Communications in Statistics—Theory and Methods, 40: 1363–1395, 2011

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610921003597211

Spatial Clustering Tests Based on the Domination Number of a New Random Digraph Family

ELVAN CEYHAN

Department of Mathematics, Koç University, Istanbul, Turkey

We use the domination number of a parametrized random digraph family called proportional-edge proximity catch digraphs (PCDs) for testing multivariate spatial point patterns. This digraph family is based on relative positions of data points from various classes. We extend the results on the distribution of the domination number of proportional-edge PCDs, and use the domination number as a statistic for testing segregation and association against complete spatial randomness. We demonstrate that the domination number of the PCD has binomial distribution when size of one class is fixed while the size of the other (whose points constitute the vertices of the digraph) tends to infinity and has asymptotic normality when sizes of both classes tend to infinity. We evaluate the finite sample performance of the test by Monte Carlo simulations and prove the consistency of the test under the alternatives. We find the optimal parameters for testing each of the segregation and association alternatives. Furthermore, the methodology discussed in this article is valid for data in higher dimensions also.

Keywords Association; Complete spatial randomness; Consistency; Delaunay triangulation; Proximity catch digraph; Proximity map; Segregation.

Mathematics Subject Classification Primary 62H30; Secondary 60D05, 05C80, 05C20, 62G10, 62G20, 62H11.

1. Introduction

In statistical literature, the problem of clustering received considerable attention. The spatial interaction between two or more classes has important implications especially for plant species; see, for example, Dixon (1994, 2002a), Stoyan and Penttinen (2000), and Perry et al. (2006). Recently, a new clustering test based on the relative allocation of points from two or more classes has been developed. The method is based on a graph-theoretic approach and is used to test the spatial pattern of complete spatial randomness (CSR) against segregation or association. Rather than the pattern of points from one-class with respect to the ground, the patterns of points from one class with respect to points from other classes are investigated. CSR is roughly defined as the lack of spatial interaction between the points in a given

Received January 29, 2009; Accepted December 28, 2009
Address correspondence to Elvan Ceyhan, Department of Mathematics, Koç University, 34450 Sarıyer, Istanbul, Turkey; E-mail: elceyhan@ku.edu.tr

study area. *Segregation* is the pattern in which points of one class tend to cluster together, i.e., form one-class clumps. On the other hand, *association* is the pattern in which the points of one class tend to occur more frequently around points from the other class.

Many methods to analyze spatial clustering have been proposed in the literature (Kulldorff, 2006). These include Ripley's K- or L-functions (Ripley, 2004), comparison of NN distances (Cuzick and Edwards, 1990), and analysis of nearest neighbor contingency tables (NNCTs) which are constructed using the NN frequencies of classes (Dixon, 1994, 2002a; Pielou, 1961). The tests (i.e., inference) based on Ripley's K- or L-functions are only appropriate when the null pattern can be assumed to be the CSR independence pattern, but not if the null pattern is the random labeling (RL) of points from an inhomogeneous Poisson pattern (Kulldorff, 2006). But, there are also variants of $K(t)$ that explicitly correct for inhomogeneity (see Baddeley et al., 2000). Cuzick and Edward's k-NN tests are designed for testing bivariate spatial interaction and mostly used for spatial clustering of cases or controls in epidemiology. Diggle's D-function is a modified version of Ripley's K-function (Diggle, 2003) and is appropriate for the case in which the null pattern is the RL of points where the points are a realization from an arbitrary point pattern. Ripley's and Diggle's functions are designed to analyze univariate or bivariate spatial interaction at various scales (i.e., inter-point distances).

In recent years, the use of mathematical graphs has also gained popularity in spatial analysis (Roberts et al., 2000), providing a way to move beyond Euclidean metrics for spatial analysis. Although only recently introduced to landscape ecology, graph theory is well suited to ecological applications concerned with connectivity or movement (Minor and Urban, 2007). Conventional graphs do not explicitly maintain geographic reference, reducing utility of other geo-spatial information. Fall et al. (2007) introduced spatial graphs that integrate a geometric reference system that ties patches and paths to specific spatial locations and spatial dimensions thereby preserving the relevant spatial information. However, after a graph is constructed using spatial data, usually the scale is lost (see for instance, Su et al., 2007). Many concepts in spatial ecology depend on the idea of spatial adjacency which requires information on the close vicinity of an object. Graph theory conveniently can be used to express and communicate adjacency information allowing one to compute meaningful quantities related to spatial point pattern. Adding vertex and edge properties to graphs extends the problem domain to network modeling (Keitt, 2007). Wu and Murray (2008) proposed a new measure based on graph theory and spatial interaction, which reflects intra-patch and inter-patch relationships by quantifying contiguity within patches and potential contiguity among patches. Friedman and Rafsky (1983) also proposed a graph-theoretic method to measure multivariate association, but their method is not designed to analyze spatial interaction between two or more classes; instead it is an extension of generalized correlation coefficient (such as Spearman's ρ or Kendall's τ) to measure multivariate (possibly nonlinear) correlation.

Priebe et al. (2001) introduced a data random digraph called *class cover catch digraph* (CCCD) in $\mathbb{R}$ and extended it to multiple dimensions. DeVinney et al. (2002), Marchette and Priebe (2003), and Priebe et al. (2003a,b) demonstrated relatively good performance of CCCDs in classification. Their methods involve *data reduction* (*condensing*) by using approximate minimum dominating sets as *prototype sets* (since finding the exact minimum dominating set is an NP-hard

problem in general—e.g., for CCCD in multiple dimensions—(see DeVinney and Priebe, 2006). For the domination number of CCCDs for one-dimensional data, a SLLN result is proved in DeVinney and Wierman (2003), and this result is extended by Wierman and Xiang (2008); furthermore, a CLT is also proved by Xiang and Wierman (2009). The asymptotic distribution of the domination number of CCCDs for non uniform data in $\mathbb{R}$ is also calculated in a rather general setting (Ceyhan, 2008a). Although intuitively appealing and easy to extend to higher dimensions, the distribution of the domination number of CCCDs is not analytically tractable for $d > 1$. As alternatives to CCCD, Ceyhan and Priebe (2003) introduced an (unparametrized) type of PCDs called *central similarity PCDs*; Ceyhan and Priebe (2005) also introduced another parametrized family of PCDs called *proportional-edge PCDs* and used the domination number of this PCD with a fixed parameter for testing spatial patterns. The relative (arc) density of the central similarity and proportional-edge PCDs are also used for testing the spatial patterns in Ceyhan et al. (2006, 2007). Ceyhan and Priebe (2007) derived the asymptotic distribution of the domination number of proportional-edge PCDs for uniform data. An extensive treatment of the PCDs based on Delaunay tessellations is available in Ceyhan (2005).

In this article, we investigate the use of the domination number of proportional-edge PCDs, whose asymptotic distribution was computed in Ceyhan and Priebe (2007) for testing spatial patterns of segregation and association. Furthermore, we extend this result for the whole range of the expansion parameter in a more general setting. In addition to the mathematical tractability and applicability to testing spatial patterns and classification, this new family of PCDs is more flexible as it allows choosing an optimal parameter for testing against various types of spatial point patterns.

We define proximity maps and the associated PCDs in Sec. 2. We present the asymptotic distribution of the domination number for uniform data in one triangle and in multiple triangles in Sec. 3. In Sec. 4, we describe the alternative patterns of segregation and association. In Sec. 5, we present the Monte Carlo simulation analysis to assess the empirical size and power performance. In Sec. 6, we suggest an adjustment for data points from the class of interest which are outside the convex hull of data from the other class. In Sec. 7, we provide an example data set. In Sec. 8, we describe the extension of proportional-edge PCDs to higher dimensions. We also provide the guidelines in using this test in Sec. 9.

2. Proximity Maps and the Associated PCDs

Our PCDs are based on the proximity maps which are defined in a fairly general setting. Let $(\Omega, \mathcal{M})$ be a measurable space. The *proximity map* $N(\cdot)$ is defined as $N : \Omega \to 2^{\Omega}$, where 2^{Ω} is the power set of Ω. The *proximity region* associated with $x \in \Omega$, denoted $N(x)$, is the image of $x \in \Omega$ under $N(\cdot)$. The points in $N(x)$ are thought of as being "closer" to $x \in \Omega$ than are the points in $\Omega \backslash N(x)$. Hence, the term "proximity" in the name *proximity catch digraph*. The Γ_1-region $\Gamma_1(\cdot) = \Gamma_1(\cdot, \mathcal{Y}) : \Omega \to 2^{\Omega}$ associates the region $\Gamma_1(x) := \{z \in \Omega : x \in N_{\mathcal{Y}}(z)\}$ with each point $x \in \Omega$. Proximity maps are the building blocks of the *proximity graphs* of Toussaint (1980); an extensive survey on proximity maps and graphs is available in Jaromczyk and Toussaint (1992).

The *proximity catch digraph* D has the vertex set $\mathcal{V} = \{p_1, p_2, \ldots, p_n\}$; and the arc set $\mathcal{A}$ is defined by $(p_i, p_j) \in \mathcal{A}$ iff $p_j \in N(p_i)$ for $i \neq j$. Notice that the proximity

catch digraph D depends on the *proximity* map $N(\cdot)$ and if $p_j \in N(p_i)$, then we call the region $N(p_i)$ (and the point p_i) *catches* point p_j. Hence, the term "catch" in the name proximity catch digraph. If arcs of the form (p_i, p_i) (i.e., loops) were allowed, D would have been called a *pseudodigraph* according to some authors (see, e.g., Chartrand and Lesniak, 1996).

In a digraph $D = (\mathcal{V}, \mathcal{A})$, a vertex $v \in \mathcal{V}$ *dominates* itself and all vertices of the form $\{u : (v, u) \in \mathcal{A}\}$. A *dominating set* S_D for the digraph D is a subset of $\mathcal{V}$ such that each vertex $v \in \mathcal{V}$ is dominated by a vertex in S_D. A *minimum dominating set* S_D^* is a dominating set of minimum cardinality and the *domination number* $\gamma(D)$ is defined as $\gamma(D) := |S_D^*|$ where $|\cdot|$ denotes the set cardinality functional. See Chartrand and Lesniak (1996) and West (2001) for more on graphs and digraphs. If a minimum dominating set is of size one, we call it a *dominating point*. Note that for $|\mathcal{V}| = n > 0$, $1 \leq \gamma(D) \leq n$, since $\mathcal{V}$ itself is always a dominating set.

We construct the proximity regions using two data sets $\mathcal{X}_n$ and $\mathcal{Y}_m$ of sizes n and m from classes $\mathcal{X}$ and $\mathcal{Y}$, respectively. Given $\mathcal{Y}_m \subseteq \Omega$, the *proximity map* $N_{\mathcal{Y}}(\cdot) : \Omega \to 2^{\Omega}$ associates a *proximity region* $N_{\mathcal{Y}}(x) \subseteq \Omega$ with each point $x \in \Omega$. The region $N_{\mathcal{Y}}(x)$ is defined in terms of the distance between x and $\mathcal{Y}_m$. More specifically, our proportional-edge proximity maps will be based on the relative position of points from $\mathcal{X}_n$ with respect to the Delaunay tessellation of $\mathcal{Y}_m$. In this article, a triangle refers to the closed region bounded by its edges; see Fig. 1 for an example with $n = 200$ $\mathcal{X}$ points iid $\mathcal{U}\big((0,1) \times (0,1)\big)$, the uniform distribution on the unit square and the Delaunay triangulation (which yields 13 triangles) is based on $m = 10$ $\mathcal{Y}$ points which are also iid $\mathcal{U}\big((0,1) \times (0,1)\big)$ and 77 of these $\mathcal{X}$ points are inside the convex hull of $\mathcal{Y}$ points.

If $\mathcal{X}_n = \{X_1, X_2, \ldots, X_n\}$ is a set of Ω-valued random variables then $N_{\mathcal{Y}}(X_i)$ and $\Gamma_1(X_i)$ are random sets. If X_i are iid then so are the random sets $N_{\mathcal{Y}}(X_i)$. The same holds for $\Gamma_1(X_i)$. We define the *data-random proximity catch digraph* D—associated

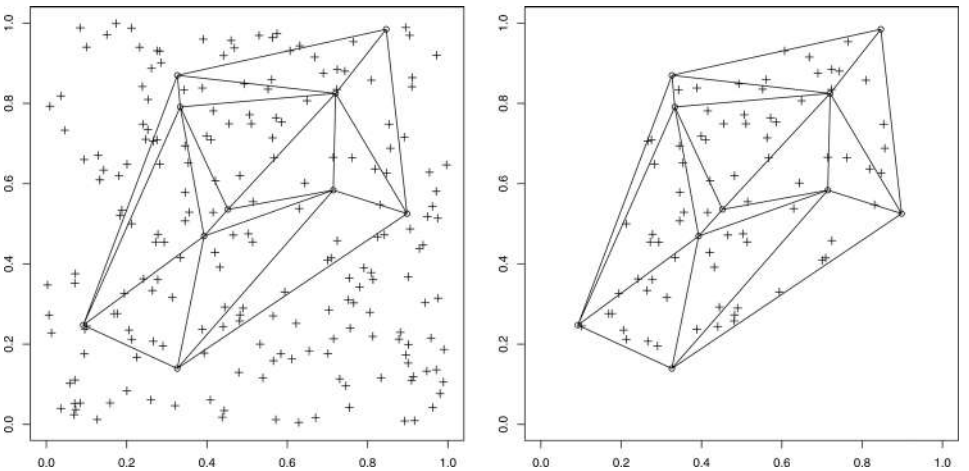

Figure 1. Plotted left is a realization of 200 $\mathcal{X}$ points (pluses, +) and the Delaunay triangulation based on 10 $\mathcal{Y}$ points (circles, ∘). Plotted right is the 77 $\mathcal{X}$ points which are in the convex hull of $\mathcal{Y}$ points. Both $\mathcal{X}_n$ and $\mathcal{Y}_m$ are random samples from $\mathcal{U}\big((0,1) \times (0,1)\big)$, the uniform distribution on the unit square.

with $N_{\mathcal{Y}}(\cdot)$—with vertex set $\mathcal{X}_n = \{X_1, X_2, \ldots, X_n\}$ and arc set $\mathcal{A}$ by

$$(X_i, X_j) \in \mathcal{A} \iff X_j \in N_{\mathcal{Y}}(X_i).$$

Since this relationship is not symmetric, a digraph is used rather than a graph. The random digraph D depends on the (joint) distribution of X_i and on the map $N_{\mathcal{Y}}(\cdot)$. For $\mathcal{X}_n = \{X_1, X_2, \ldots, X_n\}$, a set of iid random variables from F, the domination number of the associated data-random PCD based on the proximity map $N(\cdot)$, denoted $\gamma(\mathcal{X}_n, N)$, is the minimum number of point(s) that dominate all points in $\mathcal{X}_n$. The random variable $\gamma(\mathcal{X}_n, N)$ depends explicitly on $\mathcal{X}_n$ and $N(\cdot)$ and implicitly on F. Furthermore, in general, the distribution, hence the expectation $\mathbf{E}[\gamma(\mathcal{X}_n, N)]$, depends on n, F, and N; $1 \leq \mathbf{E}[\gamma(\mathcal{X}_n, N)] \leq n$. In general, the variance of $\gamma(\mathcal{X}_n, N)$ satisfies, $1 \leq \mathbf{Var}[\gamma(\mathcal{X}_n, N)] \leq n^2/4$. For example, the CCCD of Priebe et al. (2001) can be viewed as an example of PCDs.

2.1. *The Proportional-Edge Proximity Maps*

Note that in $\mathbb{R}$ the CCCDs are based on the intervals whose end points are from class $\mathcal{Y}$. This interval partitioning can be viewed as the *Delaunay tessellation* of $\mathbb{R}$ based on $\mathcal{Y}_m$. So in higher dimensions, we use the Delaunay triangulation based on $\mathcal{Y}_m$ to partition the space.

Let $\mathcal{Y}_m = \{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_m\}$ be m points in general position in $\mathbb{R}^d$ and T_i be the ith Delaunay cell for $i = 1, 2, \ldots, J_m$, where J_m is the number of Delaunay cells. Let $\mathcal{X}_n$ be a set of iid random variables from distribution F in $\mathbb{R}^d$ with support $\mathcal{S}(F) \subseteq \mathcal{C}_H(\mathcal{Y}_m)$ where $\mathcal{C}_H(\mathcal{Y}_m)$ stands for the convex hull of $\mathcal{Y}_m$. In particular, for illustrative purposes, we focus on $\mathbb{R}^2$ where a Delaunay tessellation is a *triangulation*, provided that no more than three points in $\mathcal{Y}_m$ are cocircular (i.e., lie on the same circle). Furthermore, for simplicity, let $\mathcal{Y}_3 = \{\mathrm{y}_1, \mathrm{y}_2, \mathrm{y}_3\}$ be three non collinear points in $\mathbb{R}^2$ and $T(\mathcal{Y}_3) = T(\mathrm{y}_1, \mathrm{y}_2, \mathrm{y}_3)$ be the triangle with vertices $\mathcal{Y}_3$. Let $\mathcal{X}_n$ be a set of iid random variables from F with support $\mathcal{S}(F) \subseteq T(\mathcal{Y}_3)$. If $F = \mathcal{U}(T(\mathcal{Y}_3))$, a composition of translation, rotation, reflections, and scaling will take any given triangle $T(\mathcal{Y}_3)$ to the basic triangle $T_b = T((0, 0), (1, 0), (c_1, c_2))$ with $0 < c_1 \leq 1/2$, $c_2 > 0$, and $(1 - c_1)^2 + c_2^2 \leq 1$, preserving uniformity. That is, if $X \sim \mathcal{U}(T(\mathcal{Y}_3))$ is transformed in the same manner to, say X', then we have $X' \sim \mathcal{U}(T_b)$. In fact, this will hold for any distribution F up to scale.

For $r \in [1, \infty]$, define $N_{PE}^r(\cdot, M) := N(\cdot, M; r, \mathcal{Y}_3)$ to be the (*parametrized*) *proportional-edge proximity map* with M-vertex regions as follows (see also Fig. 2 with $M = M_C$ and $r = 2$). For $x \in T(\mathcal{Y}_3) \backslash \mathcal{Y}_3$, let $v(x) \in \mathcal{Y}_3$ be the vertex whose region contains x; i.e., $x \in R_M(v(x))$. In this article, *M-vertex regions* are constructed by the lines joining any point $M \in \mathbb{R}^2 \backslash \mathcal{Y}_3$ to a point on each of the edges of $T(\mathcal{Y}_3)$. Preferably, M is selected to be in the interior of the triangle $T(\mathcal{Y}_3)^o$. For such an M, the corresponding vertex regions can be defined using the line segment joining M to e_j, which lies on the line joining y_j to M. With M_C, the lines joining M and $\mathcal{Y}_3$ are the *median lines*, that cross edges at M_j for $j = 1, 2, 3$. *M-vertex regions*, among many possibilities, can also be defined by the orthogonal projections from M to the edges; see Ceyhan (2005) for a more general definition. The vertex regions in Fig. 2 are center of mass vertex regions (i.e., M_C-vertex regions). If x falls on the boundary of two M-vertex regions, we assign $v(x)$ arbitrarily. Let $e(x)$ be the edge of $T(\mathcal{Y}_3)$ opposite of $v(x)$. Let $\ell(v(x), x)$ be the line parallel to $e(x)$ and passes through x. Let

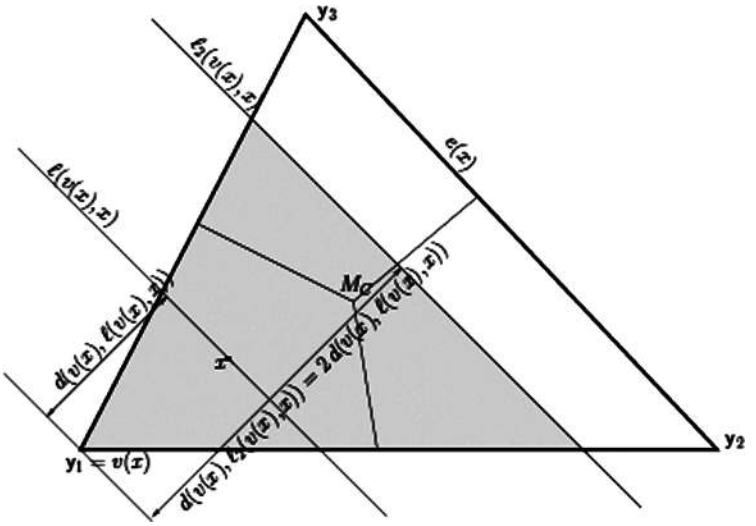

Figure 2. Construction of proportional-edge proximity region, $N_{PE}^{r=2}(x, M_C)$ (shaded region) for an x in the M_C-vertex region for y_1, $R_{M_C}(y_1)$.

$d(v(x), \ell(v(x), x))$ be the Euclidean (perpendicular) distance from $v(x)$ to $\ell(v(x), x)$. For $r \in [1, \infty)$, let $\ell_r(v(x), x)$ be the line parallel to $e(x)$ such that

$$d(v(x), \ell_r(v(x), x)) = r\, d(v(x), \ell(v(x), x)) \quad \text{and}$$
$$d(\ell(v(x), x), \ell_r(v(x), x)) < d(v(x), \ell_r(v(x), x)).$$

Let $T_r(x)$ be the triangle similar to and with the same orientation as $T(\mathcal{Y}_3)$ having $v(x)$ as a vertex and $\ell_r(v(x), x)$ as the opposite edge. Then the *proportional-edge proximity region* $N_{PE}^r(x, M)$ is defined to be $T_r(x) \cap T(\mathcal{Y}_3)$. Notice that $\ell(v(x), x)$ divides the edges of $T_r(x)$ (other than the one lies on $\ell_r(v(x), x)$) proportionally with the factor r. Hence, the name proportional-edge proximity region.

Notice that $r \geq 1$ implies $x \in N_{PE}^r(x, M)$ for all $x \in T(\mathcal{Y}_3)$. Furthermore, $\lim_{r\to\infty} N_{PE}^r(x, M) = T(\mathcal{Y}_3)$ for all $x \in T(\mathcal{Y}_3)\backslash\mathcal{Y}_3$, so we define $N_{PE}^{\infty}(x, M) = T(\mathcal{Y}_3)$ for all such x. For $x \in \mathcal{Y}_3$, we define $N_{PE}^r(x, M) = \{x\}$ for all $r \in [1, \infty]$.

The proportional-edge PCD has vertices $\mathcal{X}_n$ and arcs (x_i, x_j) iff $x_j \in N_{PE}^r(x_i, M)$; see Fig. 3 for a realization of $\mathcal{X}_n$ with $n = 7$ in one triangle (i.e., $m = 3$). Let $\gamma_n(r, M) := \gamma(\mathcal{X}_n, N_{PE}^r(\cdot, M))$. Then for $r = 3/2$, the number of arcs is 12 and the domination number $\gamma_n(r = 3/2) = 1$; and for $r = 5/4$, the number of arcs is 9 and $\gamma_n(r = 5/4) = 3$. By construction, note that as x gets closer to M (or equivalently further away from the vertices in vertex regions), $N_{PE}^r(x, M)$ increases in area, hence it is more likely for the outdegree of x to increase. So, if more $\mathcal{X}$ points are around the center M, then it is more likely for the domination number $\gamma_n(r, M)$ to decrease; on the other hand, if more $\mathcal{X}$ points are around the vertices $\mathcal{Y}_3$, then the regions get smaller, hence it is more likely for the outdegree for such points to be smaller, thereby implying $\gamma_n(r, M)$ to increase. We exploit this probabilistic behavior of $\gamma_n(r, M)$ in testing spatial patterns of segregation and association.

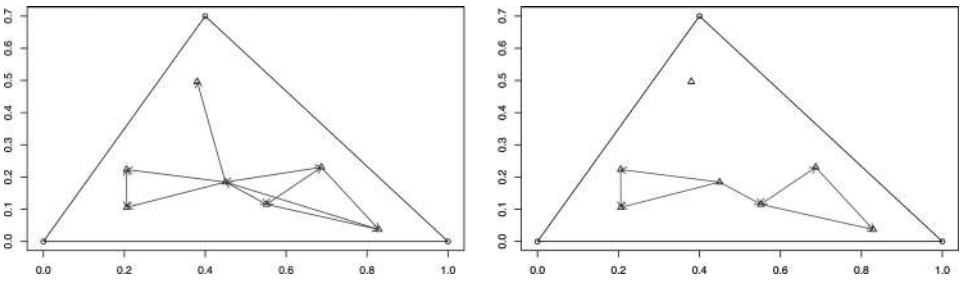

Figure 3. A realization of 7 $\mathscr{X}$ points generated iid $\mathscr{U}(T(\mathscr{Y}_3))$, the uniform distribution on $T(\mathscr{Y}_3)$ and the corresponding arcs of proportional-edge PCD with $M = M_C$ for $r = 3/2$ (left) and $r = 5/4$ (right).

Note also that, $N_{PE}^r(x, M)$ can be viewed as a *homothetic transformation* (*enlargement*) with $r \geq 1$ applied on a translation of the region $N_{PE}^{r=1}(x, M)$. Furthermore, this transformation is also an *affine similarity transformation*.

2.2. *Some Auxiliary Tools Associated with Proportional-Edge PCDs*

First, notice that $N_{PE}^r(x, M)$ is similar to $T(\mathscr{Y}_3)$ with the similarity ratio being equal to

$$\min\big(d\big(v(x), e(x)\big), rd\big(v(x), \ell(v(x), x)\big)\big)\big/d(v(x), e(x)).$$

To define the Γ_1-*region*, let $\xi_i(x)$ be the line such that $\xi_i(x) \cap T(\mathscr{Y}_3) \neq \emptyset$ and $rd(\mathrm{y}_i, \xi_i(x)) = d(\mathrm{y}_i, \ell(\mathrm{y}_i, x))$ for $i = 1, 2, 3$; see also Fig. 4. Then $\Gamma_1^r(x, M) = \bigcup_{i=1}^3 \big(\Gamma_1^r(x, M) \cap R_M(\mathrm{y}_i)\big)$ where $\Gamma_1^r(x, M) \cap R_M(\mathrm{y}_i) = \{z \in R_M(\mathrm{y}_i) : d(\mathrm{y}_i, \ell(\mathrm{y}_i, z)) \geq d(\mathrm{y}_i, \xi_i(x)\}$, for $i = 1, 2, 3$. Notice that $r \geq 1$ implies $x \in \Gamma_1^r(x, M)$. Furthermore, $\lim_{r\to\infty} \Gamma_1^r(x, M) = T(\mathscr{Y}_3)$ for all $x \in T(\mathscr{Y}_3)\backslash\mathscr{Y}_3$, and so we define $\Gamma_1^\infty(x, M) = T(\mathscr{Y}_3)$ for all such x.

For $X_i \overset{iid}{\sim} F$, with the additional assumption that the non degenerate two-dimensional probability density function f exists with support in $T(\mathscr{Y}_3)$, implies that the special cases in the construction of N_{PE}^r—X falls on the boundary of two vertex regions or on the vertices of $T(\mathscr{Y}_3)$—occur with probability zero. Note that for such an F, $N_{PE}^r(x, M)$ is a triangle a.s. and $\Gamma_1^r(x, M)$ is a star-shaped (not necessarily convex) polygon. Let $X_e := \mathrm{argmin}_{X\in\mathscr{X}_n} d(X, e)$ be the (closest) *edge extremum* for edge e (i.e., closest point among $\mathscr{X}_n$ to edge e). Then it is easily seen that $\Gamma_1^r(\mathscr{X}_n, M) = \bigcap_{i=1}^3 \Gamma_1^r(X_{e_i}, M)$, where e_i is the edge opposite vertex y_i, for $i = 1, 2, 3$. So $\Gamma_1^r(\mathscr{X}_n, M) \cap R_M(\mathrm{y}_i) = \{z \in R_M(\mathrm{y}_i) : d(\mathrm{y}_i, \ell(\mathrm{y}_i, z) \geq d(\mathrm{y}_i, \xi_i(X_{e_i}))\}$, for $i = 1, 2, 3$.

Let the closest edge extrema (if exist) be $X_{[i,1]} := \mathrm{argmin}_{X\in\mathscr{X}_n\cap R_M(\mathrm{y}_i)} d(X, e_i)$. Then $\gamma_n(r, M) \leq 3$ with probability 1, since $\mathscr{X}_n \cap R_M(\mathrm{y}_i) \subset N_{PE}^r(X_{[i,1]}, M)$ for each of $i = 1, 2, 3$. Thus,

$$1 \leq \mathbf{E}[\gamma_n(r, M)] \leq 3 \quad \text{and} \quad 0 \leq \mathbf{Var}[\gamma_n(r, M)] \leq 9/4.$$

In $T(\mathscr{Y}_3)$, drawing the lines $q_i(r, x)$ such that $d(\mathrm{y}_i, e_i) = r\, d(\mathrm{y}_i, q_i(r, x))$ for $i \in \{1, 2, 3\}$ yields another triangle, denoted as $\mathscr{T}_r$, for $r < 3/2$; see Fig. 5 for $\mathscr{T}_r$ with

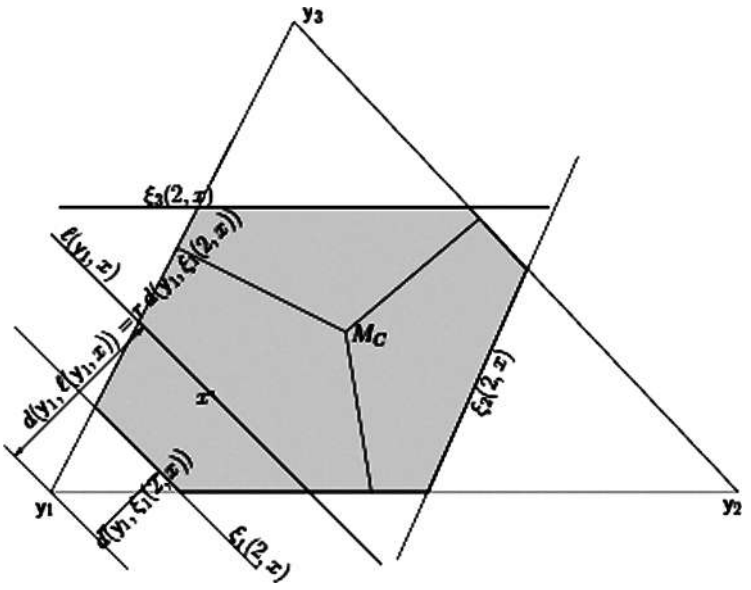

Figure 4. Construction of the Γ_1-region, $\Gamma_1^2(x, M_C)$ (shaded region).

$r = \sqrt{2}$. The functional form of $\mathcal{T}_r$ in the standard equilateral triangle is given by

$$\mathcal{T}_r = T\left(\left(\frac{3(r-1)}{2r}, \frac{\sqrt{3}(r-1)}{2r}\right), \left(\frac{3-r}{2r}, \frac{\sqrt{3}(r-1)}{2r}\right), \left(\frac{1}{2}, \frac{\sqrt{3}(2-r)}{r}\right)\right). \quad (1)$$

The triangle $\mathcal{T}_r$ given in Eq. (1) plays an important role in the distribution of the domination number of the proportional-edge PCDs.

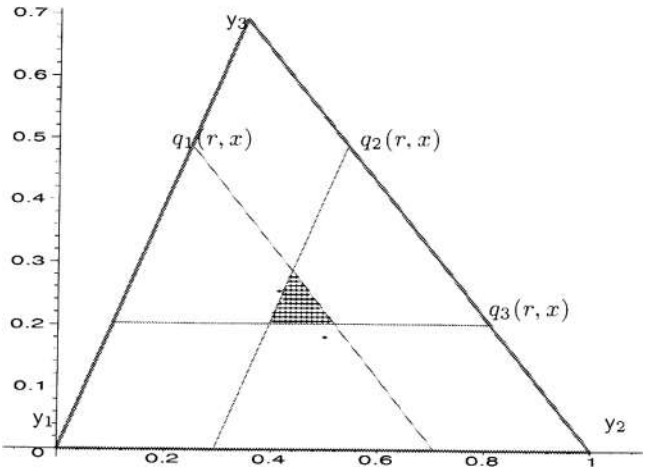

Figure 5. The triangle $\mathcal{T}_r$ with $r = \sqrt{2}$ (the hatched region).

3. The Asymptotic Distribution of Domination Number for Uniform Data

3.1. *The One-Triangle Case*

For simplicity, we consider $\mathscr{X}$ points iid uniform in one triangle only. The null hypothesis we consider is a type of *complete spatial randomness*; that is,

$$H_o : X_i \overset{iid}{\sim} \mathscr{U}(T(\mathscr{Y}_3)) \quad \text{for } i = 1, 2, \ldots, n,$$

where $\mathscr{U}(T(\mathscr{Y}_3))$ is the uniform distribution on $T(\mathscr{Y}_3)$. If it is desired to have the sample size be a random variable, we may consider a spatial Poisson point process on $T(\mathscr{Y}_3)$ as our null hypothesis. As before, let $\gamma_n(r, M)$ stand for the domination number of the PCD based on N_{PE}^r with $\mathscr{X}_n$, a set of iid random variables from $\mathscr{U}(T(\mathscr{Y}_3))$, with M-vertex regions.

We present a "geometry invariance" result for $N_{PE}^r(\cdot, M)$ where *M-vertex regions* are constructed using the line segment joining M to edge e_i on the line joining y_i to M, rather than the orthogonal projections from M to the edges. This invariance property will simplify the notation in our subsequent analysis by allowing us to consider the special case of the (standard) equilateral triangle.

Theorem 3.1 (Geometry Invariance Property)**.** *Suppose $\mathscr{X}_n$ is a set of iid random variables from $\mathscr{U}(T(\mathscr{Y}_3))$. Then for any $r \in [1, \infty]$ the distribution of $\gamma_n(r, M)$ is independent of $\mathscr{Y}_3$ and hence the geometry of $T(\mathscr{Y}_3)$.*

Proof. See Ceyhan and Priebe (2007) for the proof.

Note that geometry invariance of $\gamma_n(r = \infty, M)$ follows trivially for all $\mathscr{X}_n$ from any F with support in $T(\mathscr{Y}_3)\backslash\mathscr{Y}_3$, since for $r = \infty$, we have $\gamma_n(r = \infty, M) = 1$ a.s. Based on Theorem 3.1 we may assume that $T(\mathscr{Y}_3)$ is a standard equilateral triangle with $\mathscr{Y}_3 = \{(0, 0), (1, 0), \left(1/2, \sqrt{3}/2\right)\}$ for $N_{PE}^r(\cdot, M)$ with M-vertex regions.

Remark 3.1. Notice that we proved the geometry invariance property for $N_{PE}^r(\cdot)$ where M-vertex regions are defined with the lines joining $\mathscr{Y}_3$ to M. If we had used the orthogonal projections from M to the edges, the vertex regions (hence N_{PE}^r) would depend on the geometry of the triangle. That is, the orthogonal projections from M to the edges will not be mapped to the orthogonal projections in the standard equilateral triangle. Hence, the exact and asymptotic distribution of $\gamma_n(r, M)$ will depend on c_1, c_2 of T_b, so one needs to do the calculations for each possible combination of c_1, c_2.

The domination number $\gamma_n(r, M)$ of the PCD has the following asymptotic distribution (Ceyhan and Priebe, 2007). As $n \to \infty$,

$$\gamma_n(r, M) \xrightarrow{\mathscr{L}} \begin{cases} 2 + \text{BER}(1 - p_r) & \text{for } r \in [1, 3/2) \text{ and } M \in \{t_1(r), t_2(r), t_3(r)\}, \\ 1 & \text{for } r > 3/2 \text{ and } M \in T(\mathscr{Y}_3)^o, \\ 3 & \text{for } r \in [1, 3/2) \text{ and } M \in \mathscr{T}_r\backslash\{t_1(r), t_2(r), t_3(r)\}, \end{cases} \tag{2}$$

where $\xrightarrow{\mathscr{L}}$ stands for "convergence in law" and BER(p) stands for Bernoulli distribution with probability of success p, $\mathcal{T}_r$ and $t_i(r)$ are defined in Eq. (1), and for $r \in [1, 3/2)$ and $M \in \{t_1(r), t_2(r), t_3(r)\}$,

$$p_r = \int_0^\infty \int_0^\infty \frac{64r^2}{9(r-1)^2} w_1 w_3 \exp\left(\frac{4r}{3(r-1)}(w_1^2 + w_3^2 + 2r(r-1)w_1 w_3)\right) dw_3 w_1, \quad (3)$$

and for $r = 3/2$ and $M = M_C = (1/2, \sqrt{3}/6)$, $p_r \approx 0.7413$, which is not computed as in Eq. (3); for its computation, see Ceyhan and Priebe (2005). For example, for $r = 5/4$ and $M \in \{t_1(r) = \left(3/10, \sqrt{3}/10\right), t_2(r) = \left(7/10, \sqrt{3}/10\right), t_3(r) = \left(1/2, 3\sqrt{3}/5\right)\}$, $p_r \approx 0.6514$; see Fig. 6 for the plot of the numerically computed (by numerical integration) values of p_r as a function of r according to Eq. (3). Notice that in the non degenerate case in (2), $\mathbf{E}[\gamma_n(r, M)] = 3 - p_r$ and $\mathbf{Var}[\gamma_n(r, M)] = p_r(1 - p_r)$.

We also estimate the distribution of $\gamma_n(r, M)$ for various values of n, r, and M using Monte Carlo simulations. At each Monte Carlo replication, we generate n points iid $\mathcal{U}(T(\mathcal{Y}_3))$ and compute the value of $\gamma_n(r, M)$. The frequencies of $\gamma_n(r, M) = k$ out of $N = 1{,}000$ Monte Carlo replicates are presented in Tables 1

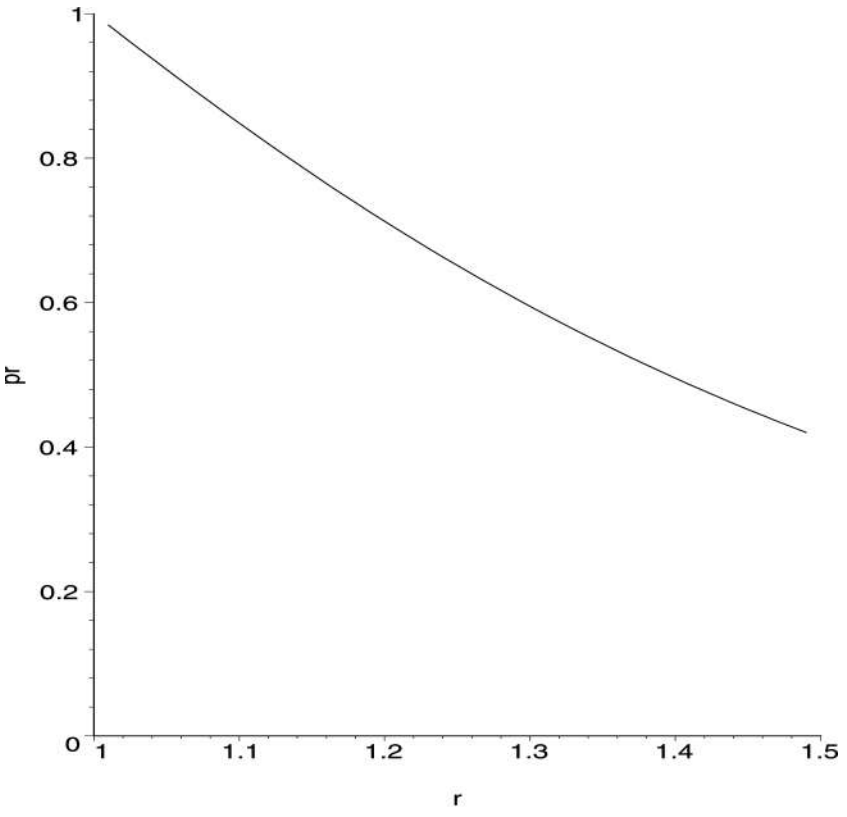

Figure 6. Plotted is the probability $p_r = \lim_{n\to\infty} P(\gamma_n(r, M) = 2)$ given in Eq. (3) as a function of r for $r \in [1, 3/2)$ and $M \in \{t_1(r), t_2(r), t_3(r)\}$.

Table 1
The number of $\gamma_n(r, M) = k$ out of $N = 1{,}000$ Monte Carlo replicates with $M = M_C$ and $r = 2$ (left) and $r = 5/4$ (right). Here, "$r = 2$ and $M = M_C$" is an example of the case "$r > 3/2$ and $M \in T(\mathcal{Y}_3)^o$", and "$r = 5/4$ and $M = M_C$" is an example of the case "$r \in [1, 3/2)$ and $M \in \mathcal{T}_r \backslash \{t_1(r), t_2(r), t_3(r)\}$"

$r = 2$ and $M = M_C$						$r = 5/4$ and $M = M_C$					
$k\backslash n$	10	20	30	50	100	$k\backslash n$	10	20	30	50	100
1	961	1000	1000	1000	1000	1	9	0	0	0	0
2	34	0	0	0	0	2	293	110	30	8	0
3	5	0	0	0	0	3	698	890	970	992	1000

and 2. Notice that as the sample size n increases, the values on these tables get closer and closer to the expected values under their asymptotic distribution.

Theorem 3.2. *Let* $\gamma_n(r, M) = \gamma(\mathcal{X}_n; \mathcal{U}(T(\mathcal{Y}_3)), N_{PE}^r, M)$. *Then* $r_1 < r_2$ *implies that* $\gamma_n(r_2, M) <^{ST} \gamma_n(r_1, M)$ *where* $<^{ST}$ *stands for "stochastically smaller than".*

Proof. Suppose $r_1 < r_2$. Then $P(\gamma_n(r_2, M) = 1) > P(\gamma_n(r_1, M) = 1)$ and $P(\gamma_n(r_2, M) = 2) > P(\gamma_n(r_1, M) = 2)$ and $P(\gamma_n(r_2, M) = 3) < P(\gamma_n(r_1, M) = 3)$. Hence, the desired result follows.

Table 2
The number of $\gamma_n(r, M) = k$ out of $N = 1{,}000$ Monte Carlo replicates with $r = 5/4$, $M = (3/5, \sqrt{3}/10)$ (top), and $M = (7/10, \sqrt{3}/10)$ (middle), and with $r = 3/2$ and $M = M_C$ (bottom). Here, "$r = 5/4$ and $M = (3/5, \sqrt{3}/10)$" is an example of the case "$r \in [1, 3/2)$ and $M \in \mathcal{T}_r \backslash \{t_1(r), t_2(r), t_3(r)\}$" with M being on the line segment joining $t_1(r)$ and $t_2(r)$; "$r = 5/4$ and $M = (7/10, \sqrt{3}/10)$" is an example of the case "$r \in [1, 3/2)$ and $M \in \mathcal{T}_r \backslash \{t_1(r), t_2(r), t_3(r)\}$" with $M = t_2(r)$; and "$r = 3/2$ and $M = M_C$" is an example of the case discussed in (Ceyhan and Priebe, 2005)

$k\backslash n$	10	20	30	50	100	500	1000	2000
$r = 5/4$ and $M = (3/5, \sqrt{3}/10)$								
1	118	60	51	39	15	1	2	1
2	462	409	361	299	258	100	57	29
3	420	531	588	662	727	899	941	970
$r = 5/4$ and $M = (7/10, \sqrt{3}/10)$								
1	174	118	82	61	22	5	1	1
2	532	526	548	561	611	617	633	649
3	294	356	370	378	367	378	366	350
$r = 3/2$ and $M = M_C$								
1	151	82	61	50	27	2	3	1
2	602	636	688	693	718	753	729	749
3	247	282	251	257	255	245	268	250

3.2. *The Multiple Triangle Case*

In this section, we present the asymptotic distribution of the domination number of the proportional-edge PCDs in multiple Delaunay triangles. Suppose $\mathcal{Y}_m = \{\mathsf{y}_1, \mathsf{y}_2, \ldots, \mathsf{y}_m\} \subset \mathbb{R}^2$ be a set of m points in general position with $m > 3$ and no more than 3 points are cocircular. Then there are $J_m > 1$ Delaunay triangles each of which is denoted as T_j (Okabe et al., 2000). We wish to investigate

$$H_o : X_i \overset{iid}{\sim} \mathcal{U}(C_H(\mathcal{Y}_m)) \quad \text{for } i = 1, 2, \ldots, n \tag{4}$$

against segregation and association alternatives (see Sec. 4). Figure 10 (middle) presents a realization of 1,000 observations independent and identically distributed as $\mathcal{U}(C_H(\mathcal{Y}_m))$ for $m = 10$ and $J_m = 13$.

Let $M_{[j]}$ be the point in T_j that corresponds to M in T_e, $\mathcal{T}_r^j$ be the triangle that corresponds to $\mathcal{T}_r$ in T_e, and $t_i^j(r)$ be the vertices of $\mathcal{T}_r^j$ that correspond to $t_i(r)$ in T_e for $i \in \{1, 2, 3\}$. Moreover, let $n_j := |\mathcal{X}_n \cap T_j|$, the number of $\mathcal{X}$ points in Delaunay triangle T_j. The digraph D is constructed using $N_{PE}^r(\cdot, M_{[j]})$ as described in Sec. 2.1, where the three points in $\mathcal{Y}_m$ defining the Delaunay triangle T_j are used as $\mathcal{Y}_{m(j)}$. Then we have $\geq J_m$ disconnected sub-digraphs. For $\mathcal{X}_n \subset \mathcal{C}_H(\mathcal{Y}_m)$, let $\gamma_{[j]}(r)$ be the domination number of the digraph induced by vertices of T_j and $\mathcal{X}_n \cap T_j$. Then the domination number of the proportional-edge PCD in J_m triangles is

$$\gamma_{n,m}(r, M) = \sum_{j=1}^{J_m} \gamma_{[j]}(r);$$

see Fig. 7 for two examples of the proportional edge PCDs based on the 77 $\mathcal{X}$ points that are in $\mathcal{C}_H(\mathcal{Y}_m)$ out of the 200 $\mathcal{X}$ points plotted in Fig. 1. The arcs are constructed for $M = M_C$ with $r = 3/2$ (left) and $r = 5/4$ (right) and the corresponding domination number values are $\gamma_{n,10}(3/2, M_C) = 22$ and

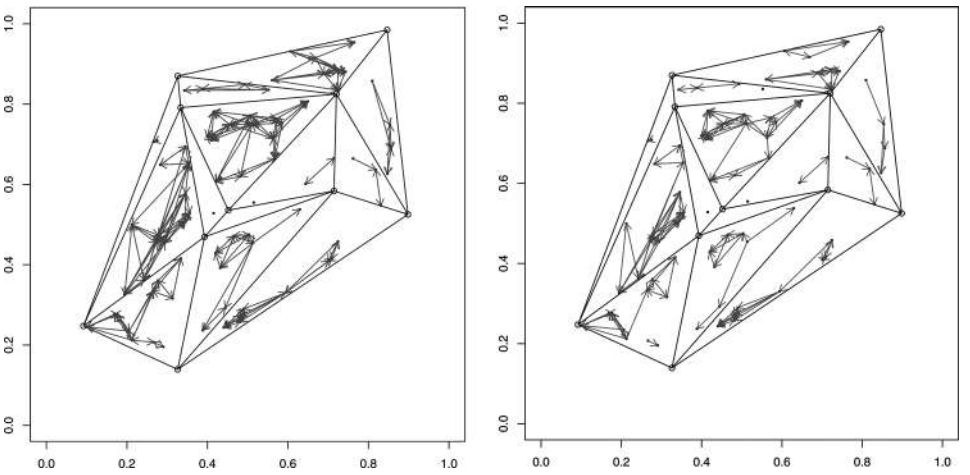

Figure 7. The arcs for the 77 $\mathcal{X}$ points (dots, **.**) in the convex hull of $\mathcal{Y}$ points (circles, ∘) given in Fig. 1 for the proportional-edge PCD with $M = M_C$ for $r = 3/2$ (left) and $r = 5/4$ (right).

$\gamma_{n,10}(5/4, M_C) = 26$. For fixed m (or fixed J_m), as $n \to \infty$, so does each n_j. Furthermore, as $n \to \infty$, each component $\gamma_{[j]}(r)$ become independent. Therefore, using Eq. (2), we can obtain the asymptotic distribution of $\gamma_{n,m}(r, M)$. For fixed J_m, as $n \to \infty$,

$$\gamma_{n,m}(r, M) \xrightarrow{\mathscr{L}} \begin{cases} 2J_m + \text{BIN}(J_m, 1 - p_r) & \text{for } M_{[j]} \in \left\{t_1^j(r), t_2^j(r), t_3^j(r)\right\} \text{ and } r \in [1, 3/2], \\ J_m & \text{for } r > 3/2 \text{ and for all } M_{[j]} \neq \mathscr{Y}_3, \\ 3J_m & \text{for } M \in \mathscr{T}_r^j \setminus \left\{t_1^j(r), t_2^j(r), t_3^j(r)\right\} \text{ and } r \in [1, 3/2), \end{cases} \tag{5}$$

where $\text{BIN}(n, p)$ stands for binomial distribution with n trials and probability of success p, for $r \in [1, 3/2)$ and $M \in \{t_1(r), t_2(r), t_3(r)\}$, p_r is given in Eq. (3) and $j = 1, 2, \ldots, J_m$. Observe that in the non degenerate case in Eq. (5), we have $\mathbf{E}[\gamma_{n,m}(r, M)] = J_m(3 - p_r)$ and $\mathbf{Var}[\gamma_{n,m}(r, M)] = J_m \cdot p_r \cdot (1 - p_r)$.

Theorem 3.3 (Asymptotic Normality)**.** *Suppose n_j and J_m are sufficiently large with $n_j \gg J_m$. Then the asymptotic null distribution of the mean domination number (per triangle) $\overline{G}(r, M) := \frac{1}{J_m} \sum_{j=1}^{J_m} \gamma_{[j]}(r) = \frac{\gamma_{n,m}(r,M)}{J_m}$ is approximately normal; i.e., for large $n_j \gg J_m$*

$$\overline{G}(r, M) \overset{approx}{\sim} \mathscr{N}(\mu, \sigma^2/J_m),$$

where $\mu = 3 - p_r$ and $\sigma^2 = p_r(1 - p_r)/J_m$.

Proof. For fixed J_m sufficiently large and each n_j sufficiently large with $n = \sum_{j=1}^{J_m} n_j \gg J_m$, $\gamma_{[j]}(r)$ are approximately independent identically distributed as in Eq. (2). Then the desired result follows from normal approximation to binomial distribution.

In Fig. 8 (top), we plot the histograms and the approximating normal curves for $\overline{G}(r, M)$ with $r = 3/2$ and $M = M_C$ for $n = 100, 1{,}000$, and $5{,}000$ $\mathscr{X}$ points generated iid $\mathscr{U}(C_H(\mathscr{Y}_m))$ where $\mathscr{Y}_m$ (which yields $J_m = 13$ triangles) is given in Fig. 1. Notice that, even though the distribution looks symmetric with $n = 100$, the normal approximation is not appropriate, since not all n_j are sufficiently large to make the binomial distribution hold as in Eq. (5), but as n increases (see $n = 1{,}000$ and $n = 5{,}000$ cases) the histograms and the corresponding normal curves become more similar indicating that the asymptotic normal approximation gets better, since all n_j are large. Larger J_m values require larger sample sizes in order to obtain approximate normality. With $J_{20} = 30$ triangles based on the Delaunay triangulation of 20 $\mathscr{Y}$ points iid uniform on the unit square (not presented), we plot the histograms and the approximating normal curves for $r = 3/2$ and $M = M_C$ in Figure 8 (bottom). Observe that with more triangles (i.e., as J_m increases), the normal approximation gets better.

As a corollary to Theorem 3.2, it follows that for $r_1 < r_2$, we have $\overline{G}(r_2, M) <^{ST} \overline{G}(r_1, M)$.

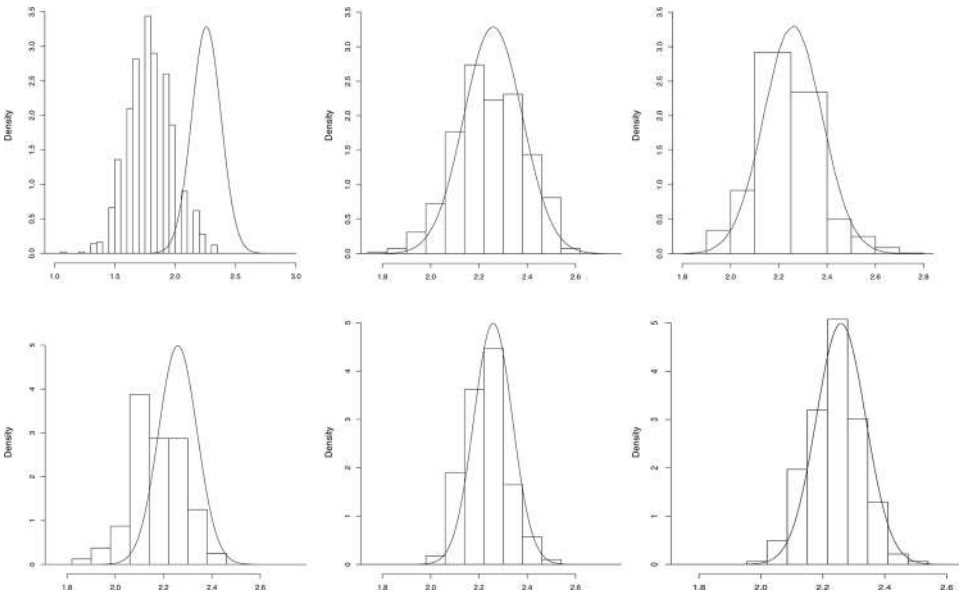

Figure 8. Depicted in the top row are $\overline{G}(r = 3/2, M = M_C) \overset{\text{approx}}{\sim} \mathcal{N}(\mu \approx 2.2587, \sigma^2/J_{10} \approx .1918/J_{10})$ for $J_{10} = 13$ and $n = 100$ (left), $n = 1{,}000$ (middle), and $n = 5{,}000$ (right). In the bottom row, depicted are $\overline{G}(r = 3/2, M = M_C) \overset{\text{approx}}{\sim} \mathcal{N}(\mu \approx 2.2587, \sigma^2/J_{20} \approx .1918/J_{20})$ for $J_{20} = 30$ and $n = 100$ (left), $n = 1{,}000$ (middle), and $n = 5{,}000$ (right). Histograms are based on 1,000 Monte Carlo replicates and the curves are the associated approximating normal curves.

4. Alternative Patterns: Segregation and Association

In a two-class setting, the phenomenon known as *segregation* occurs when members of one class have a tendency to repel members of the other class. For instance, it may be the case that one type of plant does not grow well in the vicinity of another type of plant, and vice versa. This implies, in our notation, that X_i are unlikely to be located near elements of $\mathcal{Y}$. Alternatively, association occurs when members of one class have a tendency to attract members of the other class, as in symbiotic species, so that X_i will tend to cluster around the elements of $\mathcal{Y}$, for example; see, for instance, Dixon (1994) and Coomes et al. (1999).

These alternatives can be parametrized as follows. In the one triangle case, without loss of generality let $\mathcal{Y}_3 = \{(0, 0), (1, 0), (c_1, c_2)\}$ and $T_b = T(\mathcal{Y}_3)$ with $\mathrm{y}_1 = (0, 0)$, $\mathrm{y}_2 = (1, 0)$, and $\mathrm{y}_3 = (c_1, c_2)$. For the basic triangle T_b, let $Q_\theta := \{x \in T_b : d(x, \mathcal{Y}_3) \leq \theta\}$ for $\theta \in (0, (c_1^2 + c_2^2)/2]$ and $S(F)$ be the support of F. Then consider

$$\mathcal{H}_S := \{F : S(F) \subseteq T_b \text{ and } P_F(X \in Q_\theta) < P_U(X \in Q_\theta)\}$$

and

$$\mathcal{H}_A := \{F : S(F) \subseteq T_b \text{ and } P_F(X \in Q_\theta) > P_U(X \in Q_\theta)\}$$

where P_F and P_U are probabilities with respect to distribution function F and the uniform distribution on T_b, respectively. So if $X_i \overset{iid}{\sim} F \in \mathcal{H}_S$, the pattern between $\mathcal{X}$

and $\mathcal{Y}$ points is segregation, but if $X_i \overset{iid}{\sim} F \in \mathcal{H}_A$, the pattern between $\mathcal{X}$ and $\mathcal{Y}$ points is association. For example, the distribution family

$$\mathcal{F}_S := \{F : S(F) \subset T_b \text{ and the associated pdf } f \text{ increases as } d(x, \mathcal{Y}_3) \text{ increases}\}$$

is a subset of $\mathcal{H}_S$ and yields samples from the segregation alternatives. Likewise, the distribution family

$$\mathcal{F}_A := \{F : S(F) \subset T_b \text{ and the associated pdf } f \text{ increases as } d(x, \mathcal{Y}_3) \text{ decreases}\}$$

is a subset of $\mathcal{H}_A$ and yields samples from the association alternatives.

In the basic triangle, T_b, we define the H_ε^S and H_ε^A with $\varepsilon \in \left(0, \sqrt{3}/3\right)$, for segregation and association alternatives, respectively. Under H_ε^S, $4\varepsilon^2/3 \times 100\%$ of the area of T_b is chopped off around each vertex so that the $\mathcal{X}$ points are restricted to lie in the remaining region. That is, for $\mathrm{y}_j \in \mathcal{Y}_3$, let e_j denote the edge of T_b opposite vertex y_j for $j = 1, 2, 3$, and for $x \in T_b$ let $\ell_j(x)$ denote the line parallel to e_j through x. Then define $T_j(\varepsilon) = \{x \in T_b : d(\mathrm{y}_j, \ell_j(x)) \le \varepsilon_j\}$ where $\varepsilon_1 = \frac{2\varepsilon c_2}{3\sqrt{c_2^2+(1-c_1)^2}}$, $\varepsilon_2 = \frac{2\varepsilon c_2}{3\sqrt{c_1^2+c_2^2}}$, and $\varepsilon_3 = \frac{2\varepsilon c_2}{3}$. Let $\mathcal{T}_\varepsilon := \bigcup_{j=1}^3 T_j(\varepsilon)$. Then under H_ε^S we have $X_i \overset{iid}{\sim} \mathcal{U}(T_b \backslash \mathcal{T}_\varepsilon)$. Similarly, under H_ε^A we have $X_i \overset{iid}{\sim} \mathcal{U}\left(\mathcal{T}_{\sqrt{3}/3-\varepsilon}\right)$. Thus, the segregation model excludes the possibility of any X_i occurring around a y_j, and the association model requires that all X_i occur around y_j's. The $\sqrt{3}/3 - \varepsilon$ is used in the definition of the association alternative so that $\varepsilon = 0$ yields H_o under both classes of alternatives. Thus, we have the below distribution families under this parametrization.

$$\mathcal{U}_\varepsilon^S := \{F : F = \mathcal{U}(T_b \backslash \mathcal{T}_\varepsilon)\} \quad \text{and} \quad \mathcal{U}_\varepsilon^A := \{F : F = \mathcal{U}(\mathcal{T}_{\sqrt{3}/3-\varepsilon})\}. \tag{6}$$

Clearly, $\mathcal{U}_\varepsilon^S \subsetneq \mathcal{H}_S$ and $\mathcal{U}_{\sqrt{3}/3-\varepsilon}^A \subsetneq \mathcal{H}_A$, but $\mathcal{U}_\varepsilon^S \nsubseteq \mathcal{F}_S$ and $\mathcal{U}_{\sqrt{3}/3-\varepsilon}^A \nsubseteq \mathcal{F}_A$.

These alternatives H_ε^S and H_ε^A with $\varepsilon \in \left(0, \sqrt{3}/3\right)$, can be transformed into the equilateral triangle as in Ceyhan et al. (2006, 2007).

For the standard equilateral triangle, in $T_j(\varepsilon) = \{x \in T_e : d(\mathrm{y}, \ell_j(x)) \le \varepsilon_j\}$ we have $\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = \varepsilon$. Thus, H_ε^S implies $X_i \overset{iid}{\sim} \mathcal{U}(T_e \backslash \mathcal{T}_\varepsilon)$ and H_ε^A be the model under which $X_i \overset{iid}{\sim} \mathcal{U}\left(\mathcal{T}_{\sqrt{3}/3-\varepsilon}\right)$; see Fig. 9 for a depiction of the above segregation and the association alternatives in T_e.

Remark 4.1. The geometry invariance result of Theorem 3.1 still holds under the alternatives H_ε^S and H_ε^A. In particular, the segregation alternative with $\varepsilon \in (0, \sqrt{3}/4)$ in the standard equilateral triangle corresponds to the case that in an arbitrary triangle, $\delta \times 100\%$ of the area is carved away as forbidden from the vertices using line segments parallel to the opposite edge where $\delta = 4\varepsilon^2$ (which implies $\delta \in (0, 3/4)$). But the segregation alternative with $\varepsilon \in (\sqrt{3}/4, \sqrt{3}/3)$ in the standard equilateral triangle corresponds to the case that in an arbitrary triangle, $\delta \times 100\%$ of the area is carved away as forbidden from each vertex using line segments parallel to the opposite edge where $\delta = 1 - 4(1 - \sqrt{3}\varepsilon)^2$ (which implies $\delta \in (3/4, 1)$). This argument is for the segregation alternative; a similar construction is available for the association alternative.

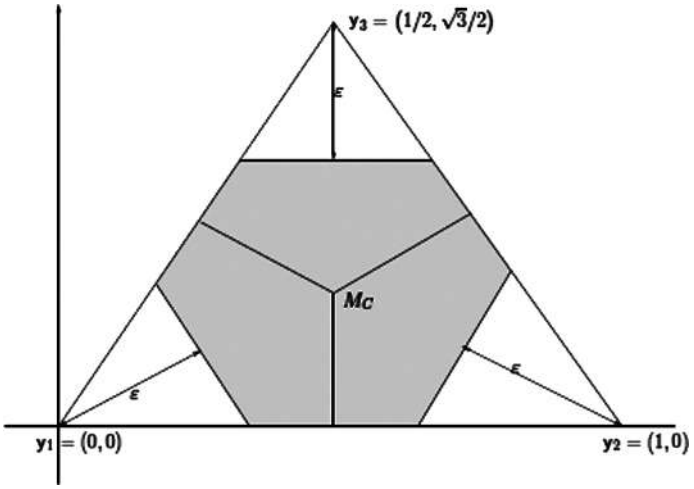

Figure 9. An example for the segregation alternative for a particular ε (shaded region), and its complement is for the association alternative (unshaded region) on the standard equilateral triangle.

4.1. *Asymptotic Distribution under the Alternatives*

Let $\gamma_n^S(F, r, M)$ be the domination number under segregation for $F \in \mathcal{H}_\theta^S$. Under this alternative with $M = M_C$, the domination number will have a multinomial distribution as $p_j^F := P(\gamma_n = j)$ for $j = 1, 2, 3$ and $p_1^F + p_2^F + p_3^F = 1$. Clearly, p_j^F values depend on the distribution F and their explicit forms for finite n or in the asymptotics are not always analytically tractable. The same holds for the domination number under association $\gamma_n^A(F, r, M)$ for $F \in \mathcal{H}_\theta^A$.

However, under the alternatives H_ε^S and H_ε^A, the asymptotic distribution of the domination number is much easier to find. Let $\gamma_n^S(\varepsilon, r, M)$ and $\gamma_n^A(\varepsilon, r, M)$ be the domination numbers under segregation and association alternatives, respectively. Under H_ε^S with $M = M_C$, the distribution of the domination number is non degenerate when $r = 3/2 - \varepsilon\sqrt{3}/2$ which implies $r \in (9/8, 3/2)$ for $\varepsilon \in (0, \sqrt{3}/4)$, and $r \in (1, 9/8)$ for $\varepsilon \in (\sqrt{3}/4, \sqrt{3}/3)$. In particular, the asymptotic distribution of the domination number for uniform data in one triangle is as follows. As $n \to \infty$, under H_ε^S with $M = M_C$ and $\varepsilon \in (0, \sqrt{3}/4)$,

$$\gamma_n^S(\varepsilon, r, M_C) \xrightarrow{\mathcal{L}} \begin{cases} 2 + \mathrm{BER}\left(1 - p_{r,\varepsilon}^S\right) & \text{for } r = 3/2 - \varepsilon\sqrt{3}/2, \\ 1 & \text{for } r > 3/2, \\ 2 & \text{for } 3/2 - \varepsilon\sqrt{3}/2 < r < 3/2, \\ 3 & \text{for } 9/8 < r < 3/2 - \varepsilon\sqrt{3}/2, \end{cases} \tag{7}$$

where $p_{r,\varepsilon}^S$ can be calculated similarly as in (3) for fixed numeric ε.

Furthermore, as $n \to \infty$, under H_ε^S with $M = M_C$ and $\varepsilon \in (\sqrt{3}/4, \sqrt{3}/3)$,

$$\gamma_n^S(\varepsilon, r, M_C) \xrightarrow{\mathcal{L}} \begin{cases} 2 + \mathrm{BER}\left(1 - p_{r,\varepsilon}^S\right) & \text{for } r = 3/2 - \varepsilon\sqrt{3}/2, \\ 1 & \text{for } r > 2 - \sqrt{3}\varepsilon, \\ 2 & \text{for } 3/2 - \varepsilon\sqrt{3}/2 < r < 2 - \sqrt{3}\varepsilon, \\ 3 & \text{for } 1 < r < 3/2 - \varepsilon\sqrt{3}/2. \end{cases} \tag{8}$$

Under H_ε^A with $M = M_C$, the domination number γ_n is non degenerate when $r = \sqrt{3}/(2\,\varepsilon)$ which implies $r > 2$ for $\varepsilon \in (0, \sqrt{3}/4)$, and $\varepsilon \in (3/2, 2)$ for $\varepsilon \in (\sqrt{3}/4, \sqrt{3}/3)$. In particular, the asymptotic distribution of the domination number for uniform data in one triangle is as follows. As $n \to \infty$, under H_ε^A with $M = M_C$ and $\varepsilon \in (0, \sqrt{3}/3)$,

$$\gamma_n^A(\varepsilon, r, M_C) \xrightarrow{\mathscr{L}} \begin{cases} 2 + \text{BER}\left(1 - p_{r,\varepsilon}^A\right) & \text{for } r = \sqrt{3}/(2\varepsilon), \\ 1 & \text{for } r > \sqrt{3}/(2\,\varepsilon), \\ 3 & \text{for } r < \sqrt{3}/(2\varepsilon), \end{cases} \tag{9}$$

where $p_{r,\varepsilon}^A$ can be calculated similarly as in (3) for fixed numeric ε. However, for finite n, $\gamma_n^A(\varepsilon, r, M_C)$ is also non degenerate for $\sqrt{3}/(2\varepsilon) - 1 < r < \sqrt{3}/(2\varepsilon)$.

Under segregation with general M, suppose $M \in T_e \setminus \bigcup_{y \in \mathscr{Y}_e} T(y, \varepsilon)$ (i.e., M is in the support of $\mathscr{X}$ points under H_ε^S). Then for fixed $r = r_o$ for which γ_n is non degenerate under CSR (i.e., r_o is a value such that $M \in \{t_1(r_o), t_2(r_o), t_3(r_o)\}$), then γ_n is non degenerate under H_ε^S if $r = r_o(2 - 4/(\sqrt{3}\varepsilon))$. For $r_o \in (4/3, 3/2)$, if $M \notin T_e \setminus \bigcup_{y \in \mathscr{Y}_e} T(y, \varepsilon)$ and $\varepsilon > \frac{3}{2}\left(1 - \frac{1}{2r}\right)$, then $\gamma_n \to 1$ in probability as $n \to \infty$; and the same also holds if $\sqrt{3}\left(1 - \frac{1}{r}\right) < \varepsilon < \frac{3}{2}\left(1 - \frac{1}{2r}\right)$. γ_n is non degenerate when $r = r_o(2 - 4\,\varepsilon/\sqrt{3})$. For general M, if $\varepsilon \in (0, \sqrt{3}/4)$, then γ_n is non degenerate when $r = r_o(1 - \varepsilon/\sqrt{3})$.

Under association with general M, when $M \notin \bigcup_{y \in \mathscr{Y}_e} T(y, \varepsilon)$ then γ_n is non degenerate when $r = r_o$ (i.e., M is not in the support of $\mathscr{X}$ points under H_ε^A). If $M \in \bigcup_{y \in \mathscr{Y}_e} T(y, \varepsilon)$ then γ_n is non degenerate when $r = \frac{\sqrt{3}(r_o - 2)}{2\varepsilon(r_o - 1) + \sqrt{3}(2r_o - 3)}$.

Theorem 4.1 (Stochastic Ordering). *Let $\gamma_n^S(\varepsilon, r, M)$ be the domination number under the segregation alternative with $\varepsilon > 0$. Then with $\varepsilon_j \in (0, \sqrt{3}/3)$, $j = 1, 2$, $\varepsilon_1 > \varepsilon_2$ implies that $\gamma_n^S(\varepsilon_1, r, M) <^{ST} \gamma_n^S(\varepsilon_2, r, M)$.*

Proof. Note that for $\varepsilon_1 > \varepsilon_2$ and finite n, $P(\gamma_n^S(\varepsilon_1, r, M) = 1) > P(\gamma_n^S(\varepsilon_2, r, M) = 1)$ and $P(\gamma_n^S(\varepsilon_1, r, M) = 2) > P(\gamma_n^S(\varepsilon_2, r, M) = 2)$, hence the desired result follows.

Note that for Theorem 4.1 to hold in the limiting case when $r \in [1, 3/2]$ and $M \in \{t_1(r), t_2(r), t_3(r)\}$, $\varepsilon_1 \in I_i(r)$ and $\varepsilon_2 \in I_j(r)$ should hold for $i < j$ where $I_1(r) = ((2 - r)/\sqrt{3}, \sqrt{3}/3)$, $I_2(r) = ((3 - 2r)/\sqrt{3}, (2 - r)/\sqrt{3})$, and $I_3(r) = (0, (3 - 2r)/\sqrt{3})$. For $\varepsilon \in \left(0, \sqrt{3}/4\right]$, $\gamma_n^S(\varepsilon, r, M) \to 2$ in probability as $n \to \infty$, and for $\varepsilon \in (\sqrt{3}/4, \sqrt{3}/3)$, $\gamma_n^S(\varepsilon, r, M) \to 1$ in probability as $n \to \infty$.

Similarly, the stochastic ordering result of Theorem 4.1 holds for association for all ε and $n < \infty$, with the inequalities being reversed.

Remark 4.2. The Alternatives in the Multiple Triangle Case. In the multiple triangle case, the segregation and association alternatives, H_ε^S and H_ε^A with $\varepsilon \in (0, \sqrt{3}/3)$, are defined as in the one-triangle case, in the sense that, when each triangle (together with the data in it) is transformed to the standard equilateral triangle as in Theorem 3.1, we obtain the same alternative pattern described above.

Let $\gamma_{n,m}^S(\varepsilon, r, M)$ and $\gamma_{n,m}^A(\varepsilon, r, M)$ be the domination numbers under segregation and association alternatives in the multiple triangle case with m triangles, respectively. The extensions of their distributions from Eqs. (7), (8), and (9) are

similar to the extension of the distribution of the domination number from one-triangle to multiple-triangle case under the null hypothesis in Sec. 3.2. Furthermore, the stochastic ordering result of Theorem 4.1 extends in a straightforward manner.

4.2. *The Test Statistics and Their Distributions*

A translated form of the domination number of the PCD is a test statistic for the segregation/association alternative:

$$B_{n,m} := \begin{cases} \gamma_n(r, M) - 2J_m = \sum_{j=1}^{J_m} \gamma_{[j]}(r) - 2J_m & \text{if } \gamma_n(r, M) > 2J_m, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Rejecting for extreme values of $B_{n,m}$ is appropriate, since under segregation we expect $B_{n,m}$ to be small, while under association we expect $B_{n,m}$ to be large. Using this test statistic the critical value for finite J_m and large n for the one-sided level α test against segregation is given by b_α, the $\alpha \times 100$th percentile of $\text{BIN}(J_m, 1 - p_r)$ (i.e., the test rejects for $B_{n,m} \leq b_\alpha$), and against association, the test rejects for $B_{n,m} \geq b_{1-\alpha}$.

Similarly, the mean domination number (per triangle) of the PCD which is defined as $\overline{G}(r, M) := \frac{1}{J_m} \sum_{j=1}^{J_m} \gamma_{[j]}(r)$, can also be used as a test statistic for the segregation/association alternative when $n \gg J_m$ and both n and J_m are sufficiently large. Using the standardized test statistic

$$S_{n,m} = \sqrt{J_m}(\overline{G}(r, M) - \mu)/\sigma, \tag{11}$$

where $\mu = 3 - p_r$ and $\sigma^2 = p_r(1 - p_r)$, the asymptotic critical value for the one-sided level α test against segregation is given by $z_\alpha = \Phi^{-1}(\alpha)$ where $\Phi(\cdot)$ is the standard normal distribution function. The test rejects for $S_{n,m} < z_\alpha$. Against association, the test rejects for $S_{n,m} > z_{1-\alpha}$.

Depicted in Fig. 10 are the segregation with $\delta = 3/16$, CSR, and association with $\delta = 1/4$ realizations for $m = 10$ and $J_m = 13$, and $n = 1{,}000$. The associated mean domination numbers with $r = 3/2$ are $2.000, 2.1538$, and 3.000, for the segregation alternative, null realization, and the association alternatives,

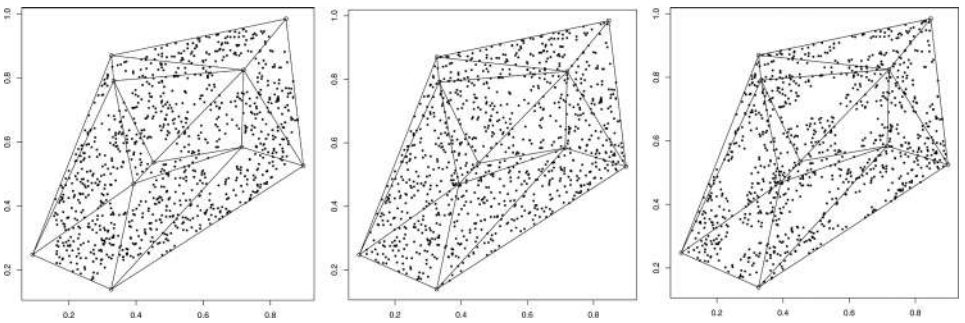

Figure 10. A realization of segregation (left), CSR (middle), and association (right) for $|\mathcal{Y}| = 10$, $J_{10} = 13$, and $n = 1{,}000$.

respectively, yielding p-values ≈ 0.000, 0.6139, and ≈ 0.000 based on binomial approximation, and p-values 0.0166, 0.3880, and < 0.0001 based on normal approximation. We also present a Monte Carlo power investigation in Sec. 5 for these cases.

Theorem 4.2 (Consistency-I)**.** *Let $\gamma^S_{n,m}(F, r, M)$ and $\gamma^A_{n,m}(F, r, M)$ be the domination numbers under segregation and association alternatives in the multiple triangle case with m triangles, respectively. The test against segregation with $F \in \mathcal{H}_S$ which rejects for $S_{n,m} < z_\alpha$ and the test against association with $F \in \mathcal{H}_A$ which rejects for $S_{n,m} > z_{1-\alpha}$ are consistent.*

Proof. Given $F \in \mathcal{H}_S$. Let $\gamma_{n,m}(\mathcal{U}, r, M)$ be the domination number for $\mathcal{X}_n$ being a random sample from $\mathcal{U}(T(\mathcal{Y}_3))$. Then $P(\gamma^S_{n,m}(F, r, M) = 1) \geq P(\gamma_{n,m}(\mathcal{U}, r, M) = 1)$; $P(\gamma^S_{n,m}(F, r, M) \leq 2) \geq P(\gamma^S_{n,m}(\mathcal{U}, r, M) \leq 2)$; and $P(\gamma^S_{n,m}(F, r, M) = 3) \leq P(\gamma^S_{n,m}(\mathcal{U}, r, M) = 3)$. Hence, $S_{n,m} < 0$ with probability 1, as $n \gg m \to \infty$. Hence consistency follows from the consistency of tests which have asymptotic normality. The consistency against the association alternative can be proved similarly.

Below, we provide a result which is stronger, in the sense that it will hold for finite m and $n \to \infty$.

Theorem 4.3 (Consistency-II)**.** *Let $\gamma^S_{n,m}(\varepsilon, r, M)$ and $\gamma^A_{n,m}(\varepsilon, r, M)$ be the domination numbers under segregation and association alternatives H^S_ε and H^A_ε in the multiple triangle case with m triangles, respectively. Let $J^*(\alpha, \varepsilon) := \left\lceil \left(\frac{\sigma \cdot z_\alpha}{\overline{G}(r,M) - \mu} \right)^2 \right\rceil$, where $\lceil \cdot \rceil$ is the ceiling function and ε-dependence is through $\overline{G}(r, M)$ under a given alternative. Then the test against H^S_ε which rejects for $S_{n,m} < z_\alpha$ is consistent for all $\varepsilon \in (0, \sqrt{3}/3)$ and $J_m \geq J^*(\alpha, \varepsilon)$, and the test against H^A_ε which rejects for $S_{n,m} > z_{1-\alpha}$ is consistent for all $\varepsilon \in (0, \sqrt{3}/3)$ and $J_m \geq J^*(1 - \alpha, \varepsilon)$.*

Proof. Let $\varepsilon > 0$. Under H^S_ε, $\gamma^S_n(\varepsilon, r, M)$ is degenerate in the limit as $n \to \infty$, which implies $\overline{G}(r, M)$ is a constant a.s. In particular, for $\varepsilon \in (0, \sqrt{3}/4]$, $\overline{G}(r, M) = 2$ and for $\varepsilon \in (\sqrt{3}/4, \sqrt{3}/3)$, $\overline{G}(r, M) \leq 2$ a.s. as $n \to \infty$. Then the test statistic $S_{n,m} = \sqrt{J_m}(\overline{G}(r, M) - \mu)/\sigma$ is a constant a.s. and $J_m \geq J^*(\alpha, \varepsilon)$ implies that $S_{n,m} < z_\alpha$ a.s. Hence, consistency follows for segregation.

Under H^A_ε, as $n \to \infty$, $\overline{G}(r, M) = 3$ for all $\varepsilon \in (0, \sqrt{3}/3)$, a.s. Then $J_m \geq J^*(1 - \alpha, \varepsilon)$ implies that $S_{n,m} > z_{1-\alpha}$ a.s., hence consistency follows for association.

Consistency in the sense of Theorems 4.2 and 4.3 also follows for $B_{n,m}$ similarly.

Remark 4.2 (Asymptotic Efficiency)**.** Pitman asymptotic efficiency (PAE) provides for an investigation of "local (around H_o) asymptotic power". This involves the limit as $n \to \infty$ as well as the limit as $\varepsilon \to 0$. A detailed discussion of PAE is available in Kendall and Stuart (1979) and Eeden (1963). For segregation or association alternatives H^S_ε and H^A_ε the PAE is not applicable because the Pitman conditions (Eeden, 1963) are not satisfied by the test statistic, $\overline{G}(r, M)$.

Hodges–Lehmann asymptotic efficiency analysis (Hodges and Lehmann, 1956) and asymptotic power function analysis (Kendall and Stuart, 1979) are not applicable here either. However, when $M = M_C$ (which also implies $r = 3/2$), for ε small and n large enough, this test is very sensitive for both alternatives because

$\gamma_n^S(\varepsilon, 3/2, M_C) \to 2$ in probability as $n \to \infty$ for segregation and $\gamma_n^A(\varepsilon, 3/2, M_C) \to 3$ in probability as $n \to \infty$ for association. That is, the test statistic becomes degenerate in the limit for all $\varepsilon > 0$ but in the correct direction for both alternatives. On the other hand, when $M \neq M_C$ (i.e., $r \neq 3/2$) this test is very sensitive for the segregation alternative since $\gamma_n^S(\varepsilon, r, M) \to 2$ in probability as $n \to \infty$; the same holds for the association alternative, but the test is not as sensitive as in the segregation case, since we only have $\gamma_n^A(\varepsilon, r, M) <^{ST} \gamma_n(r, M)$.

However, the variance of $\gamma_n(r, M)$ is minimized when $p_r = 1/2$, which happens when $r \approx 1.395$ (obtained numerically). Hence, we expect the test to have higher power under the alternatives for r around 1.40.

5. Monte Carlo Simulation Analysis

5.1. *Empirical Size Analysis under CSR*

For the null pattern of CSR, we generate n $\mathscr{X}$ points iid $\mathscr{U}(C_H(\mathscr{Y}_{10}))$ where $\mathscr{Y}_{10}$ is the set of the 10 $\mathscr{Y}$ points in Fig. 1. We calculate and record the domination number $\gamma_n(r, M)$ and the mean domination number (per triangle), $\overline{G}(r, M)$ for $r = 1.00, 1.01, 1.02, \ldots, 1.49$ at each Monte Carlo replicate. We repeat the Monte Carlo procedure $N_{mc} = 1{,}000$ times for each of $n = 500, 1{,}000, 2{,}000$. Using the critical values based on the binomial distribution for the domination number and the normal approximation for $\overline{G}(r, M)$, we calculate the empirical size estimates for both right- and left-sided tests. The empirical sizes significantly smaller (larger) than .05 are deemed conservative (liberal). The asymptotic normal approximation to proportions is used in determining the significance of the deviations of the empirical sizes from .05. For these proportion tests, we also use $\alpha = .05$ as the significance level. With $N_{mc} = 1{,}000$, empirical sizes less than .039 are deemed conservative, greater than .061 are deemed liberal at $\alpha = .05$ level. The empirical sizes together with upper and lower limits of liberalness and conservativeness are plotted in Fig. 11. Observe that right-sided tests are liberal with being less liberal when sample size n increases, and it has about the nominal level for most r values between 1.1 and 1.4. The left-sided test tends to be liberal for small r, and conservative for large r, but has about the desired nominal level for r around 1.2 and 1.3.

Since p_r has a different form when $r = 3/2$, we estimate the empirical sizes for $r = 3/2$ separately. The size estimates for $n = 500, 1{,}000,$ and $2{,}000$ relative to segregation and association alternatives are presented in Table 3. Based on the Monte Carlo simulations under CSR, the use of domination number for $r \in (1.45, 1.50)$ is not recommended, as the test is extremely liberal for the segregation (i.e., left-sided) alternative, while it is extremely conservative for the association (i.e., right-sided) alternative. This deviation from the nominal level for the test is due to the fact that for $r \in (1.45, 1.50)$ much larger sample sizes are required for the binomial and the normal approximations to hold.

5.2. *Empirical Power Analysis under the Alternatives*

To compare the distribution of the test statistic under CSR, and the segregation and association alternatives, we generate n points iid $\mathscr{U}(C_H(\mathscr{Y}_m))$ under CSR, iid uniformly on the support that corresponds to $H^S_{\sqrt{3}/8}$, and iid uniformly on the

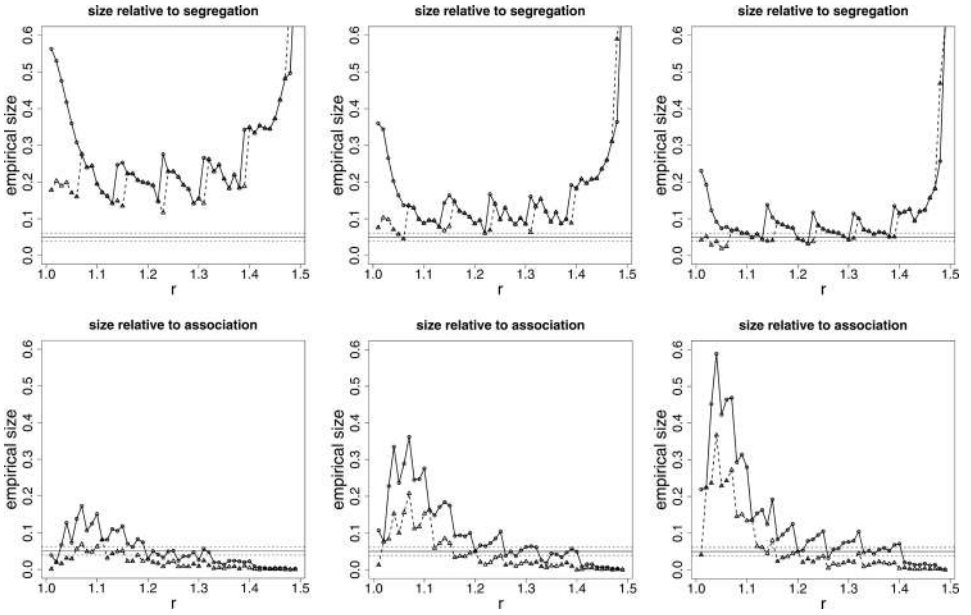

Figure 11. The empirical size estimates for the left-sided alternative (i.e., relative to segregation) and the right-sided alternative (i.e., relative to association) with $n = 500$ (left), $n = 1{,}000$ (middle), and $n = 2{,}000$ (right) under the CSR pattern. The empirical sizes based on the binomial distribution are plotted in circles (∘) and joined with solid lines, and those based on the normal approximation are plotted in triangles (△) and joined with dashed lines. The horizontal lines are located at .039 (upper threshold for conservativeness), .050 (nominal level), and .061 (lower threshold for liberalness).

support that corresponds to $H^A_{\sqrt{3}/21}$ for each triangle based on the same $\mathcal{Y}_m$ points. Under each case, we generate $n = 1{,}000$ points with $J_{10} = 13$ and $n = 5{,}000$ points with $J_{20} = 30$ for 500 Monte Carlo replicates. The kernel density estimates of $\overline{G}(r = 3/2, M = M_C)$ are presented in Figs. 12 and 13. In Fig. 12, we observe empirically that even under mild segregation we obtain considerable separation between the kernel density estimates under null and segregation cases for moderate J_m and n values suggesting high power at $\alpha = .05$.

In Fig. 13, we observe that even in mild association we obtain considerable separation for moderate J_m and n values suggesting high power (with $J_{10} = 13$ and $n = 1{,}000$, the empirical critical value is 2.46, $\hat{\alpha} = .034$ and empirical power is $\hat{\beta} = 1.0$ and with $J_{20} = 30$, $n = 5{,}000$, the empirical critical value is 2.36, $\hat{\alpha} = .04$ and empirical power is $\hat{\beta} = 1.0$).

For the segregation alternatives, we consider the following three cases: $\varepsilon = \sqrt{3}/8$, $\varepsilon = \sqrt{3}/4$, $\varepsilon = 2\sqrt{3}/7$ in the 13 Delaunay triangles obtained by the 10 $\mathcal{Y}$ points in Fig. 1. We generate $n = 500, 1{,}000, 2{,}000, 5{,}000$ in the convex hull of $\mathcal{Y}_{10}$ at each Monte Carlo replication. We estimate the empirical power of the tests for $r = 1.00, 1.01, 1.02, \ldots, 1.49$ values using $N_{mc} = 1{,}000$ replicates. The power estimates based on the binomial distribution and normal approximation under $H^S_{\sqrt{3}/8}$ for $n = 1{,}000, 2{,}000, 5{,}000$ are plotted in Fig. 14. Observe that the power estimates are about 1.0 for $r \gtrsim 1.15$. Considering the empirical size and power estimates together, we recommend r values around 1.22 or 1.30 for the segregation alternatives.

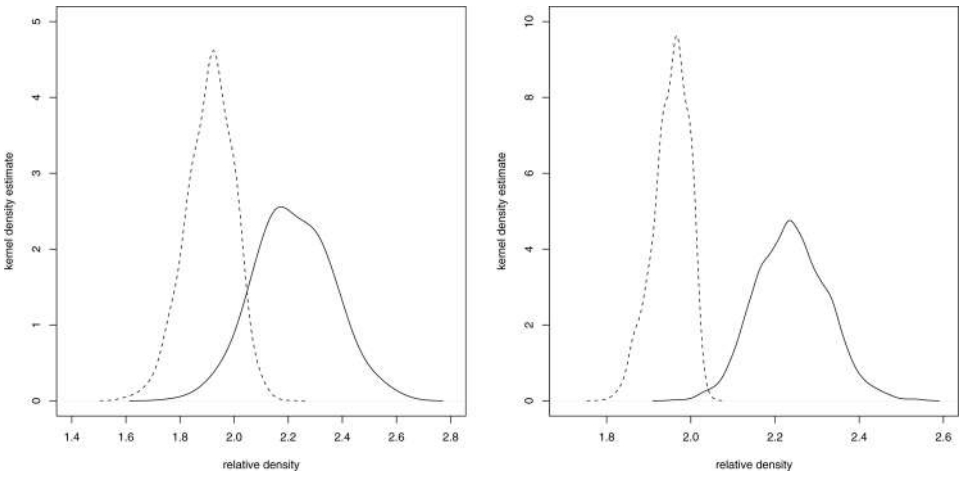

Figure 12. Two Monte Carlo experiments against the segregation alternatives $H^S_{\sqrt{3}/8}$ with $\delta = 1/16$. Depicted are kernel density estimates of $\overline{G}(r = 3/2, M = M_C)$ for $J = 13$ and $n = 1{,}000$ with 1,000 replicates (left) and $J_{20} = 30$ and $n = 5{,}000$ with 1,000 replicates (right) under the null (solid) and segregation alternative (dashed).

For the association alternatives, we consider the following three cases: $\varepsilon = 5\sqrt{3}/24$, $\varepsilon = \sqrt{3}/12$, $\varepsilon = \sqrt{3}/21$ in the 13 Delaunay triangles obtained by the 10 $\mathcal{Y}$ points in Fig. 1. We generate $n = 500, 1{,}000, 2{,}000, 5{,}000$ in the convex hull of $\mathcal{Y}_{10}$ at each Monte Carlo replication. We estimate the empirical power of the tests for $r = 1.00, 1.01, 1.02, \ldots, 1.49$ values using $N_{mc} = 1{,}000$ replicates. The power estimates based on the binomial distribution and normal approximation under

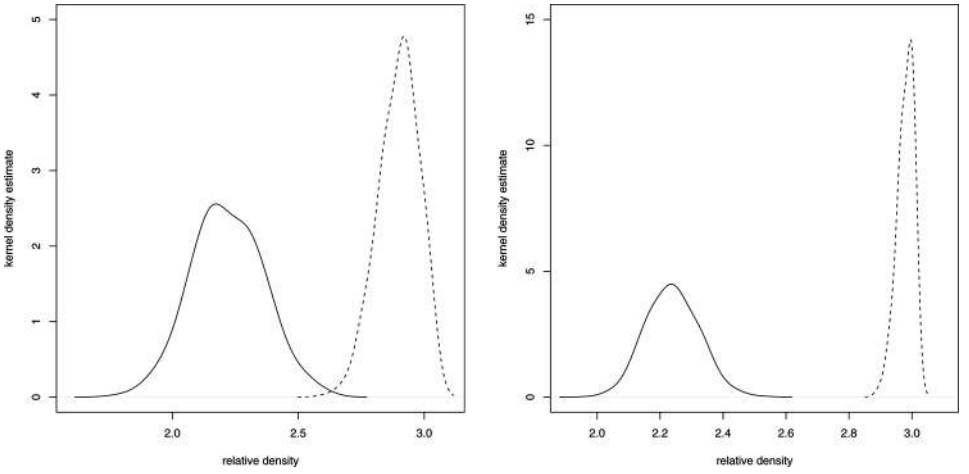

Figure 13. Two Monte Carlo experiments against the association alternatives $H^A_{\sqrt{3}/21}$, i.e., $\delta = 16/49$. Depicted are kernel density estimates of $\overline{G}(r = 3/2, M = M_C)$ for $J = 13$ and $n = 1{,}000$ with 500 replicates (left) and $J_{20} = 30$ and $n = 5{,}000$ with 100 replicates under the null (solid) and association alternative (dashed).

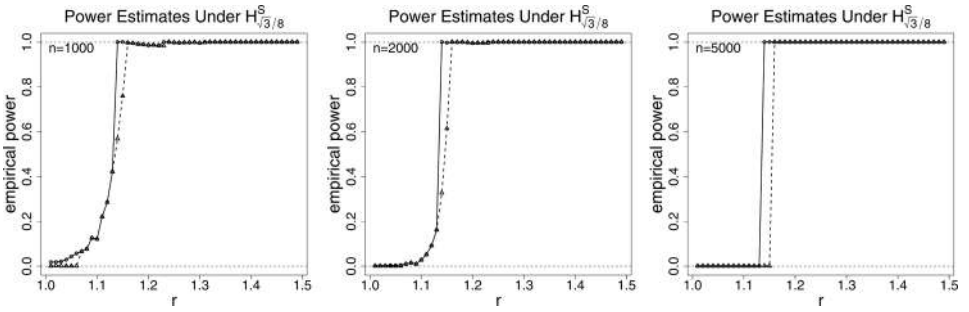

Figure 14. The empirical power estimates under segregation with $\varepsilon = \sqrt{3}/8$, $\varepsilon = \sqrt{3}/4$ and $n = 1{,}000$ (left), $n = 2{,}000$ (middle), and $n = 5{,}000$ (right). The power estimates based on the binomial distribution are plotted in circles (∘) and joined with solid lines, and those based on the normal approximation are plotted in triangles (△) and joined with dashed lines.

$H^A_{5\sqrt{3}/24}$ for $n = 1{,}000, 2{,}000, 5{,}000$ are plotted in Fig. 15. Observe that the power estimates are about 1.0 for $r \gtrsim 1.33$, but the power performance is poor for r between 1.1 and 1.33. Considering the empirical size and power estimates together, we recommend r values around 1.35 for the association alternatives.

The empirical power estimates for $r = 3/2$ and $M = M_C$ are presented in Table 3 also.

Remark 5.1. The choice of the null pattern in Sec. 3.2 and the conditions in Theorem 3.3 seem to be somewhat stringent; i.e., $\mathscr{X}$ points are assumed to be uniformly distributed in the convex hull of $\mathscr{Y}$ points, which might not be realistic in practice. However, if the supports of distributions of $\mathscr{X}$ and $\mathscr{Y}$ points do not intersect, or mildly intersect, then it is clear that the null hypothesis is violated (i.e., two classes are segregated) which is easily detected by the test statistics $B_{n,m}$ or $S_{n,m}$ (see Eqs. (10) and (11)) as they tend to be smaller under segregation than expected under CSR. When their supports have non-empty intersection, then either the $\mathscr{X}$ points are segregated from the $\mathscr{Y}$ points, or follow CSR, or are associated with the $\mathscr{Y}$ points in this intersection. Then we only consider the $\mathscr{Y}$ points in this support

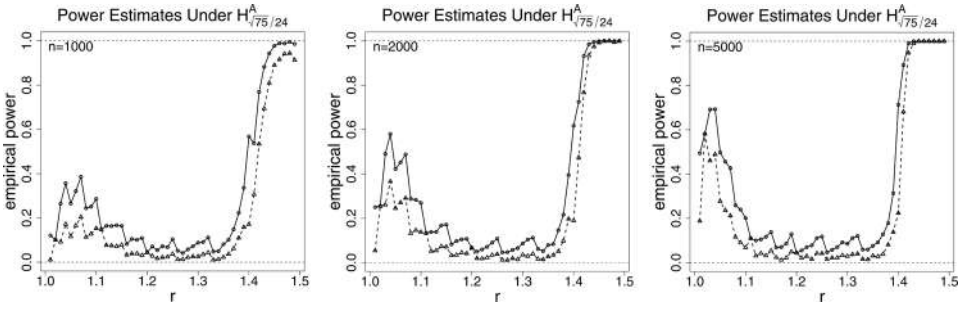

Figure 15. The empirical power estimates under association with $\varepsilon = 5\sqrt{3}/24$, $\varepsilon = \sqrt{3}/12$ and $n = 1{,}000$ (left), $n = 2{,}000$ (middle), and $n = 5{,}000$ (right). The power estimates based on the binomial distribution are plotted in circles (∘) and joined with solid lines, and those based on the normal approximation are plotted in triangles (△) and joined with dashed lines.

Table 3
The empirical size and power estimates for $r = 3/2$ and $M = M_C$ under the null and alternatives. n stands for number of $\mathscr{X}$ points, $\hat{\alpha}^S$ for empirical size relative segregation, $\hat{\alpha}^A$ for empirical size relative to association, $\hat{\beta}_1^S$, $\hat{\beta}_2^S$, and $\hat{\beta}_3^S$ for empirical power estimates under H_ε^S with $\varepsilon = \sqrt{3}/8$, $\varepsilon = \sqrt{3}/4$, and $\varepsilon = 2\sqrt{3}/7$, respectively, $\hat{\beta}_1^A$, $\hat{\beta}_2^A$, and $\hat{\beta}_3^A$ for empirical power estimates under H_ε^A with $\varepsilon = 5\sqrt{3}/24$, $\varepsilon = \sqrt{3}/12$, and $\varepsilon = \sqrt{3}/21$, respectively

Empirical size and power estimates for $r = 3/2$ and $M = M_C$								
n	$\hat{\alpha}^S$	$\hat{\alpha}^A$	$\hat{\beta}_1^S$	$\hat{\beta}_2^S$	$\hat{\beta}_3^S$	$\hat{\beta}_1^A$	$\hat{\beta}_2^A$	$\hat{\beta}_3^A$
500	0.161	0.062	0.961	1.000	1.000	1.000	1.000	0.997
1000	0.071	0.082	0.975	1.000	1.000	1.000	1.000	1.000
2000	0.049	0.081	0.995	1.000	1.000	1.000	1.000	1.000

intersection, then our inference will be local (i.e., restricted to this intersection). If one takes all of the $\mathscr{Y}$ points, then our inference will be a global one (i.e., for the entire support of $\mathscr{Y}$ points).

6. Correction for $\mathscr{X}$ Points Outside the Convex Hull of $\mathscr{Y}_m$

Our null hypothesis in (4) is rather restrictive, in the sense that, it might not be that realistic to assume the support of $\mathscr{X}$ being $C_H(\mathscr{Y}_m)$ in practice. Until now, our inference is restricted to the $C_H(\mathscr{Y}_m)$. However, crucial information from the data (hence power) might be lost since a substantial proportion of $\mathscr{X}$ points, denoted π_{out}, might fall outside the $C_H(\mathscr{Y}_m)$. We investigate the effect of π_{out} (or restriction to the $C_H(\mathscr{Y}_m)$) on our tests and propose an empirical correction to mitigate this based on an extensive Monte Carlo simulation study.

We consider the following six cases to investigate how the removal of points outside $C_H(\mathscr{Y}_m)$ affects the empirical size and power performance of the tests. We only consider $r = 1.35$ and $r = 1.5$ which have better size and power performances compared to others. In each case, at each Monte Carlo replication, we generate $\mathscr{X}_n$ and $\mathscr{Y}_m$ independently as random samples from $\mathscr{U}(\mathscr{S}_X)$ and $\mathscr{U}(\mathscr{S}_Y)$, respectively, for various values of n and m where $\mathscr{S}_X$ and $\mathscr{S}_Y$ are the support sets of $\mathscr{X}$ and $\mathscr{Y}$ points, respectively. We take $\mathscr{S}_Y = (0, 1) \times (0, 1)$ and manipulate $\mathscr{S}_X$ in each case to simulate CSR and various forms of deviations from CSR. We repeat the generation procedure N_{mc} times for each combination of m and n. At each Monte Carlo replication, we record the number of $\mathscr{X}$ points outside $C_H(\mathscr{Y}_m)$ and the domination number, $\gamma_{m,n}(r)$.

Case 1. In this case, we also set $\mathscr{S}_X = (0, 1) \times (0, 1)$.

Case 2. $\mathscr{S}_X = (-\delta, 1 + \delta) \times (-\delta, 1 + \delta)$ for $\delta \in \{.1, .25, .5\}$.

Case 3. $\mathscr{S}_X = (0, 1) \times (0, 1 + \delta)$ for $\delta \in \{.1, .25, .5\}$,

Case 4. $\mathscr{S}_X = (0, 1) \times (\delta, 1 + \delta)$ for $\delta \in \{.1, .25, .5\}$.

Case 5. Given a realization of $\mathscr{Y}$ points, $\mathscr{Y}_m = \{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_m\}$, from $\mathscr{U}(\mathscr{S}_Y = (0, 1) \times (0, 1))$, $\mathscr{S}_X = [(-\delta, 1 + \delta) \times (-\delta, 1 + \delta)] \setminus \bigcup_{i=1}^m B(\mathrm{y}_i, \varepsilon)$ with $\delta = \frac{1}{2\sqrt{\lambda}} = \frac{1}{2\sqrt{m}}$ which the expected interpoint distance in a homogeneous Poisson process with

intensity (expected number of points per unit area) λ (Dixon, 2002b) and $\varepsilon = \delta/k$ for $k = 1.5, 2.0$,

Case 6. Given a realization of $\mathcal{Y}$ points, $\mathcal{Y}_m = \{\mathsf{y}_1, \mathsf{y}_2, \ldots, \mathsf{y}_m\}$, from $\mathcal{U}(\mathcal{S}_Y)$, $\mathcal{S}_X = \bigcup_{i=1}^{m} B(\mathsf{y}_i, \varepsilon)$ with $\varepsilon = \delta/k$, $\delta = \frac{1}{2\sqrt{m}}$, and $k = 1.0, 1.5$.

Notice that in Case 1 both $\mathcal{X}_n$ and $\mathcal{Y}_m$ have the same support. By construction the two classes follow CSR independence with very different relative abundances (i.e., number of $\mathcal{X}$ points being larger than number of $\mathcal{Y}$ points). In Cases 2 and 3, the support of $\mathcal{X}_n$ contains (but larger than) the support of $\mathcal{Y}_m$, which suggests segregation of $\mathcal{X}$ points from $\mathcal{Y}$ points, at least when we move away from the support of $\mathcal{Y}$ points (which is the unit square). However, when we restrict our attention to $C_H(\mathcal{Y}_m)$ or the unit square, we have CSR or CSR independence, respectively. Furthermore, the larger the δ value, the larger the level of segregation of $\mathcal{X}$ from $\mathcal{Y}$. In Case 4, the support of $\mathcal{X}_n$ and $\mathcal{Y}_m$ overlap, but neither is a subset of the other, which suggests segregation between $\mathcal{X}$ and $\mathcal{Y}$ points. When we restrict our attention to $C_H(\mathcal{Y}_m)$, there is still segregation between $\mathcal{X}$ and $\mathcal{Y}$ points. Furthermore, the larger the δ value, the larger the level of segregation between $\mathcal{X}$ and $\mathcal{Y}$ points. In Case 5, $\mathcal{X}$ points are segregated from $\mathcal{Y}$ points both in and outside $C_H(\mathcal{Y}_m)$. Furthermore, the larger the δ value, the larger the level of segregation of $\mathcal{X}$ points from $\mathcal{Y}$ points. Finally, in Case 6, $\mathcal{X}$ points are associated with $\mathcal{Y}$ points. Furthermore, the smaller the δ value, the larger the level of association of $\mathcal{X}$ points with $\mathcal{Y}$ points.

In Case 1 (i.e., the benchmark case), we consider $n = 100, 200, \ldots, 900, 1{,}000, 2{,}000, \ldots, 9{,}000, 10{,}000$ for each of $m = 10, 20, \ldots, 50$. We generate $N_{mc} = 1{,}000$ replication for each n, m combination. In the other cases, we consider $n = 100, 500, 1{,}000$ for $m = 10$ and $n = 500, 1{,}000$ for $m = 20$; and we generate $N_{mc} = 10{,}000$ replication for each n, m combination.

In Cases 1–6, we estimate the proportion of $\mathcal{X}$ points outside the $C_H(\mathcal{Y}_m)$. For each m, n combination we average (over n) this proportion which is denoted as $\hat{\pi}_{out}$. We present the estimated (mean) proportions $\hat{\pi}_{out}$ for Case 1 in Table 4 and for Cases 2–6 in technical report (Ceyhan, 2009b). In Cases 2–5, $\hat{\pi}_{out}$ values are larger than that in Case 1, while in Case 6, $\hat{\pi}_{out}$ values are smaller than that in Case 1.

For Case 1, we model the relationship between $\hat{\pi}_{out}$ and m. Our simulation results suggest that $\hat{\pi}_{out} \approx 1.7932/m + 1.2229/\sqrt{m}$. Notice that as $m \to \infty$, $\hat{\pi}_{out} \to 0$. We present the actual fitted values denoted $\hat{\pi}_{fit}$ based on this model in Table 4.

Based on our Monte Carlo simulation results (presented in Ceyhan, 2009b) we propose a coefficient to adjust for the proportion of $\mathcal{X}$ points outside $C_H(\mathcal{Y}_m)$,

Table 4
The (mean) proportion of $\mathcal{X}$ points outside the $C_H(\mathcal{Y}_m)$ which is denoted as $\hat{\pi}_{out}$ and the fitted values $\hat{\pi}_{fit}$ for various m values in Case 1

m	10	20	30	40	50
$\hat{\pi}_{out}$	0.56	0.37	0.29	0.23	0.20
$\hat{\pi}_{fit}$	0.57	0.36	0.28	0.24	0.21

namely,

$$C_{ch} := 1 - (p_{out} - \mathbf{E}[\hat{\pi}_{out}]) \tag{12}$$

where p_{out} is the observed and $\mathbf{E}[\hat{\pi}_{out}] \approx 1.7932/m + 1.2229/\sqrt{m}$ is the expected (under the conditions stated in Case 1) proportion of $\mathscr{X}$ points outside $C_H(\mathscr{Y}_m)$. For the binomial test statistic in Eq. (10), we suggest

$$B_{n,m}^{ch} := \begin{cases} (\gamma_n(r, M) - 2J_m) \cdot C_{ch} = \left(\sum_{j=1}^{J_m} \gamma_{[j]}(r) - 2J_m \right) \cdot C_{ch} & \text{if } \gamma_n(r, M) \cdot C_{ch} > 2J_m, \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

For the mean domination number (per triangle) of the PCD, we suggest

$$S_{n,m}^{ch} = S_{n,m} \cdot C_{ch}. \tag{14}$$

This (convex hull) adjustment slightly affects the empirical size estimates in Case 1, since p_{out} and $\mathbf{E}[\hat{\pi}_{out}]$ values are very similar. In Cases 2–5, there is segregation when all data points are considered, and p_{out} values tend to be larger than $\mathbf{E}[\hat{\pi}_{out}]$ values, and in Case 6 (which is the simulation of the association case), p_{out} values tend to be smaller than $\mathbf{E}[\hat{\pi}_{out}]$ values. Hence, in Cases 2–6, the adjustment seems to correct the power estimates in the desired direction, thereby increasing the power estimates; see Ceyhan (2009b) for more detail.

Remark 6.1 (Correction for Small Samples)**.** The distributional results in Eqs. (2) and (5) might require large n for the convergence to hold. In particular, it might be necessary for the number of $\mathscr{X}$ points per Delaunay triangle to be larger than 100 as a practical guide which implies very large samples from $\mathscr{X}$ are needed for a large number of $\mathscr{Y}$ points. Hence, it might be necessary to propose a correction in the test statistics for small n also. Based on our extensive Monte Carlo simulations (of Case 1 above), we suggest that the test statistic $S_{n,m}$ in Eq. (11) can be adjusted as $S_{n,m}^{adj} := \frac{S_{n,m} - a_{n,m}}{b_{n,m}}$. We provide the explicit forms of $a_{n,m}$ and $b_{n,m}$ for $m = 10, 20, \ldots, 50$ in Ceyhan (2009b). For example for $m = 10$, $S_{n,m}$ in Eq. (11) can be adjusted as $S_{n,m}^{adj} := \frac{S_{n,m} - a_{n,m}}{b_{n,m}}$ where $a_{n,m} = -8.80/(n/J_m) - 30.94/\sqrt{n/J_m} + 9.09/\sqrt[3]{n/J_m}$ and $b_n = 1 - 18.81/(n/J_m) + 16.26/\sqrt{n/J_m} - 4.42/\sqrt[3]{n/J_m}$. Observe that as expected $S_{n,m}^{adj}$ converges to $S_{n,m}$ as $n \to \infty$ for each m value considered provided $n/J_m \to \infty$ which is a requirement in our testing framework. See the technical report Ceyhan (2009b) for further details.

7. Example Data Set

We illustrate the method on a forestry data set (namely, swamp tree data). Good and Whipple (1982) considered the spatial patterns of tree species along the Savannah River, South Carolina, U.S.A. From this data, Dixon (2002b) used a single 50 m × 200 m rectangular plot to illustrate his nearest neighbor contingency table (NNCT) methods. All live or dead trees with 4.5 cm or more dbh (diameter at breast height) were recorded together with their species. Hence, it is an example of

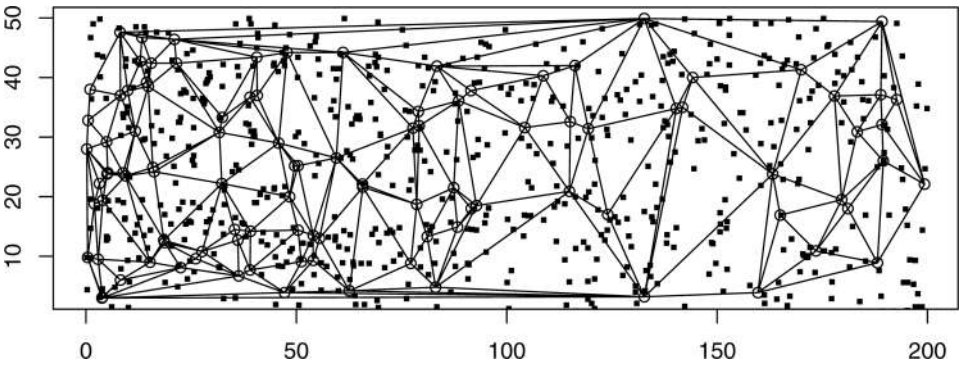

Figure 16. The scatter plot of the locations of dead trees (circle ∘) and live trees (black squares ■) in the swamp tree data. The Delaunay triangulation is based on the locations of the dead trees.

a realization of a marked multi-variate point pattern. The plot contains 13 different tree species, 4 of which comprising over 90% of the 734 tree stems; see Ceyhan (2009a) for more detail on the data.

In this article, we investigate the spatial interaction of live trees with dead ones (i.e., live trees are taken to be the $\mathscr{X}$ points, while dead trees are taken to be the $\mathscr{Y}$ points; hence, Delaunay triangulation is based on the locations of dead trees). The study area contains 630 dead and 104 live trees; see also Fig. 16 which is suggestive of segregation of live trees from dead trees.

We calculate the domination number, $\gamma_{n,m}(r)$, for $r = 1.01, 1.02, 1.03, \ldots, 1.49, 1.50$ values (not presented) and the proportion of live trees outside the convex hull of dead trees to be $p_{out} = 0.12$ (the expected proportion is $\pi_{out} = 0.14$). Without convex hull adjustment, the $\gamma_{n,m}(r)$ values range from 176–206 and with convex hull adjustment, they range from 179.25–209.81 (since $p_{out} < \pi_{out}$). The corresponding test statistics of Eqs. (10), (11), (13), and (14) (i.e., binomial and normal approximation test statistics with and without convex hull correction) yield very small p-values ($p < .0001$) for the left-sided alternative for each r. We also perform a Monte Carlo randomization test as follows. First we calculate the domination number, denoted γ_{obs}, for the current data set. Then we randomly assign 104 of the trees as "dead trees" (without replacement) and the remaining trees as "live trees", then calculate the domination number for the live trees within the convex hull of the dead trees. We repeat this procedure 999 times. Combining the observed γ_{obs} with these Monte Carlo randomization γ values, we obtain 1,000 γ values. We sort these γ values and determine the rank of the γ_{obs} value. This rank divided by 1,000 (or 1 minus rank divided by 1,000) will yield the estimated p-value for the left sided alternative (or the right sided alternative). Here, we also apply the convex hull correction by determining the proportion of live trees outside the convex hull of dead trees, and multiply the γ values by the correction coefficient in Eq. (12). Then we determine the estimated p-value for these convex-hull-corrected γ values as before. With $r = 1.5$ our Monte Carlo randomization procedure (without convex hull correction) yields $p \leq .002$ and with convex hull correction we get $p \leq .005$. With $r = 1.35$, our Monte Carlo randomization procedure (without convex hull correction) yields $p \leq .002$ and with convex hull correction we get $p \leq .004$. Hence, there is evidence for significant segregation of live trees from dead ones,

Table 5
The NNCT for swamp tree data (left) and the corresponding percentages (right)

		NN				NN		
		Live	Dead	Sum		Live	Dead	
Base	Live	536	96	632	Live	85%	15%	86%
	Dead	73	31	104	Dead	70%	30%	14%
	Sum	609	127	736		83%	17%	100%

which might indicate that the factors that cause trees to die (e.g., the soil content or quality) force the trees to cluster in favorable locations.

We also analyze the same data in a 2×2 NNCT with Dixon's overall test of segregation (Dixon, 2002a) and Ceyhan's NNCT-test (Ceyhan, 2008b). See Table 5 for the corresponding NNCT and the percentages (observe that the row sum for live trees is 632 instead of 630 due to ties in nearest neighbor (NN) distances). The cell percentages are relative to the row sums (i.e., number of dead or live trees) and marginal percentages are relative to the overall sum. Notice that the table is suggestive of segregation especially for the nearest NN with respect to the dead trees. That is, live trees seem to be segregated from the dead trees. Dixon's overall test statistic is $C_D = 19.13$ ($p = 0.0001$) and Ceyhan's test is $C_N = 10.01$ ($p = 0.0016$), both of which are suggestive of significant deviation from CSR independence. Considering the NNCT and Ceyhan's cell-specific tests (Ceyhan, 2008b), the test statistics for cells $1, 1$ and $2, 2$ are 3.15 and 3.16, respectively, with the corresponding p-values being $p = 0.0008$ for both cells. Dixon's cell-specific tests are not presented as they are not robust to differences in relative abundances (Ceyhan, 2008b). These results support the claim that live trees are significantly segregated from dead trees. So, NNCT-analysis and our domination number approach give similar results about the spatial interaction of live trees with dead ones. However, NNCT and our domination number approach answer different questions. More specifically, NNCT-tests in this example tests both directions of the spatial interaction, while the domination number approach only tests the spatial interaction of live trees with the dead ones, but not vice versa.

8. Extension of Proportional-Edge Proximity Regions to Higher Dimensions

The extension to $\mathbb{R}^d$ for $d > 2$ with $M = M_C$ is provided in Ceyhan and Priebe (2005), the extension for general M is similar: Let $\mathcal{Y} = \{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_{d+1}\}$ be $d+1$ non coplanar points. Denote the simplex formed by these $d+1$ points as $\mathcal{S}(\mathcal{Y})$. For $r \in [1, \infty]$, define the r-factor proximity map as follows. Given a point x in $\mathcal{S}(\mathcal{Y})$, let $Q_{\mathrm{y}}(M, x)$ be the polytope with vertices being the $d(d+1)/2$ points on the edges, the vertex y and x so that the faces of $Q_{\mathrm{y}}(M, x)$ are formed by $d-1$ line segments each of which joining one of $\mathcal{Y}$ points, say y_i, to M and that are between M and the face opposite y_i. That is, the vertex region for vertex v is the polytope with vertices given by v and such points on the edges. Let $v(x)$ be the vertex in whose region x falls. If x falls on the boundary of two vertex regions, we assign $v(x)$ arbitrarily. Let $\varphi(x)$ be the face opposite to vertex $v(x)$, and $\eta(v(x), x)$ be the hyperplane

parallel to $\varphi(x)$ which contains x. Let $d(v(x), \eta(v(x), x))$ be the (perpendicular) Euclidean distance from $v(x)$ to $\eta(v(x), x)$. For $r \in [1, \infty)$, let $\eta_r(v(x), x)$ be the hyperplane parallel to $\varphi(x)$ such that $d(v(x), \eta_r(v(x), x)) = r\, d(v(x), \eta(v(x), x))$ and $d(\eta(v(x), x), \eta_r(v(x), x)) < d(v(x), \eta_r(v(x), x))$. Let $\mathcal{S}_r(x)$ be the polytope similar to and with the same orientation as $\mathcal{S}(\mathcal{Y})$ having $v(x)$ as a vertex and $\eta_r(v(x), x)$ as the opposite face. Then the r-factor proximity region $N_{\mathcal{Y}}^r(x) := \mathcal{S}_r(x) \cap \mathcal{S}(\mathcal{Y})$. Also, let $\zeta_j(x)$ be the hyperplane such that $\zeta_j(x) \cap \mathcal{S}(\mathcal{Y}) \neq \emptyset$ and $r\, d(\mathrm{y}_j, \zeta_j(x)) = d(\mathrm{y}_j, \eta(\mathrm{y}_j, x))$ for $j = 1, 2, \ldots, d+1$. Then the Γ_1-region is $\Gamma_1^r(x) = \bigcup_{j=1}^{d+1}(\Gamma_1^r(x) \cap R_M(\mathrm{y}_j))$, where $\Gamma_1^r(x) \cap R_M(\mathrm{y}_j) = \{z \in R_M(\mathrm{y}_j) : d(\mathrm{y}_j, \eta(\mathrm{y}_j, z)) \geq d(\mathrm{y}_j, \zeta_j(x)\}$, for $j = 1, 2, \ldots, d+1$.

Let $X_\varphi := \mathrm{argmin}_{X \in \mathcal{X}_n} d(X, \varphi)$ be the closest point among $\mathcal{X}_n$ to face φ. Then it is easily seen that $\Gamma_1^r(\mathcal{X}_n, M) = \bigcap_{i=1}^{d+1} \Gamma_1^r(X_{\varphi_i}, M)$, where φ_i is the face opposite vertex y_i, for $i = 1, 2, \ldots, d$. So $\Gamma_1^r(\mathcal{X}_n, M) \cap R_M(\mathrm{y}_i) = \{z \in R_M(\mathrm{y}_i) : d(\mathrm{y}_i, \eta(\mathrm{y}_i, z) \geq d(\mathrm{y}_i, \Xi_i(X_{\varphi_i}))\}$, for $i = 1, 2, \ldots, d$.

Let the domination number be $\gamma_n(r, F, M, d) := \gamma_n(\mathcal{X}_n; F, N_{PE}^r, d)$ and the closest face extrema (if exists) be $X_{[i,1]} := \mathrm{argmin}_{X \in \mathcal{X}_n \cap R_M(\mathrm{y}_i)} d(X, \varphi_i)$. Then $\gamma_n(r, M) \leq d+1$ with probability 1, since $\mathcal{X}_n \cap R_M(\mathrm{y}_i) \subset N_{PE}^r\left(X_{[i,1]}, M\right)$ for each of $i = 1, 2, \ldots, d$.

In $\mathcal{S}(\mathcal{Y})$, drawing the hypersurfaces $Q_i(r, x)$ such that $d(\mathrm{y}_i, \varphi_i) = rd(\mathrm{y}_i, Q_i(r, x))$ for $i \in \{1, 2, \ldots, d\}$ yields another polytope, denoted as $\mathcal{P}_r$, for $r < (d+1)/d$. Let $\gamma_n(r, M, d) := \gamma(\mathcal{X}_n, N_{PE}^r, M, d)$ be the domination number of the PCD based on the extension of $N_{PE}^r(\cdot, M)$ to $\mathbb{R}^d$. Then we conjecture the following.

Conjecture 8.1. Suppose $\mathcal{X}_n$ is set of iid random variables from the uniform distribution on a simplex in $\mathbb{R}^d$. Then as $n \to \infty$, the domination number $\gamma_n(r, M, d)$ in the simplex satisfies

$$\gamma_n(r, M, d) \xrightarrow{\mathcal{L}} \begin{cases} d + \mathrm{BER}(1 - p_{r,d}) & \text{for } r \in [1, (d+1)/d) \text{ and} \\ & M \in \{t_1(r), t_2(r), \ldots, t_{d+1}(r)\}, \\ \leq (d-1) & \text{for } r > (d+1)/d \text{ and } M \in \mathcal{S}(\mathcal{Y})^o, \\ d+1 & \text{for } r \in [1, (d+1)/d) \text{ and} \\ & M \in \mathcal{P}_r \backslash \{t_1(r), t_2(r), \ldots, t_{d+1}(r)\}, \end{cases} \tag{15}$$

where $p_{r,d}$ can be calculated by intensive numerical integration as in the calculation of Eq. (3) and for $r = (d+1)/d$ and $M = M_C$, $p_{r,d}$ will be different from the continuous extension of Eq. (15).

9. Discussion and Conclusions

In this article, we consider the asymptotic distribution of the domination number of proportional-edge proximity catch digraphs (PCDs), for testing bivariate spatial point patterns of segregation and association. To our knowledge the PCD-based methods are the only graph theoretic tools for testing spatial patterns in literature (Ceyhan and Priebe, 2005; Ceyhan et al., 2006, 2007). The proportional-edge PCDs lend themselves for such a purpose, because of the geometry invariance property for uniform data on Delaunay triangles. Let the two samples of sizes n and m be from classes $\mathcal{X}$ and $\mathcal{Y}$, respectively, with $\mathcal{X}$ points being used as the vertices of the PCDs and $\mathcal{Y}$ points being used in the construction of Delaunay triangulation. For the domination number approach to be appropriate, n should be much larger

compared to m. This implies that n tends to infinity while m is assumed to be fixed. That is, the imbalance in the relative abundance of the two classes should be large for our method. Such an imbalance usually confounds the results of other spatial interaction tests. Furthermore, we can also use the normal approximation to binomial distribution for the domination number, provided n is much larger than m, with both sizes tending to infinity. Therefore, as long as $n \gg m \to \infty$, we can remove the conditioning on m.

The null hypothesis is assumed to be CSR of $\mathscr{X}$ points, i.e., the uniformness of $\mathscr{X}$ points in the convex hull of $\mathscr{Y}$ points. Although we have two classes here, the null pattern is not the CSR independence, since for finite m, we condition on m and the locations of the $\mathscr{Y}$ points (assumed not co-circular) are irrelevant. That is, the $\mathscr{Y}$ points can result from any pattern that results in a unique Delaunay triangulation.

There are many types of parametrizations for the alternatives. The particular parametrization of the alternatives in Eq. (6) is chosen so that the distribution of the domination number under the alternatives would be geometry invariant (i.e., independent of the geometry of the support triangles). The more natural alternatives (i.e., the alternatives that are more likely to be found in practice) can be similar to or might be approximated by our parametrization. Because in any segregation alternative, the $\mathscr{X}$ points will tend to be further away from $\mathscr{Y}$ points and in any association alternative $\mathscr{X}$ points will tend to cluster around the $\mathscr{Y}$ points. Such patterns can be detected by the test statistics based on the domination number, since under segregation (whether it is parametrized as in Sec. 4 or not) we expect them to be smaller, and under association (regardless of the parametrization) they tend to be larger.

By construction our method uses only the $\mathscr{X}$ points in $C_H(\mathscr{Y}_m)$ (the convex hull of $\mathscr{Y}$ points) which might cause substantial data (hence information) loss. To mitigate this, we propose a correction for the proportion of $\mathscr{X}$ points outside $C_H(\mathscr{Y}_m)$, because the pattern inside $C_H(\mathscr{Y}_m)$ might not be the same as the pattern outside $C_H(\mathscr{Y}_m)$. We suggest analysis with our domination number approach in two steps: (i) analysis restricted to $C_H(\mathscr{Y}_m)$, which provides inference only for $\mathscr{X}$ points in $C_H(\mathscr{Y}_m)$; (ii) overall analysis with convex hull correction (i.e., for all $\mathscr{X}$ points with respect to $\mathscr{Y}_m$). When the number of Delaunay triangles based on $\mathscr{Y}$ points, denoted J_m, is less than 30, we recommend the use of binomial distribution as $n \to \infty$ (i.e., for large n); when J_m is larger than 30, we recommend the use of normal approximation as $n \to \infty$. For small samples, one might use Monte Carlo simulation or randomization with our approach or apply a finite sample correction as in Remark 6.1. In the case of small samples with some $\mathscr{X}$ points existing outside $C_H(\mathscr{Y}_m)$, convex hull correction can be implemented first, and then the small sample correction. Furthermore, when testing against segregation we recommend the parameter $r \approx 1.3$, while for testing against association we recommend the parameter $r \approx 1.35$ as they exhibit the best performance in terms of size and power. The proportional-edge PCDs have applications in classification. This can be performed building discriminant regions in a manner analogous to the procedure proposed in Priebe et al. (2003a).

Acknowledgments

Supported by DARPA as administered by the Air Force Office of Scientific Research under contract DOD F49620-99-1-0213 and by ONR Grant N00014-95-1-0777 and by TUBITAK Kariyer Project Grant 107T647. We also thank anonymous

referees, whose constructive comments and suggestions greatly improved the presentation and flow of this article.

References

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

Ceyhan, E. (2005). An investigation of proximity catch digraphs in delaunay tessellations. Also available as technical monograph titled "Proximity Catch Digraphs: Auxiliary Tools, Properties, and Applications" by VDM Verlag, PhD thesis, The Johns Hopkins University, Baltimore, MD.

Ceyhan, E. (2008a). The distribution of the domination number of class cover catch digraphs for non-uniform one-dimensional data. *Discrete Mathematics* 308:5376–5393.

Ceyhan, E. (2008b). Overall and pairwise segregation tests based on nearest neighbor contingency tables. *Computat. Statist. Data Anal.* 53(8):2786–2808.

Ceyhan, E. (2009a). Class-specific tests of segregation based on nearest neighbor contingency tables. *Statistica Neerlandica* 63(2):149–182.

Ceyhan, E. (2009b). Spatial clustering tests based on domination number of a new random digraph family. arXiv:0909.3034 [math.ST]. *Technical Report # KU-EC-09-6*, Koç University, Istanbul, Turkey.

Ceyhan, E., Priebe, C. (2003). Central similarity proximity maps in Delaunay tessellations. *Proc. Joint Statist. Meeting, Statist. Comput. Sec. Amer. Statist. Assoc.*, San Francisco, CA, August 3–7.

Ceyhan, E., Priebe, C. E. (2005). The use of domination number of a random proximity catch digraph for testing spatial patterns of segregation and association. *Statist. Probab. Lett.* 73:37–50.

Ceyhan, E., Priebe, C. E. (2007). On the distribution of the domination number of a new family of parametrized random digraphs. *Model Assist. Statist. Applic.* 1(4):231–255.

Ceyhan, E., Priebe, C. E., Marchette, D. J. (2007). A new family of random graphs for testing spatial segregation. *Can. J. Statist.* 35(1):27–50.

Ceyhan, E., Priebe, C. E., Wierman, J. C. (2006). Relative density of the random r-factor proximity catch digraphs for testing spatial patterns of segregation and association. *Computat. Statist. Data Anal.* 50(8):1925–1964.

Chartrand, G., Lesniak, L. (1996). *Graphs & Digraphs*. Boca Raton, FL: Chapman & Hall/CRC Press LLC.

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

Cuzick, J., Edwards, R. (1990). Spatial clustering for inhomogeneous populations (with discussion). *J. Roy. Statist. Soc. Ser. B* 52:73–104.

DeVinney, J., Priebe, C. E. (2006). A new family of proximity graphs: Class cover catch digraphs. *Discr. Appl. Math.* 154(14):1975–1982.

DeVinney, J., Priebe, C. E., Marchette, D. J., Socolinsky, D. (2002). Random walks and catch digraphs in classification. *Proc. 34th Symp. Interface: Computing Science and Statistics*, Vol. 34. Available at: http://www.galaxy.gmu.edu/interface/I02/I2002 Proceedings/DeVinneyJason/DeVinneyJason.paper.pdf

DeVinney, J., Wierman, J. C. (2003). A SLLN for a one-dimensional class cover problem. *Statist. Probab. Lett.* 59(4):425–435.

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

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

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

Dixon, P. M. (2002b). Nearest neighbor methods. In: El-Shaarawi, A. H., Piegorsch, W. W., eds. *Encyclopedia of Environmetrics*. New York: John Wiley & Sons Ltd., Vol. 3, pp. 1370–1383.

Eeden, C. V. (1963). The relation between Pitman's asymptotic relative efficiency of two tests and the correlation coefficient between their test statistics. *Ann. Mathemat. Statist.* 34(4):1442–1451.

Fall, A., Fortin, M. J., Manseau, M., O'Brien, D. (2007). Ecosystems. *Int. J. Geograph. Inform. Sci.* 10(3):448–461.

Friedman, J. H., Rafsky, L. C. (1983). Graph-theoretic measures of multivariate association and prediction. *Ann. Statist.* 11(2):377–391.

Good, B. J., Whipple, S. A. (1982). Tree spatial patterns: South Carolina bottomland and swamp forests. *Bull. Torrey Botanical Club* 109:529–536.

Hodges, J. L. J., Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the *t*-test. *Ann. Mathemat. Statist.* 27(2):324–335.

Jaromczyk, J. W., Toussaint, G. T. (1992). Relative neighborhood graphs and their relatives. *Proc. IEEE* 80:1502–1517.

Keitt, T. (2007). Introduction to spatial modeling with networks. Presented at the *Workshop on Networks in Ecology and Beyond Organized by the PRIMES (Program in Interdisciplinary Math, Ecology and Statistics)* at Colorado State University, Fort Collins, Colorado.

Kendall, M., Stuart, A. (1979). *The Advanced Theory of Statistics*. 4th ed. Vol. 2. London: Griffin.

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

Marchette, D. J., Priebe, C. E. (2003). Characterizing the scale dimension of a high dimensional classification problem. *Pattern Recogn.* 36(1):45–60.

Minor, E. S., Urban, D. L. (2007). Graph theory as a proxy for spatially explicit population models in conservation planning. *Ecolog. Applic.* 17(6):1771–1782.

Okabe, A., Boots, B., Sugihara, K., Chiu, S. N. (2000). *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. New York: Wiley.

Perry, G., Miller, B., Enright, N. (2006). A comparison of methods for the statistical analysis of spatial point patterns in plant ecology. *Plant Ecol.* 187(1):59—-82.

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

Priebe, C. E., DeVinney, J. G., Marchette, D. J. (2001). On the distribution of the domination number of random class cover catch digraphs. *Statist. Probab. Lett.* 55:239–246.

Priebe, C. E., Marchette, D. J., DeVinney, J., Socolinsky, D. (2003a). Classification using class cover catch digraphs. *J. Classific.* 20(1):3–23.

Priebe, C. E., Solka, J. L., Marchette, D. J., Clark, B. T. (2003b). Class cover catch digraphs for latent class discovery in gene expression monitoring by DNA microarrays. *Computat. Statist. Data Anal. Visual.* 43-4:621–632.

Ripley, B. D. (2004). *Spatial Statistics*. New York: Wiley-Interscience.

Roberts, S. A., Hall, G. B., Calamai, P. H. (2000). Analysing forest fragmentation using spatial autocorrelation, graphs and GIS. *Int. J. Geograph. Inform. Sci.* 14(2): 185–204.

Stoyan, D., Penttinen, A. (2000). Recent applications of point process methods in forestry statistics. *Statist. Sci.* 15(1):61–78.

Su, W. Z., Yang, G. S., Yao, S. M., Yang, Y. B. (2007). Scale-free structure of town road network in southern Jiangsu Province of China. *Chin. Geograph. Sci.* 17(4): 311–316.

Toussaint, G. T. (1980). The relative neighborhood graph of a finite planar set. *Patt. Recogn.* 12(4):261–268.

West, D. B. (2001). *Introduction to Graph Theory*. 2nd ed. Englewood Cliffs, NJ: Prentice Hall.

Wierman, J. C., Xiang, P. (2008). A general SLLN for the one-dimensional class cover problem. *Statist. Probab. Lett.* 78(9):1110–1118.

Wu, X., Murray, A. T. (2008). A new approach to quantifying spatial contiguity using graph theory and spatial interaction. *Int. J. Geograph. Inform. Sci.* 22(4):387–407.

Xiang, P., Wierman, J. C. (2009). A CLT for a one-dimensional class cover problem. *Statist. Probab. Lett.* 79(2):223–233.