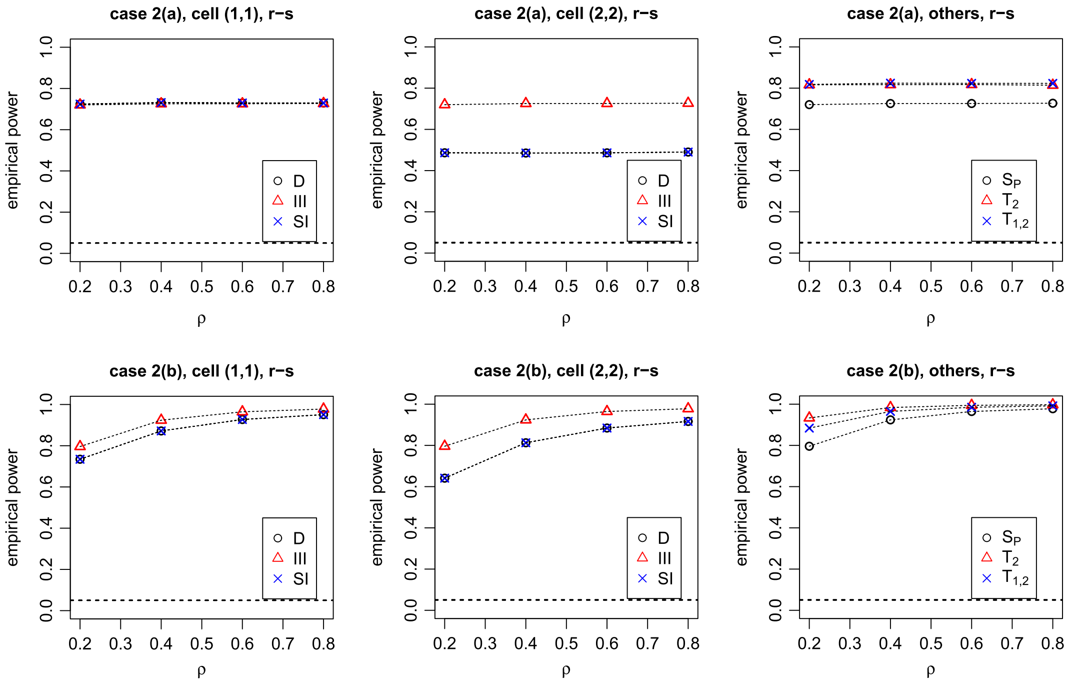


Figure 7. Empirical power estimates under the non-random labeling case 2(a) (top row) and case 2(b) (bottom row) with $\pi_i = 0.3$, $\pi_u = 0.5$ and $\rho = 0.2, 0.4, 0.6, 0.8$. The dashed horizontal lines and the legend labeling in both rows are as in Figure 6.

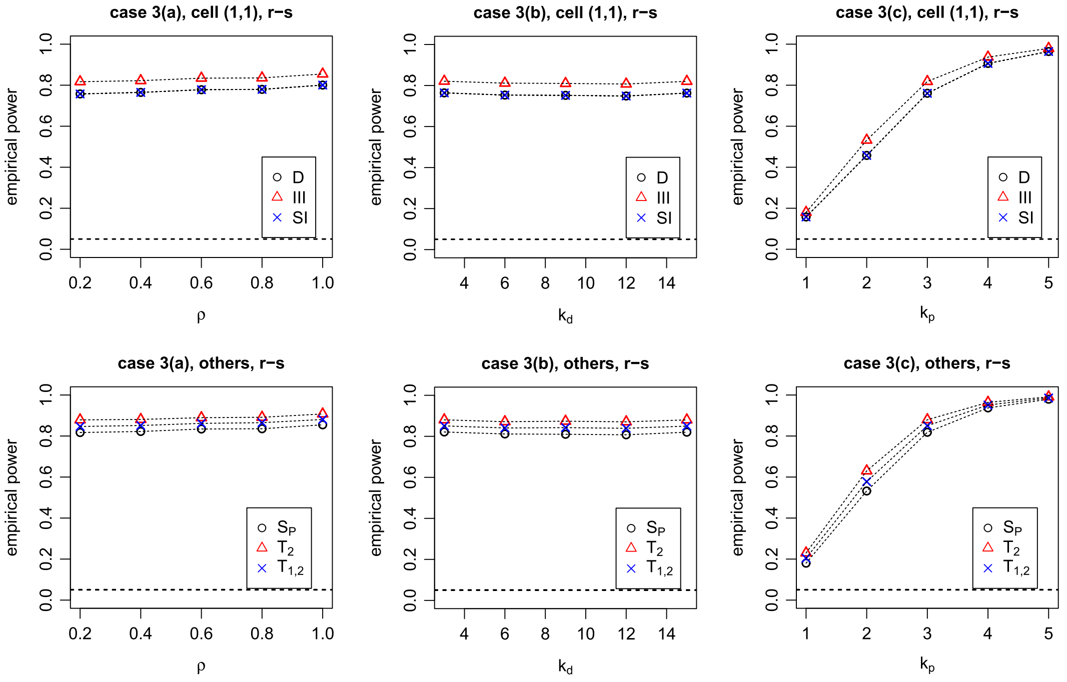
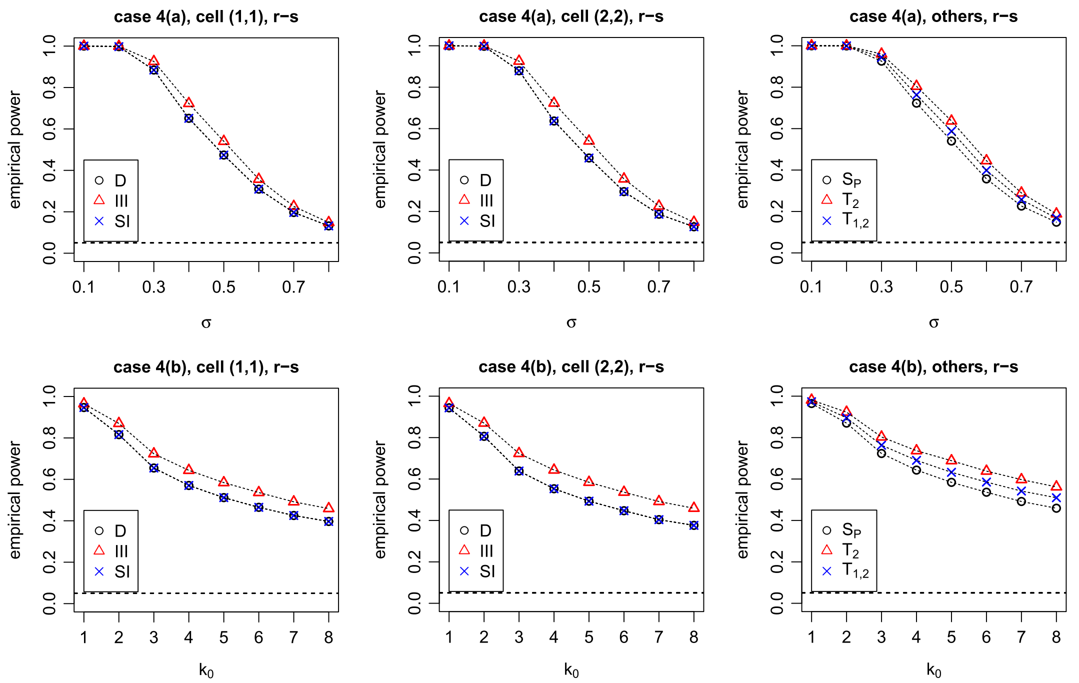


Figure 8. Empirical power estimates under the non-RL cases 3(a)-(c). In case 3(a), we take $\rho = 0.2, 0.4, \ldots, 1.0$, $k_p = 3$, and $k_d = 1$; in case 3(b), we take $\rho = 0.8$, $k_p = 3$, and $k_d = 3, 6, \ldots, 15$; and in case 3(c), we take $\rho = 0.8$, $k_p = 1, 2, \ldots, 5$, and $k_d = 1$. The dashed horizontal line and legend labeling are as in Figure 6.

type III test has higher power; and among others, Cuzick–Edwards's tests, T_2, and $T_{1,2}$, have higher power (with the former having highest power).

We presented the empirical power estimates under cases 2(a) and (b) in Figure 7. In case 2(a), the power estimates are almost constant, with Cuzick–Edwards's tests having power around 0.80, Dixon's cell $(2, 2)$ test and segregation index having power around 0.50, and all others having power around 0.70.



case 4(a), cell (1,1), r–s

case 4(a), cell (2,2), r–s

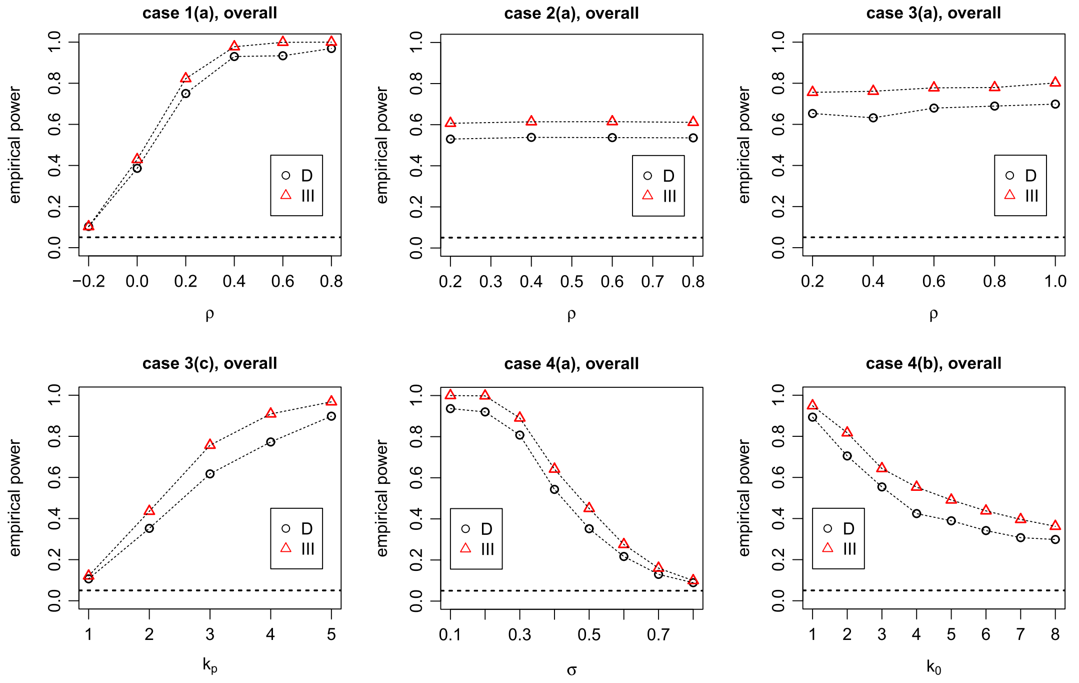
case 4(a), others, r–s

case 4(b), cell (1,1), r–s

case 4(b), cell (2,2), r–s

case 4(b), others, r–s

Figure 9. Empirical power estimates under the non-random labeling cases 4(a)–(b). In case 4(a) (top row), we take $k_0 = 3$ and $\sigma = 0.1, 0.2, \ldots, 0.8$; and in case 4(b) (bottom row), we take $k_0 = 1, 2, \ldots, 8$ and $\sigma = 0.4$. The dashed horizontal line and legend labeling are as in Figure 6.



case 1(a), overall

case 2(a), overall

case 3(a), overall

case 3(c), overall

case 4(a), overall

case 4(b), overall

Figure 10. Empirical power estimates for the overall nearest neighbor contingency table tests under the non-random labeling cases 1–4. D stands for Dixon's overall test, and III stands for type III overall test.

In case 2(b), the power estimates are higher compared with case 2(a), and they increase as ρ increases. Among the cell-related tests, type III test has higher power (and S_P has about the same power as the type III tests). Cuzick–Edwards's tests have the higher power estimates, with T_2 having the highest power. In this type of non-RL, the power seems not to depend on ρ if only the first NN is labeled according to the probabilities.

We presented the empirical power estimates under cases 3(a)–(c) in Figure 8. In cases 3(a) and (b), notice that the power estimates slightly increase as ρ increases, but it seems that the power estimates (hence, the level of segregation) does not crucially depend on k_d or ρ. In case 3(c), the power estimates tend to increase as k_p increases. Hence, as k_p increases, the probability of assigning the same label to NNs increases. In all these cases, among cell-specific tests, type III test has the highest power, and among others T_2 has highest power.

We presented the empirical power estimates under cases 4(a) and (b) in Figure 9. In case 4(a), notice that as σ increases, the power estimates tend to decrease. That is, when σ increases (with n_1, n_2, and k_0 being fixed), the probability of assigning the same label to NNs of the source points decreases, hence, the level of segregation; thereby, the power of the tests decreases as well. In case 4(b), as number of source points, k_0, increases, the power estimates tend to decrease. That is, when the number of source points increases (with n_1, n_2, and σ being fixed), the (relative) probability of assigning the same label to NNs of the source points decreases, hence, the level of segregation and the power of the tests decrease as well. In both cases, among cell-related tests, type III test has higher power, and among others, T_2 has higher power estimates.

The empirical power estimates of the NNCT overall tests under cases 1-4 are presented in Figure 10. The power estimates for cases 1(b), 2(b), and 3(b) are similar to those for cases 1(a), 2(a), and 3(a), respectively, hence are not presented. In all these cases, type III overall test has higher power estimates compared with Dixon's overall test. In cases 1(a) and (b), cases 2(a) and (b), the power estimates increase with increasing ρ (in all cases, the power estimates are higher with $k = 1$ compared to $k = 3$). In case 3(a) (respectively, (b)), the power estimates does not seem to depend on the parameter ρ (respectively, k_d). In case 3 (c), power estimates increase as k_p increases; in case 4(a) (respectively, (b)), power estimates decrease as σ (respectively, k_0) increases.

Remark

We also compute empirical power estimates based on Monte Carlo critical values. Under case 1(a) of RL pattern with $n_1 = n_2 = 100$, we compute the 95^{th} empirical percentiles of the test statistics computed in the Monte Carlo simulations and use these as the Monte Carlo critical values. For example, the empirical power (based on Monte Carlo critical value) for S_P is calculated for the right-sided alternative as $\frac{1}{N_{mc}} \sum_{i=1}^{N_{mc}} \mathbf{I}(S_{P,i} > S_{P,crit}^{mc})$ where we have $N_{mc} = 100000$ and $S_{P,crit}^{mc}$ is the 95^{th} empirical percentile of Pielou's coefficient of segregation under RL case 1(a) with $n_1 = n_2 = 100$. The power estimation for the other tests is similar. We observed that the power estimates using the asymptotic critical values and those using the Monte Carlo critical values are very similar for all tests. Hence, we only present the power estimates with the asymptotic critical values.

7. Example data sets

7.1. Childhood leukemia data

This data set consists of spatial locations of 62 cases of childhood leukemia in the North Humberside region of the UK, between the years 1974 and 1982 inclusive [11]. From the same region, we selected a random sample of 143 controls using the completely randomized design. We analyze the spatial clustering of leukemia cases with respect to controls in this data with the tests considered earlier. We plotted the locations of the points in the study region in Figure 11, and the segregation indices (together with standard errors) are provided in Table II. The figure is suggestive of mild clustering of leukemia cases, and the indices together with their standard errors suggest only mild segregation (if any). Here, the indices and their standard errors are sufficient for an initial clustering assessment, because either the indices have zero expected value (as in S_P) or their expectation is approximately zero (and tending to zero with increasing class sizes) as in Dixon's segregation indices.

The appropriate null hypothesis is the RL pattern, because it is reasonable to assume that some process affects a posteriori the population of North Humberside region so that some of the individuals get to be cases, while others continue to be healthy (i.e., they are controls) [37]. In Table III, we present the test statistics and the associated p-values based on of asymptotic critical values and on Monte Carlo randomization. The latter is estimated as follows. We compute the test statistics for the original data, and the labels are randomly assigned to the points 10,000 times. At each random assignment, we compute the test statistics and find how many times they equal or exceed the test statistics in the original data. This number divided by 10000 yields the p-values based on Monte Carlo randomization. Notice that both versions of p-values are similar for each test (except for T_2 and $T_{1,2}$). Observe that only T_2 and

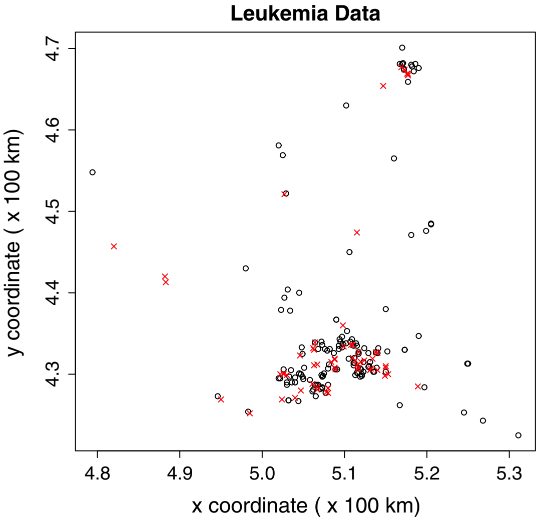


Figure 11. The scatter plots of the locations of cases (crosses ×) and controls (circles ○) in North Humberside leukemia data set.

Table II. Pielou's coefficient of segregation and Dixon's segregation indices (± standard errors) for the right-sided alternatives for North Humberside leukemia data.

Segregation indices for leukemia data				
S_P	S_{11}^D	S_{22}^D	$S_{11}^{D,c}$	$S_{22}^{D,c}$
0.1348 (±0.090)	0.3548 (±0.314)	0.2362 (±.175)	0.3420 (±0.272)	0.2317 (±0.251)

Table III. The corrected versions and the test statistics and the associated p-values for the right-sided alternatives for North Humberside leukemia data.

	Z_{11}^D	Z_{22}^D	Z_{11}^{III}	Z_{22}^{III}	Z_{11}^S	Z_{22}^S	Z_P	T_2	$T_{1,2}$	C_D	C_{III}
Test statistics for leukemia data	1.2021	1.2829	1.4568	1.4590	1.1292	1.3482	1.4983	2.6263	2.1206	2.2604	2.1254
Associated p-values, with asymptotic critical values	0.1147	0.0998	0.0726	0.0723	0.1294	0.0888	0.0670	0.0043	0.0170	0.3230	0.1449
Associated p-values, with Monte Carlo randomization	0.1365	0.0743	0.0784	0.0780	0.1294	0.1100	0.0726	0.0211	0.0696	0.4460	0.1462

Z_{ii}^D (Z_{ii}^{III}) is Dixon's (type III) cell-specific test for cell (i,i), Z_{ii}^S is the standardized version of Dixon's segregation indices for cell (i,i), $i = 1,2$, Z_P is the standardized version of Pielou's coefficient of segregation, T_2 is Cuzick–Edwards's 2 NN test, $T_{1,2}$ is Cuzick–Edwards's combined test for $k = 1,2$, C_D, and C_{III} are Dixon's and type III overall tests, respectively.

$T_{1,2}$ are significant at 0.05 level, while all others are not. Hence, we conclude that there is no significant segregation of cases at small scales (about the first NN distances), but cases tend to cluster significantly at larger scales (about the second NN distances). The standardized versions of the corrected segregation indices are $Z_{11}^{D,c} = 1.2591$ and $Z_{22}^{D,c} = 1.076$ with the p-values for the right-sided alternative are 0.1040 and 0.1410, respectively. The corresponding p-values based on Monte Carlo randomization are 0.1294 and 0.1100, respectively.

On the basis of the tests discussed earlier, we conclude that the cases and controls do not exhibit significant clustering (i.e., segregation) at small scales. On the basis of Cuzick–Edwards's tests, we find that the cases are significantly segregated around k NN distances for $k = 2$. In particular, average NN distance for leukemia data is 700 m (±1400 m), and the aforementioned analysis summarizes the pattern for about $t = 1000$ m, except for T_2 and $T_{1,2}$ where T_2 summarizes the pattern at about 1350 m (because the average 2-NN distance is 1342 m), and $T_{1,2}$ for distances between 1000 and 1350 m.

7.2. *Liver data*

This data set consists of spatial locations of 761 cases of a liver disease in a region of interest and 3044 controls in the same region [2]. We analyze the spatial clustering of liver disease cases with respect to the healthy controls. The locations of the points are plotted in Figure 12 and the segregation indices (together with standard errors) are provided in Table IV. Observe that the plot of locations is suggestive of strong clustering of cases, and the indices together with the standard errors support this initial assessment.

As in the leukemia data set, the appropriate null hypothesis is again the RL pattern. In Table V, we present the test statistics and the associated p-values based on asymptotic critical values and on Monte Carlo randomization where the latter is estimated as in Section 7.1. Both versions of p-values are similar for each test. Observe that all tests except Z^D_{22} and Z^S_{22} are significant at 0.05 level (but their Monte Carlo randomized versions are significant), implying significant segregation of cases at small scales (about the first NN distances) and at larger scales about the second NN distances. That is, cases tend to cluster significantly at smaller scales. The standardized versions of the corrected segregation indices are $Z^{D,c}_{11} = 2.9077$ and $Z^{D,c}_{22} = 1.3813$ with the p-values for the right-sided alternative are 0.0018 and 0.0836, respectively. The corresponding p-values based on Monte Carlo randomization are 0.0004 and 0.0348, respectively.

The aforementioned tests indicate a significant segregation of cases and controls, and segregation of cases from controls seems to be much stronger compared with that of controls from cases. This implies a significant clustering of cases at smaller scales around the average first NN distance. Similarly, Cuzick–Edwards's tests also imply significant segregation of cases and controls around k NN distances for $k = 2$. In particular, average NN distance for liver data is 34.24 (± 61.20), and the aforementioned analysis summarizes the pattern for about $t = 35$, except for T_2 and $T_{1,2}$ where T_2 summarizes the pattern at about 50 (because the average 2-NN distance is 52.20) and $T_{1,2}$ for distances between 35 and 50 units. Notice that by construction, Cuzick–Edwards's tests for T_k with $k > 1$ and T_S with $\{1\} \subsetneq S$

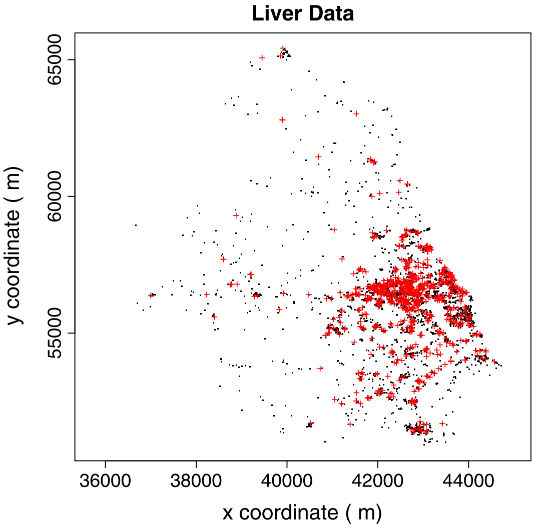


Figure 12. The scatter plots of the locations of cases (pluses +) and controls (dots ·) in Diggle's liver data set.

Table IV. Pielou's coefficient of segregation and Dixon's segregation indices (± standard errors) for the right-sided alternatives for Diggle's liver data.

Segregation indices for liver data				
S_P	S^D_{11}	S^D_{22}	$S^{D,c}_{11}$	$S^{D,c}_{22}$
0.0654 (± 0.025)	0.3410 (± 0.117)	0.0712 (± 0.051)	0.3393 (± 0.117)	0.0711 (± 0.051)

Table V. The corrected versions and the test statistics and the associated p-values for the right-sided alternatives for Diggle's liver data.

	Z_{11}^{D}	Z_{22}^{D}	Z_{11}^{III}	Z_{22}^{III}	Z_{11}^{S}	Z_{22}^{S}	Z_P	T_2	$T_{1,2}$	C_D	C_{III}
Test statistics for liver data	3.2024	1.3520	3.2732	3.2729	2.9055	1.3814	2.5737	9.1854	7.7709	10.9096	10.7134
Associated p-values, with asymptotic critical values	0.0007	0.0882	0.0005	0.0005	0.0018	0.0836	0.0050	< 0.0001	< 0.0001	0.0043	0.0011
Associated p-values, with Monte Carlo randomization	0.0004	0.0348	0.0004	0.0004	0.0004	0.0348	0.0020	< 0.0001	< 0.0001	0.0007	0.0007

The labels of the tests are as in Table III.

provide information not available by the other tests considered. However, this comes with a huge computational cost, because for liver data, it took about 7 h to compute the Cuzick–Edwards tests T_1, T_2, and $T_{1,2}$ in a HP Pavilion dv6 (Core i7 3720QM Processor 2.6 GHz, 8-GB RAM) laptop, but the other NNCT-based tests took only about 5 min. The time difference was not that crucial for leukemia data as Cuzick–Edwards's test took about 8 s, whereas NNCT-tests took only about 0.5 s. Our simulations indicate that NNCT tests have $O(n^2)$ computing time, but Cuzick–Edwards's tests T_1, T_2 and $T_{1,2}$ together have $O(n^{5/2})$ computing time. Hence, when the number of cases or controls is large (more than a few hundred), Cuzick–Edwards's tests are not computationally feasible, but the NNCT tests still are.

8. Discussion and conclusions

In this article, we propose the use of two segregation indices, namely, Pielou's coefficient of segregation [15] and Dixon's segregation indices [17] as tests to detect segregation between two classes, in particular to detect significance of disease clustering. We derive their asymptotic distributions under RL of cases and controls to given locations, and compare these tests with some other distance-based tests (such as Dixon's and type III cell-specific and overall tests, and Cuzick–Edwards's k NN and combined tests) in terms of empirical size and power via extensive Monte Carlo simulations. The tests related to NNCTs (i.e., Pielou's coefficient of segregation, Dixon's segregation indices, Dixon's and type III cell-specific, and overall tests) are for testing interaction at smaller scales about the first NN distance, and T_1 is equivalent to Dixon's cell $(1, 1)$ test while T_2 is for the interaction at about the second NN distance, and $T_{1,2}$ combines the interaction information at the first and second NN distances.

We investigate the effect of the clustering (i.e., level of clustering and number of clusters) of the background points (on which RL is applied) and the effect of the differences in relative abundances on the size of these tests. Our simulation results suggest that there is no increasing or decreasing trend in size when the number of clusters or level of clustering increases. On the other hand, the differences in relative abundances have a much stronger influence on the size performance of the tests. For the tests of small-scale interaction (around the first NN distance), we observe that Pielou's coefficient of segregation and type III overall tests seem to be robust to differences in relative abundances with Pielou's coefficient of segregation being more robust. On the other hand, for tests of higher-scale interaction (around or up to the second NN distance), T_2 and $T_{1,2}$ are both robust, with $T_{1,2}$ being more robust. Furthermore, among cell-related and overall tests, type III tests have better size performance, and when all tests are considered, Pielou's coefficient of segregation and T_2 and $T_{1,2}$ have better size performance.

We introduce four new non-RL algorithms yielding clustering of cases (or segregation between the classes) after the algorithm is executed on the background points. With these non-RL alternatives, we assess the power performance of the tests and see that type III tests and Cuzick–Edwards's tests have higher power than others (also we notice that Pielou's coefficient of segregation has power estimates close to Cuzick–Edwards's tests, although slightly lower). As for the computational complexity, Cuzick–Edwards's tests require much longer time and hence not so feasible for large sample sizes; on the other hand, the tests based on NNCTs require reasonable times even if sample sizes are on the order of thousands.

The methodology introduced in this article can also be used to test the deviations from CSR independence. But in this setting, the tests would be conditional on the values of Q and R, which are no longer fixed, but random quantities. Furthermore, the methodology is also applicable to test the spatial interaction at other contexts (e.g., the spatial interaction between plant species in ecology). In these contexts, the left-sided (or association) alternative could also be of practical interest.

Our simulation study suggests that Dixon's segregation indices do not fare well in testing spatial clustering. Hence, Dixon's segregation indices should not be employed with the asymptotic critical values in testing spatial clustering, but its Monte Carlo randomized version can be used. On the other hand, Pielou's coefficient of segregation performs similar to the best performing tests based on NN distances (at the scale, it is intended to work, i.e., at about the first NN distance). Considering both size and power performance of the tests together, for the interaction at small scales (around the first NN distance), we recommend Pielou's coefficient of segregation. In fact, if the relative abundances of the classes are similar, either type III tests or Pielou's coefficient of segregation can be employed, but if the relative abundances of the classes are different, Pielou's coefficient of segregation is recommended. For the interaction at higher scales, we recommend Cuzick–Edwards's k NN test with $k > 1$ and combined tests T_S with $\{1\} \subsetneq S$ for testing segregation (or disease clustering) against RL with the caveat of their computational cost in time.

Acknowledgements

I would like to thank an anonymous associate editor and two referees, whose constructive comments and suggestions greatly improved the presentation and flow of the paper. The research agency TUBITAK via Project # 111T767 and the European Commission under the Marie Curie International Outgoing Fellowship Programme via Project # 329370 titled PRinHDD supported this research.

References

1. Ripley BD. *Spatial Statistics*, 2nd ed. Wiley-Interscience: New York, 2004.
2. Diggle PJ. *Statistical Analysis of Spatial Point Patterns*, 2nd ed. Hodder Arnold Publishers: London, 2003.
3. Banerjee S, Dey DK. Editorial for the special issue on spatial statistics. *Statistical Methodology* 2012; **9**(1-2):115–116.
4. LeSage J, Banerjee S, Fischer MM, Congdon P. Spatial statistics: methods, models & computation. *Computational Statistics & Data Analysis* 2009; **53**(8):2781–2785.
5. Waller LA, Gotway CA. *Applied Spatial Statistics for Public Health Data*. Wiley-Interscience: NJ, 2004.
6. Lawson A, Denison D. *Spatial Cluster Modelling*. CRC Press: Baco Raton, Florida, 2002.
7. Rogerson PA. Statistical methods for the detection of spatial clustering. *Statistics in Medicine* 2006; **25**:811–823.
8. Gómez-Rubio V, Ferrándiz J, López A. Detecting clusters of diseases with R. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, Vienna, Austria, March 20–22, 2003. ISSN 1609-395X Kurt Hornik, Friedrich Leisch &amp; Achim Zeileis (eds.) Available from: http://www.ci.tuwien.ac.at/Conferences/DSC-2003/.
9. Potthoff RF, Whittinghill M. Testing for homogeinity: II. The Poisson distribution. *Acta Jutlandica* 1966; **53**:183–190.
10. Besag J, Newell J. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A* 1991; **154**:143–155.
11. Cuzick J, Edwards R. Spatial clustering for inhomogeneous populations (with discussion). *Journal of the Royal Statistical Society, Series B* 1990; **52**:73–104.
12. Centers for Disease Control and Prevention. Guidelines for investigating clusters of health events. Morbidity and mortality weekly report. *39(RR-11)*, 1-23 1990.
13. Tango T. *Scan Statistics, Statistics for Industry and Technology, Section: Detection of Disease Clustering*. John Wiley: Chichester, 2009.
14. Dixon PM. Nearest neighbor methods. In *Encyclopedia of Environmetrics*, El-Shaarawi AH, Piegorsch WW (eds). John Wiley & Sons Ltd.: NY, 2002; **3**: 1370–1383.
15. Pielou EC. Segregation and symmetry in two-species populations as studied by nearest-neighbor relationships. *Journal of Ecology* 1961; **49**(2):255–269.
16. Dixon PM. Testing spatial segregation using a nearest-neighbor contingency table. *Ecology* 1994; **75**(7):1940–1948.
17. Dixon PM. Nearest-neighbor contingency table analysis of spatial segregation for several species. *Ecoscience* 2002; **9**(2):142–151.
18. Ceyhan E. New tests of spatial segregation based on nearest neighbor contingency tables. *Scandinavian Journal of Statistics* 2010; **37**:147–165.
19. Moran PAP. The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B* 1948; **10**:243–251.
20. Geary RC. The contiguity ratio and statistical mapping. *The Incorporated Statistician* 1954; **5**:115–145.
21. Whittemore AS, Friend N, Byron W, Brown JR, Holly EA. A test to detect clusters of disease. *Biometrika* 1987; **74**:631–635.

22. Tango T. Comparison of general tests for spatial clustering. In *Disease Mapping and Risk Assessment for Public Health*, chapter 8. Wiley: New York, 1999; 111–117.
23. Openshaw S, Charlton M, Wymer C, Craft AW. A mark I geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems* 1987; **1**:335–358.
24. Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. *Statistics in Medicine* 1995; **14**:799–810.
25. Kulldorff M, Tango T, Park P. Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis* 2003; **42**:665–684.
26. Eshel G. *Spatiotemporal Data Analysis*. Princeton University Press: Princeton, NJ, 2012.
27. Tango T. The detection of disease clustering in time. *Biometrics* 1984; **40**(1):15–26.
28. Kryscio RJ, Lefèvre C. Measuring the severity of disease clustering using Tango's index. *Mathematical Biosciences* 1991; **107**(2):235–247.
29. Woillez M, Poulard JC, Rivoirard J, Petitgas P, Bez N. Indices for capturing spatial patterns and their evolution in time, with application to European hake (Merluccius merluccius) in the Bay of Biscay. *ICES Journal of Marine Science* 2007; **64**(3):537–550.
30. Li H, Reynolds JF. A new contagion index to quantify spatial patterns of landscapes. *Landscape Ecology* 1993; **8**(3):155–162.
31. Ceyhan E. Segregation indices for disease clustering. arXiv:1310.0364 [stat.ME]. *Technical Report # KU-EC-13-1*, Koç University, Istanbul, Turkey, 2013.
32. Kulldorff M. Tests for spatial randomness adjusted for an inhomogeneity: a general framework. *Journal of the American Statistical Association* 2006; **101**(475):1289–1305.
33. Song C, Kulldorff M. Power evaluation of disease clustering tests. *International Journal of Health Geographics* 2003; **2**(1):9–16.
34. Ceyhan E. New cell-specific and overall tests of spatial interaction based on nearest neighbor contingency tables. arXiv:1206.1850v1 [stat.ME]. *Technical Report # KU-EC-12-1*, Koç University, Istanbul, Turkey, 2012.
35. Searle SR. *Matrix Algebra Useful for Statistics*. Wiley-Intersciences: New York, 2006.
36. Baddeley AJ, Turner R. spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software* 2005; **12**(6):1–42.
37. Goreaud F, Pélissier R. Avoiding misinterpretation of biotic interactions with the intertype K_{12}-function: population independence vs. random labelling hypotheses. *Journal of Vegetation Science* 2003; **14**(5):681–692.