Testing Spatial Clustering Using Relative Density of Two Random Geometric Digraph Families

Elvan Ceyhan
Department of Mathematics, College of Sciences,
Koç University, 34450 Sarıyer, Istanbul, Turkey.
elceyhan@ku.edu.tr

Abstract

We compare the relative density of two parameterized random geometric digraph families called proportional edge and central similarity *proximity catch digraphs* (PCDs). In this article, we compare finite sample performance of the tests by Monte Carlo simulations and asymptotic performance by Pitman asymptotic efficiency. We find the optimal expansion parameters of the PCDs for testing each alternative in finite samples and in the limit as the sample size tends to infinity. As a result of our comparison, we demonstrate that in terms of empirical power (i.e., for finite samples) relative density of central similarity PCD has better performance (which occurs for expansion parameter values larger than one) under the segregation alternative, while relative density of proportional edge PCD has better performance under the association alternative. The methods are also illustrated in a real-life data set from plant ecology.

Keywords: association, complete spatial randomness, consistency, Delaunay triangulation, Pitman asymptotic efficiency, proximity catch digraphs, segregation, U-statistic

1 Introduction

Spatial clustering has received considerable attention in the statistical literature (Cressie (1993) and Diggle (2003)). Recently, the use of mathematical graphs has gained popularity in spatial analysis (Roberts et al. (2000)) although it potentially reduces the benefit of other geo-spatial information, since a graph ignores the geographic reference. However, graphs provide the means to go beyond the usual Euclidean metrics for spatial analysis. For example, graphs are potentially useful to ecological applications concerned with connectivity or movement. Furthermore, many concepts in spatial ecology depend on the idea of spatial adjacency which requires information on the close vicinity of an object. See Ceyhan (2011) and references therein for further information.

In recent years, a new clustering approach has been developed which uses vertex-random digraphs called proximity catch digraphs (PCDs) and is based on the relative positions of the data points from various classes. Priebe et al. (2001) introduced the class cover catch digraphs (CCCDs) which is a special type of PCDs and gave the exact and the asymptotic distribution of the domination number of the CCCDs in $\mathbb{R}$. The CCCD approach is extended to multiple dimensions by Marchette and Priebe (2003), and Priebe et al. (2003),who demonstrated relatively good performance of it in classification. The proportional edge and central similarity proximity maps are introduced based on the appealing properties CCCDs.

We describe the two particular PCD families in Section 2, provide the asymptotic distribution of relative density of the PCDs for uniform data in Section 3. We describe the alternative patterns of segregation and association, propose tests based on relative density of PCDs for testing segregation/association, provide the asymptotic normality and consistency of the tests under the alternatives, present the empirical size of the PCD tests, empirical power under the alternatives, and asymptotic efficiency in Ceyhan (2010). We present discussion and conclusions in Section 4.

2 The Proximity Map Families and the Associated PCDs

We first define proximity maps and PCDs in a fairly general setting. A pointed set is a pair (S, p) where S is a set and p a distinguished point. Then a catch digraph is a directed graph whose vertices are pointed sets with an arc from vertex $(N(u), u)$ to vertex $(N(v), v)$ whenever $v \in N(u)$. Hence $N(u)$ *catches* v. For PCDs, the sets $N(\cdot)$ are based on the proximity maps which are defined as follows. For a measurable space $(\Omega, \mathcal{M})$ with $\wp(\cdot)$ representing the power set function, given $\mathcal{Y}_m \subseteq \Omega$, the *proximity map* $N(\cdot) = N(\cdot, \mathcal{Y}_m) : \Omega \to \wp(\Omega)$ defines a *proximity region* $N(x) \subseteq \Omega$ for each point $x \in \Omega$. The region $N(x)$ is defined based on the dissimilarity between x and $\mathcal{Y}_m$. Then we define the vertex-random PCD, D, with vertex set $\mathcal{V} = \{X_1, X_2, \dots, X_n\}$ and arc set $\mathcal{A}$ by $(X_i, X_j) \in \mathcal{A} \iff X_j \in N(X_i)$. An extensive treatment of the proximity graphs is presented in Toussaint (1980) and Jaromczyk and Toussaint (1992).

For $\Omega = \mathbb{R}^d$ let $\mathcal{Y}_m = \{\mathsf{y}_1, \mathsf{y}_2, \dots, \mathsf{y}_m\}$ be m points in general position in $\mathbb{R}^d$. The space, $\mathbb{R}^d$, is partitioned by the Delaunay tessellation of class $\mathcal{Y}$ points which is the Delaunay triangulation in $\mathbb{R}^2$ provided that no more than three points in $\mathcal{Y}_m$ are cocircular (i.e., lie on the same circle). Then let T_i be the i^{th} Delaunay cell for $i = 1, 2, \dots, J_m$. 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 $\mathbb{R}^2$. For simplicity, we consider the one triangle case first. Let $\mathcal{Y}_3 = \{\mathsf{y}_1, \mathsf{y}_2, \mathsf{y}_3\}$ be three non-collinear points in $\mathbb{R}^2$ and $T(\mathcal{Y}_3) = T(\mathsf{y}_1, \mathsf{y}_2, \mathsf{y}_3)$ be the triangle with vertices $\mathcal{Y}_3$. Let $\mathcal{X}_n$ be a set of iid random variables from F with support $\mathcal{S}(F) \subseteq T(\mathcal{Y}_3)$ and $\mathcal{U}(T(\mathcal{Y}_3))$ be the uniform distribution on $T(\mathcal{Y}_3)$.

The **proportional edge proximity maps** are defined in detail in Ceyhan (2010); we provide the definition briefly here for the sake of completeness. For the expansion parameter $r \in [1, \infty]$, we define the *proportional edge* proximity map with expansion parameter r, denoted $N_{PE}(x, r)$ as follows; see also Figure 1 (left). Using line segments from the center of mass of $T(\mathcal{Y}_3)$ to the midpoints of its edges, we partition $T(\mathcal{Y}_3)$ into "vertex regions" $R_V(\mathsf{y}_1)$, $R_V(\mathsf{y}_2)$, and $R_V(\mathsf{y}_3)$. For $x \in T(\mathcal{Y}_3) \setminus \mathcal{Y}_3$, let $v(x) \in \mathcal{Y}_3$ be the vertex in whose region x falls, so $x \in R_V(v(x))$. If x falls on the boundary of two vertex regions, we assign $v(x)$ arbitrarily to one of the adjacent regions. Let $e(x)$ be the edge of $T(\mathcal{Y}_3)$ opposite $v(x)$. Let $\ell(x)$ be the line parallel to $e(x)$ through x. Let $d(v(x), \ell(x))$ be the Euclidean distance from $v(x)$ to $\ell(x)$. For $r \in [1, \infty)$, let $\ell_r(x)$ be the line parallel to $e(x)$ such that $d(v(x), \ell_r(x)) = r d(v(x), \ell(x))$ and $d(\ell(x), \ell_r(x)) < d(v(x), \ell_r(x))$. Let $T_{PE}(x, r)$ be the triangle similar to and with the same orientation as $T(\mathcal{Y}_3)$ having $v(x)$ as a vertex and $\ell_r(x)$ as the opposite edge. Then the *proportional edge* proximity region $N_{PE}(x, r)$ is defined to be $T_{PE}(x, r) \cap T(\mathcal{Y}_3)$. Notice that $r \geq 1$ implies $x \in N_{PE}(x, r)$.

The **central similarity proximity maps** were defined with expansion parameter $\tau \leq 1$

Below, we provide a definition for much wider range of the expansion parameter $\tau \in (0, \infty]$. Define $N_{CS}(x, \tau)$ to be the *central similarity proximity map* with expansion parameter τ as follows; see also Figure 1 (right). Let e_j be the edge opposite vertex y_j for $j = 1, 2, 3$, and let "edge regions" $R_E(e_1)$, $R_E(e_2)$, $R_E(e_3)$ partition $T(\mathcal{Y}_3)$ using line segments from the center of mass of $T(\mathcal{Y}_3)$ to the vertices. For $x \in (T(\mathcal{Y}_3))^o$, let $e(x)$ be the edge in whose region x falls; $x \in R_E(e(x))$. If x falls on the boundary of two edge regions we assign $e(x)$ arbitrarily. For $\tau > 0$, the central similarity proximity region $N_{CS}(x, \tau)$ is defined to be the triangle $T_{CS}(x, \tau) \cap T(\mathcal{Y}_3)$ with the following properties:

(i) For $\tau \in (0, 1]$, the triangle $T_{CS}(x, \tau)$ has an edge $e_\tau(x)$ parallel to $e(x)$ such that $d(x, e_\tau(x)) = \tau\, d(x, e(x))$ and $d(e_\tau(x), e(x)) \leq d(x, e(x))$ and for $\tau > 1$, $d(e_\tau(x), e(x)) < d(x, e_\tau(x))$ where $d(x, e(x))$ is the Euclidean distance from x to $e(x)$,

(ii) the triangle $T_{CS}(x, \tau)$ has the same orientation as and is similar to $T(\mathcal{Y}_3)$,

(iii) the point x is at the center of mass of $T_{CS}(x, \tau)$.

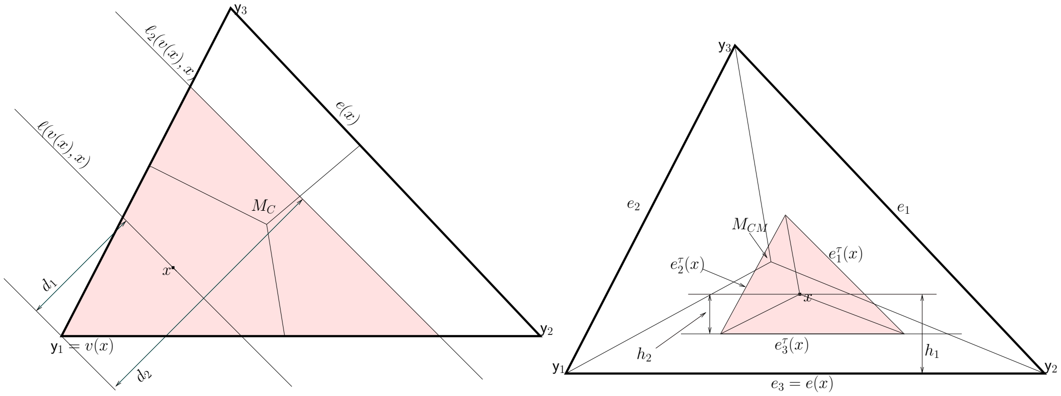

Figure 1: Plotted in the left is the illustration of the construction of proportional edge proximity region, $N_{PE}(x, r = 2)$ (shaded region) for an $x \in R_V(\mathsf{y}_1)$ where $d_1 = d(v(x), \ell(v(x), x))$ and $d_2 = d(v(x), \ell_2(v(x), x)) = 2\, d(v(x), \ell(v(x), x))$; and in the right is the illustration of the construction of central similarity proximity region, $N_{CS}(x, \tau = 1/2)$ (shaded region) for an $x \in R_E(e_3)$ where $h_2 = d(x, e_3^\tau(x)) = \frac{1}{2}\, d(x, e(x))$ and $h_1 = d(x, e(x))$.

3 The Asymptotic Distribution of Relative Density

The *relative density* of a digraph $D = (\mathcal{V}, \mathcal{A})$ of order $|\mathcal{V}| = n$, denoted $\rho(D)$, is defined as

$$\rho(D) = \frac{|\mathcal{A}|}{n(n-1)}$$

where $|\cdot|$ stands for set cardinality (Janson et al. (2000)). Thus $\rho(D)$ represents the ratio of the number of arcs in the digraph D to the number of arcs in the complete symmetric digraph of order n, which is $n(n-1)$. If $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$, then the relative density of the associated data-random PCD, denoted $\rho(\mathcal{X}_n; h, N)$, is shown to be a U-statistic

$$\rho(\mathcal{X}_n; h, N) = \frac{1}{n(n-1)} \sum\sum_{i<j} h_{ij} \qquad (1)$$

where

$$h_{ij} := h(X_i, X_j; N) = \mathbf{I}\{(X_i, X_j) \in \mathcal{A}\} + \mathbf{I}\{(X_j, X_i) \in \mathcal{A}\} = \mathbf{I}\{X_j \in N(X_i)\} + \mathbf{I}\{X_i \in N(X_j)\}.$$

Since the digraph is asymmetric, h_{ij} is defined as the number of arcs in D between vertices X_i and X_j, in order to produce a symmetric kernel with finite variance (Lehmann (1988)). Moreover, by a central limit theorem for U-statistics (Lehmann (1988)) it has been proved that

$$\sqrt{n}(\rho_n - \mathbf{E}[\rho_n]) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbf{Cov}[h_{12}, h_{13}]) \tag{2}$$

provided $\mathbf{Cov}[h_{12}, h_{13}] > 0$ where $\mathcal{N}(\mu, \sigma^2)$ stands for the normal distribution with mean μ and variance σ^2 and $\mathbf{E}[\rho_n] = \frac{1}{2}\mathbf{E}[h_{12}]$

For simplicity, we consider $\mathcal{X}$ points iid uniform in one triangle only. The extension to multiple triangles is presented in Ceyhan (2010). The null hypothesis is a type of *complete spatial randomness* (CSR); that is,

$$H_o : X_i \overset{iid}{\sim} \mathcal{U}(T(\mathcal{Y}_3)) \text{ for } i = 1, 2, \ldots, n. \tag{3}$$

The central limit theorem for U-statistics establishes the asymptotic normality under the uniform null hypothesis. For our proximity maps and uniform null hypothesis, the asymptotic null distribution of $\rho_{{}_{PE}}(n, r)$ (or $\rho_{{}_{CS}}(n, \tau)$) can be derived as a function of r (or τ). Let $\mu_{{}_{PE}}(r) := \mathbf{E}[\rho_{{}_{PE}}(n, r)]$ and $\nu_{{}_{PE}}(r) := \mathbf{Cov}[h_{12}, h_{13}]$. Notice that $\mu_{{}_{PE}}(r) = \mathbf{E}[h_{12}]/2 = P(X_2 \in N_{PE}(X_1, r))$ is the probability of an arc occurring between any pair of vertices, hence is called *arc probability* also. Similarly, let $\mu_{{}_{CS}}(\tau) := \mathbf{E}[\rho_{{}_{CS}}(n, \tau)]$, then $\mu_{{}_{CS}}(\tau) = P\big(X_2 \in N_{CS}(X_1, \tau)\big)$ and let $\nu_{{}_{CS}}(\tau) := \mathbf{Cov}[h_{12}, h_{13}]$.

By detailed geometric probability calculations, the means and the asymptotic variances of the relative density of the PCDs were calculated explicitly (see Ceyhan (2010)).

The forms of the mean functions are depicted together in Figure 2 (left). Note that $\mu_{{}_{PE}}(r)$ is monotonically increasing in r, since $N_{PE}(x, r)$ increases in size with r for all $x \in R_V(\mathsf{y}_j) \setminus \mathscr{R}_S(N_{PE}(\cdot, r), M_C)$, where $\mathscr{R}_S(N_{PE}(\cdot, r), M_C) := \{x \in T(\mathcal{Y}_3) : \; N_{PE}(x, r) = T(\mathcal{Y}_3)\}$. Note also that $\mu_{{}_{CS}}(\tau)$ is monotonically increasing in τ, since $N_{CS}(x, \tau)$ increases in size with τ for all $x \in R_E(e_j) \setminus \mathscr{R}_S(N_{CS}(\cdot, \tau), M_C)$, where $\mathscr{R}_S(N_{CS}(\cdot, \tau), M_C) := \{x \in T(\mathcal{Y}_3) : \; N_{CS}(x, \tau) = T(\mathcal{Y}_3)\}$. The asymptotic variance functions are depicted together in Figure 2 (right).

4 Discussion

We compare the relative density of two proximity catch digraphs (PCDs), namely, PE-PCDs and CS-PCDs for testing bivariate spatial patterns of segregation and association against complete spatial randomness (CSR). For finite samples, we assess the empirical size and power of the relative density of the PCDs by extensive Monte Carlo simulations. For the PE-PCDs, the optimal expansion parameters (in terms of appropriate empirical size and high power) are about 1.5 under mild segregation and values in $(2, 3)$ under moderate to severe segregation; and about 2 under association. On the other hand, for CS-PCDs, the optimal parameters are about 7 under segregation, and about 1 under association. Furthermore, we have shown that relative density of CS-PCDs has better empirical size performance; and

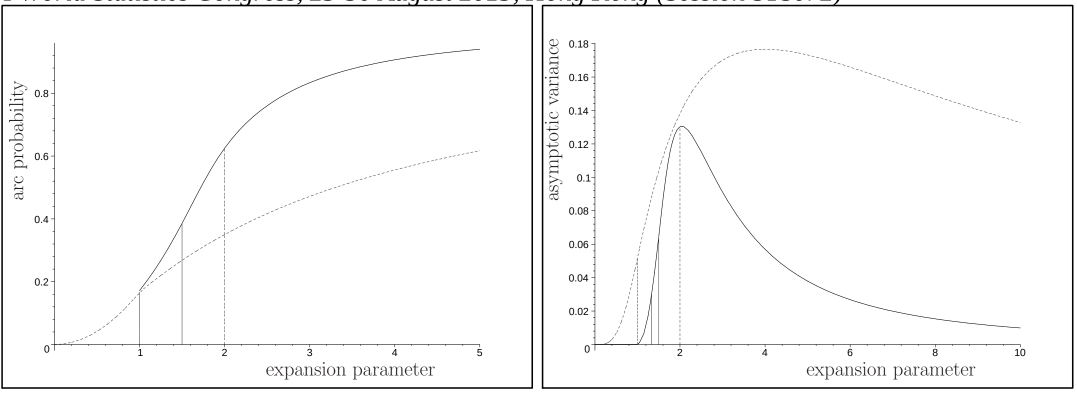

Figure 2: Asymptotic null means (i.e., arc probabilities) (left) and variances (right) as a function of the expansion parameters for relative density of PE-PCDs (solid line) and CS-PCDs (dashed line). The vertical lines indicate the endpoints of the intervals in the piecewise definition of the functions. Notice that the vertical and horizontal axes are differently scaled.

also, it has higher power against the segregation alternatives. On the other hand, relative density of PE-PCDs has higher power against the association alternatives.

For the two samples with 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. The null hypothesis is assumed to be CSR of $\mathcal{X}$ points, i.e., the uniformness of $\mathcal{X}$ points in the convex hull of $\mathcal{Y}$ points, $C_H(\mathcal{Y}_m)$. Although we have two classes here, the null pattern is not the CSR independence, since for finite m, we condition on relative areas of the Delaunay triangles based on $\mathcal{Y}$ points (assumed to have no more than three co-circular points). The relative density of the two PCD families lend themselves for spatial pattern testing conveniently, because of the geometry invariance property for uniform data on Delaunay triangles.

We also compare the asymptotic relative efficiency of the relative densities of the two PCD families. Based on Pitman asymptotic efficiency, we have shown that in general the relative density of PE-PCDs is asymptotically more efficient under segregation, while relative density of CS-PCDs is more efficient under association. However, this result is for $n \to \infty$ under very mild deviations from CSR. Besides for the above optimal expansion parameter values (optimal with respect to empirical size and power), the asymptotic efficiency and empirical power analysis yields the same ordering in terms of performance.

For the relative density approach to be appropriate, the size of $\mathcal{X}$ points (i.e., n) should be much larger compared to size of $\mathcal{Y}$ points (i.e., 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 to be appropriate. Such an imbalance usually confounds the results of other spatial interaction tests. Furthermore, by construction our method uses only the $\mathcal{X}$ points in $C_H(\mathcal{Y}_m)$ which might cause substantial data (hence information) loss. To mitigate this, we propose a correction for the proportion of $\mathcal{X}$ points outside $C_H(\mathcal{Y}_m)$, because the pattern inside $C_H(\mathcal{Y}_m)$ might not be the same as the pattern outside $C_H(\mathcal{Y}_m)$. We suggest a two-stage analysis with our relative density approach: (i) analysis for $C_H(\mathcal{Y}_m)$, which provides inference restricted to $\mathcal{X}$ points in $C_H(\mathcal{Y}_m)$, (ii) overall analysis with convex hull correction (i.e., for all $\mathcal{X}$ points with respect to $\mathcal{Y}_m$). We recommend the use of normal approximation if $n \approx 10 \times m$ or more, although Monte Carlo simulations suggest smaller n might also work here.

Acknowledgments

Supported 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. Most of the Monte Carlo simulations presented in this article were executed at Koç University High Performance Computing Laboratory.

References

Ceyhan, E. (2010). A comparison of two proximity catch digraph families in testing spatial clustering. arXiv:1010.4436v1 [math.CO]. Technical Report # KU-EC-10-3, Koç University, Istanbul, Turkey.

Ceyhan, E. (2011). Spatial clustering tests based on domination number of a new random digraph family. *Communications in Statistics - Theory and Methods*, 40(8):1363–1395.

Cressie, N. A. C. (1993). *Statistics for Spatial Data.* Wiley, New York.

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

Janson, S., Łuczak, T., and Ruciński, A. (2000). *Random Graphs.* Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons, Inc., New York.

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

Lehmann, E. L. (1988). *Nonparametrics: Statistical Methods Based on Ranks.* Prentice-Hall, Upper Saddle River, NJ.

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

Priebe, C. E., DeVinney, J. G., and Marchette, D. J. (2001). On the distribution of the domination number of random class cover catch digraphs. *Statistics & Probability Letters*, 55:239–246.

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

Roberts, S. A., Hall, G. B., and Calamai, P. H. (2000). Analysing forest fragmentation using spatial autocorrelation, graphs and GIS. *International Journal of Geographical Information Science*, 14(2):185–204.

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