A SCALE-SENSITIVE TEST OF ATTRACTION AND REPULSION BETWEEN SPATIAL POINT PATTERNS[*]

Tony E. Smith
Department of Electrical and Systems Engineering
University of Pennsylvania
Philadelphia, PA 19104

January 9, 2004

Abstract

There exist a variety of tests for attraction and repulsion effects between spatial point populations, most notably those involving either nearest-neighbor or cell-count statistics. Diggle and Cox (1981) showed that for the nearest-neighbor approach, a powerful test could be constructed using Kendall's rank correlation coefficient. In the present paper, this approach is extended to cell-count statistics in a manner paralleling the K-function approach of Lotwick and Silverman (1982). The advantage of the present test is that, unlike nearest-neighbors, one can identify the spatial scales at which repulsion or attraction are most significant. In addition, it avoids the torus-wrapping restrictions implicit in the Monte Carlo testing procedure of Lotwick and Silverman. Examples are developed to show that this testing procedure can in fact identify both attraction and repulsion between the same pair of point populations at different scales of analysis.

[*] The author is indebted to Sergio Rey for helpful comments on an earlier draft of this paper.

1. Introduction

There currently exist a variety of tests for the presence of attraction and repulsion effects between spatial point populations, most notably those involving either nearest-neighbor or cell-count statistics (as reviewed for example in Cressie, 1993, section 8.6). The advantage of *nearest-neighbor* approaches is that it is often possible to obtain exact (or at least asymptotic) distributions for certain test statistics under the null hypothesis of statistically independent populations. Most notable here is the approach of Diggle and Cox (1981), who showed that a powerful nearest-neighbor test of independence between two spatial point patterns could be constructed using Kendall's rank correlation coefficient. But by their very nature, nearest-neighbor statistics yield only a local description of point-pattern relationships. While it is theoretically possible to incorporate more global relationships by analyzing the joint distributions of say the first k nearest neighbors of points, there is general agreement that more powerful approaches can be constructed in terms of the *cross K-function* developed by Ripley (1976,1977). Most notable among these approaches is the Monte Carlo testing procedure developed by Lotwick and Silverman (1982). As with all K-function methods, this testing procedure yields a family of cell-count statistics at each scale specified by the user, and hence allows comparative analysis of pattern relationships over a range of relevant scales. However, (as discussed further below) this particular procedure is only applicable to a limited range of situations, and also suffers from certain theoretical problems that can lead to questionable results.

Hence the main objective of the present paper is to propose a testing procedure that combines certain positive features of each of these approaches.

2. Review of Two Testing Procedures

To motivate this procedure, it is convenient to begin by considering each of the above approaches in more detail. Consider a pair of point patterns $X_j = \{x_{ij} = (x_{1ij}, x_{2ij}) \in S : i = 1, .., n_j\}$, $j = 1, 2$ within some bounded region of the plane, $S \subset \mathbb{R}^2$. Statistically, these patterns are treated as a realization of some underlying bivariate point process on S. Each pattern X_j is thus a realization of the associated marginal point process. In this context, patterns X_1 and X_2 are said to be *independent* if and only if these marginal processes are statistically independent. To test for deviations from independence, we now consider each of the above testing procedures in turn.

2.1. Diggle-Cox Test

The procedure proposed by Diggle and Cox (1981) starts with the observation that if these marginal processes are independent, and if the random variables, d_1 and d_2, denote the distance from a random point in S to the nearest point in patterns X_1 and X_2, respectively, then d_1 and d_2 will also be independent random variables. This independence property can then be tested in a number of ways.

A simple nonparametric approach that does not rely on the particular distance values is Kendall's *rank-correlation coefficient*, τ. In particular, if for a sample of n random (uniformly distributed) *reference points* in S, we let $D_j = (d_{kj} : k = 1, .., n)$ denote the corresponding distances from each reference point k to the nearest point in pattern X_j , $j = 1, 2$, and for any pair of these reference points, k, h, let $s_{kh}^{(j)} = sign\,(d_{kj} - d_{hj})$,$j = 1, 2$, and let $s_{kh} = s_{kh}^{(1)} \cdot s_{kh}^{(2)}$, then $s_{kh} = 1(s_{kh} = -1)$ if and only if the ordering of nearest-neighbor distances to patterns $X_1$and X_2 is the same (opposite) for points k and h [with ties in either ordering being denoted by $s_{kh} = 0$]. In terms of these signed indices, Kendall's τ amounts to an appropriately scaled sum of signs:

$$\tau(D_1, D_2) = \frac{\sum_{k=1}^{n-1} \sum_{h=k+1}^{n} s_{kh}}{\sqrt{\left(\sum_{k=1}^{n-1} \sum_{h=k+1}^{n} |s_{kh}^{(1)}|\right) \left(\sum_{k=1}^{n-1} \sum_{h=k+1}^{n} |s_{kh}^{(2)}|\right)}} \tag{2.1}$$

From a testing viewpoint, this statistic has the advantage of being asymptotically normally distributed under the null hypothesis of independence, with mean zero and variance (Kendall, 1962, p.51):

$$\sigma^2 = \frac{1}{18} n(n-1)(2 * n + 5) \tag{2.2}$$

This not only provides a simple test of independence, but also provides one-sided tests of attraction versus repulsion between the two point patterns. In particular, if there is strong *attraction* between patterns, then for any random location in S, the presence (absence) of nearby points in one pattern will tend to imply the presence (absence) of nearby points in the other, so that the rank correlation, $\tau(D_1, D_2)$, between nearest-neighbor distances should be significantly *positive*. Similarly, if there is strong *repulsion* between patterns, then the presence of nearby points in one pattern will tend to imply the absence of nearby points in the other, and visa versa, so that $\tau(D_1, D_2)$, should be significantly *negative*. Diggle and Cox (1981) have shown that for simulated bivariate point processes exhibiting either attraction or repulsion, this testing procedure is more powerful than a range of selected competitors (including Rayleigh's test, the Wilcoxon test and a two-sample Kolmogorov-Smirnov test).

As a simple illustration of this test, consider the pair of point patterns illustrated in Figure 2.1 below, where pattern X_1 corresponds to the dots, pattern X_2 corresponds to the circles, and region S is defined by the square boundary shown. Here it seems clear that the these patterns are too much in agreement to be consistent with independence. Hence one expects to find significant attraction. However, closer inspection shows that the minimum spacing between these two patterns is too large to be random. In fact these patterns were simulated by a modified "simple inhibition process" (Diggle, 2003, section 5.6) in which Pattern X_1 is first generated randomly, and then Pattern X_2 is generated by locating a point randomly in the ring about each point in Pattern X_1 with inner radius, $r_1 = 5$, and outer radius, $r_2 = 8$. Rejection sampling is then used to ensure that the distance from each X_2 point to *all* points in X_1 is at least 5. Hence there is *small-scale repulsion* between these patterns at interpoint distances below 5, and *large-scale attraction* at interpoint distances above 8. [As one concrete illustration, imagine that these points represent the individual locations of plant species requiring similar soil conditions, but having incompatible root systems. Then these species might tend to appear together, but with only those plants surviving that have sufficient room to grow.]

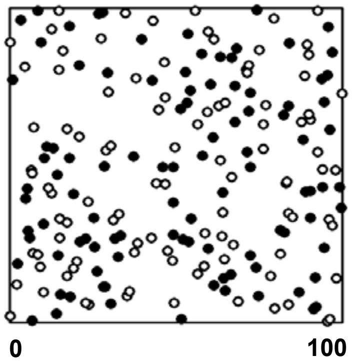

Figure 2.1: A Two-Population Pattern

To apply the Diggle-Cox Test here, an additional set of 200 reference points was randomly generated, and the rank correlation between nearest-neighbor distance rankings was computed to be $\tau(D_1, D_2) = -0.15839$. This negative correlation is highly significant, with P-value $< .0001$, and suggests that there is indeed small-scale repulsion between these patterns. This result serves to illustrate both the strength and weakness of the Diggel-Cox Test. For while this test is very powerful for discerning *small-scale relationships*, its reliance on nearest-neighbor statistics necessarily obscures relationships at larger scales. More generally it should be clear that structural differentiation between point patterns at different scales cannot be captured by any single statistic.

2.2. Lotwick-Silverman Test

As an alternative approach, Lotwick and Silverman (1982) proposed an application of cross K-functions to analyze pattern relationships over a range of spatial scales. In addition, they proposed a method for simulating a statistical population that is roughly consistent with the null hypothesis of independence. But unlike the simple hypothesis of "complete spatial randomness" used to test for clustering versus uniformity in single point patterns, it is virtually impossible to simulate exact independence between pairs of patterns without knowing their full marginal distributions. Rather than estimating these unknown marginals, Lotwick and Silverman attempt to preserve all marginal properties by considering only random shifts of one pattern relative to the other. To avoid the obvious boundary problems incurred in such a procedure, they assume that the boundary is rectangular (as in Figure 2.1) and then wrap this rectangle on a torus (donut) to eliminate the edges. One pattern can then be randomly shifted relative to the other on the surface of this torus. For our present purposes, however, it is useful to develop this process in terms of an equivalent version that is more easily seen graphically. In particular, random shifts on a torus can be equivalently represented by creating a mosaic of nine shifted copies of one pattern square, as shown in Figure 2.2 below, and randomly locating a copy of the second pattern square inside the boundary defined by the dotted lines.[1] Here the pattern of dots (X_1) in Figure 2.1 is used for the mosaic, and a possible random location of the pattern of circles (X_2) is shown by the heavy-line box inside the admissible region defined by the dotted lines.[2] If we denote this mosaic as pattern $\widetilde{X}_1$ (containing nine shifted copies of pattern X_1), then the key assumption of this approach is that the underlying

[1] For non-square rectangles, S, the dotted region continues to be defined by the centers of the corner rectangles of the mosaic, as shown in Figure 2.1.

[2] The pattern of circles in this box is not shown.

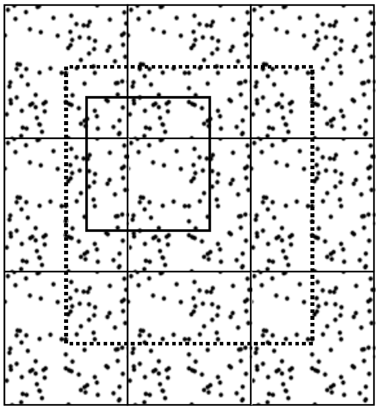

Figure 2.2: Mosaic for Random Shifts

marginal point processes are sufficiently *homogeneous* (stationary) across space to ensure that pattern $\widetilde{X}_1$ is a good statistical representative of marginal process 1 over this larger region.[3] If so, then one can treat the realization, X_2, of process 2 as being randomly embedded inside this larger realization, $\widetilde{X}_1$, of process 1. This not only yields independent joint realizations of the marginal processes, but also avoids many of the usual boundary problems by allowing comparisons of points in X_2 with points in $\widetilde{X}_1$ beyond its borders.

By sampling many random locations of X_2, one can in principle compute any type of test statistic for each sample, and thereby build up sampling distributions of these statistics. As mentioned above, Lotwick and Silverman employ sample estimates of Ripley's cross K-function in order to capture relationships at different scales. For our present purposes, it is enough to say that for each distance, $r > 0$, the estimated cross K-function value, $\widehat{K}_{12}(r)$, is proportional to the fraction of point pairs (x_{1i}, x_{2j}) from the respective patterns, X_1 and X_2, that are within distance r of each other. If this fraction is "larger than expected" at distance r under independence, then this suggests that there is some degree of attraction between members of X_1 and X_2 at *scale* r. Similarly, if they are "lower than expected" under independence, then there is some degree of *repulsion* at this scale.

To make these ideas precise, we first define $\widehat{K}_{12}(r)$ for any two patterns X_1 and

[3] More precisely it is assumed that these marginal processes are invariant under rigid motions of the plane.

X_2 as follows. If $I(E)$ denotes the indicator function for event, E, $[I(E) = 1$ if E is true and zero otherwise], and if we drop the constant factor of proportionality,[4] then we may set

$$\widehat{K}_{12}(r) = \frac{1}{n_1 n_2} \sum\nolimits_{i=1}^{n_1} \sum\nolimits_{j=1}^{n_2} I\left(\|x_{1i} - x_{2j}\| \le r\right) \tag{2.3}$$

Here r is restricted to be no greater than half the width of the square S, say $r(S)$.[5] Within this range, disks of radius r about any point in the randomly located small pattern X_2 must lie totally inside the mosaic pattern $\widetilde{X}_1$ (and hence exhibit no overlap on the torus).

Given this statistic, one can then estimate its sampling distribution under the null hypothesis of independence by computing its value for a number of random-shift simulations. First, for each simulation, $s = 1,..,N$, let the points of the randomly shifted X_2 pattern be denoted by $X_2^{(s)} = \{x_{2j}^{(s)} : j = 1,..,n_2\}$. Next observe that for each point x_{1j} in X_1 there are exactly nine copies of x_{1j} in $\widetilde{X}_1$. It should also be clear from Figure 2.2 that for each disk of radius $r \le r(S)$ about any point $x_{2j}^{(s)} \in X_2^{(s)}$, there is *at most one* copy of each X_1 point that is inside this disk. Hence if we now consider all pairs of points $\left(\widetilde{x}_{1i}, x_{2j}^{(s)}\right)$with $x_{2j}^{(s)} \in X_2^{(s)}$ and with $\widetilde{x}_{1i}$ denoting the copy of $x_{1i} \in X_1$ in $\widetilde{X}_1$ that is *closest* to $x_{2j}^{(s)}$, then in terms of these pairs, the relevant K-function values for each simulation, $s = 1,..,N$, are given by

$$\widehat{K}_{12}^{(s)}(r) = \frac{1}{n_1 n_2} \sum\nolimits_{i=1}^{n_1} \sum\nolimits_{j=1}^{n_2} I\left(\left\|\widetilde{x}_{1i} - x_{2j}^{(s)}\right\| \le r\right) \;\; , \;\; 0 < r \le r(S) \tag{2.4}$$

Thus $\widehat{K}_{12}^{(s)}(r)$ is again the fraction of those (relevant) pairs from each pattern that are within distance r of each other.[6] This is mathematically equivalent to

[4] The proportionality factor here is simply the area of region R. Note that while the averaging factor, $1/(n_1 n_2)$, could also be dropped here, it is kept in order to facilitate the interpretation of the statistic.

[5] More generally, if S is a rectangular box with length, l, and width, w, then the maximum radius is defined to be $r(S) = \min(l/2, w/2)$.

[6] The advantage of this definition is that it preserves the original pattern sizes n_1 and n_2. However, one could also define $\widehat{K}_{12}^{(s)}(r)$ directly in terms of the two patterns $\widetilde{X}_1$ and $X_2^{(s)}$ as

$$\widehat{K}_{12}^{(s)}(r) = \frac{1}{\widetilde{n}_1 n_2} \sum\nolimits_{i=1}^{\widetilde{n}_1} \sum\nolimits_{j=1}^{n_2} I\left(\left\|\widetilde{x}_{1i} - x_{2j}^{(s)}\right\| \le r\right)$$

the values obtained by shifts on the torus, but allows a somewhat simpler two-dimensional interpretation.

Finally, recalling that the initial X_2 pattern is precisely the "non-shifted" copy, $X_2^{(0)}$, lying on the center square of the mosaic, we may now compute the *observed* cross K-function values, $\widehat{K}_{12}^{(0)}(r)$, by simply extending (2.4) to the case $s = 0$. The question is then whether this value is typical of the population defined by the sample values $\{\widehat{K}_{12}^{(s)}(r) : s = 1, .., n\}$. To answer this question, one may simply rank these values and determine whether or not $\widehat{K}_{12}^{(0)}(r)$ is an extreme value in this ranking. More precisely, if $\widehat{K}_{12}^{(0)}(r)$ were just another random-shift sample, then the probability that it would be at least the m^{th} biggest in this list of $N+1$ values is simply $m/(N+1)$. This is precisely the P-value for a one-sided test of independence against the alternative hypothesis of "attraction" between patterns, which we now designate as the *attraction P-value*, $P_{att}(r)$. For example, if $N = 99$ and say $m = 4$ then the chance of getting a value as large as $\widehat{K}_{12}^{(0)}(r)$ is estimated to be $P_{att}(r) = 4/100 = .04$, providing evidence of significant attraction between these patterns at scale r. Similarly, if $m = 96$, then by reversing the tests and considering a one sided test of independence again the alternative hypothesis of "repulsion", one would obtain a *repulsion P-value* of $P_{rep}(r) = 1 - P_{att}(r) = 1 - .96 = .04$, thus indicating significant repulsion at scale r.

To see how this works in the example of Figure 2.1 above, a series of $N = 999$ simulations produced attraction P-values at selected distance, r, between 0 and 30 [$< 50 = r(S)$ in Figure 2.1].[7] These P-values are plotted in Figure 2.3 below. Here the lower dashed line denotes significant *attraction* P-values (below .05). Similarly the upper dashed line denotes attraction P-values above .95, corresponding to significant *repulsion* P-values (below .05). As with the Diggle-Cox test, this test shows significant small-scale repulsion. But now one identify the *scale* at which this repulsion occurs, namely at distances around 5. Equally important is the fact that significant *attraction* can now be observed for distances starting at around 8. So this testing procedure clearly yields more information about the overall

where $\widetilde{x}_{1i}$ is now an arbitrary element of the mosaic $\widetilde{X}_1$. It can readily be verified that all values in this summation are zero except those appearing in expression (2.4). Hence this definition differs from (2.4) only by the proportionality factor $n_1/\widetilde{n}_1 = 1/9$.

[7] At this point it should be emphasized that each set of simulated values will of course produce slightly different results. However for simulations of the order of 1000 samples, the variation between sample runs turns out to be minimal, with essentially the same regions of significance. This is true of all subsequent simulations in this paper. Hence the sampling standard adopted is $N = 999$, and the instances shown are taken to be typical representatives.

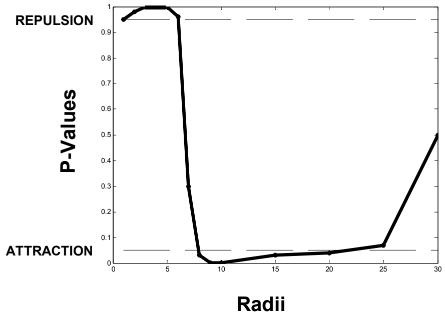

Figure 2.3: P-Values for the Lotwick-Silverman Test

structural relation between these patterns, and in this case turns out to reflect the underlying probability model with remarkable accuracy.

However, this random-shift (or torus-wrapping) procedure is not without its drawbacks. As Diggle (2003, section 1.3) observes, the juxtaposition of rectangular regions in a mosaic can sometimes create unintended structure in the point patterns. This is well illustrated by a second data set involving patterns of both healthy and diseased myrtle trees shown in Figure 2.4 below.[8] Here, healthy myrtles are shown by circles and diseased myrtles are shown by dots. It is evident

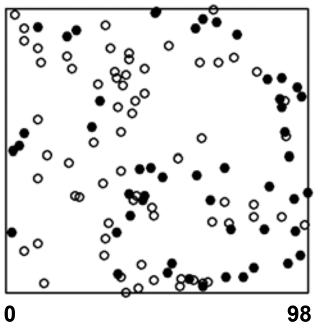

Figure 2.4: Healthy and Diseased Myrtles

[8] This data is also taken from Diggle, and is part of a larger data set available on his web site: *http://www.maths.lancs.ac.uk/˜diggle/pointpatterns/Datasets.*

from the figure that diseased trees tend to occur in clumps (suggesting perhaps that there is some local contagion in the spread of this disease). Hence one would expect to find some degree of *repulsion* between these point patterns. However, an analysis of this data using the Lotwick-Silverman test (again with $N = 999$) produced the results shown in Figure 2.5 below. Here there is no evidence of significant repulsion. Part of the reason for this can be seen by examining the mosaic generated by this pattern, as shown in Figure 2.6 below. Here only the the diseased trees are shown. First notice from Figure 2.4 above that many of the diseased clumps of trees happen to be near the border of region S. In Figure 2.6 this is seen to have the effect of a creating a number of larger clumps (such as those

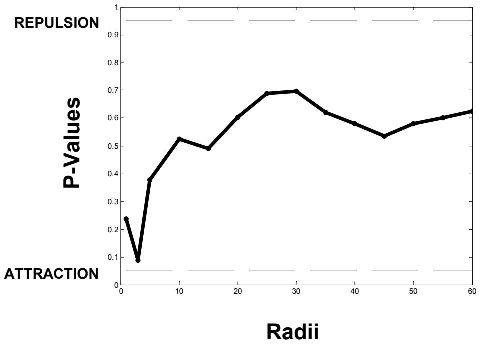

Figure 2.5: Random-Shift Analysis of Myrtles

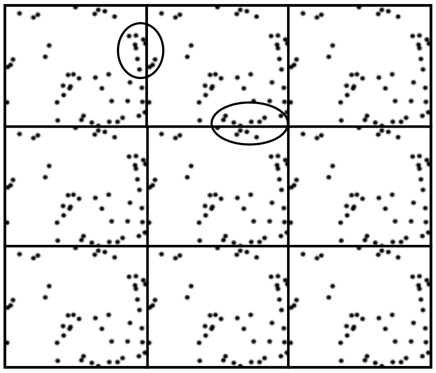

Figure 2.6: Mosaic for Myrtles

circles in the figure). Hence the mosaic pattern of diseased trees tends to look even more "clumply" than the original pattern. This has the effect of increasing the clumpiness of point counts for random shifts, thus making the clumpiness of the original point counts look *less significant*. This example thus serves to show that while torus-wrapping can in principle eliminate certain boundary problems, it can also have the effect of creating new ones.

3. A Combined Testing Procedure

Given these preliminary observations, we now consider a new testing procedure which is designed to combine certain positive features of both the Diggle-Cox test and the Lotwick-Silverman test. As in Diggle-Cox, this procedure starts with the generation of a random set of reference points with respect to which both patterns X_1 and X_2 can be described in a symmetric way. But rather than recording nearest-neighbor distances, the present procedure focuses on cell counts over a range of distance values as in Lotwick-Silverman. Hence rather than looking at rank correlations between nearest-neighbor pairs, this approach looks at the rank correlations between cell-count pairs at each relevant distance. These correlations constitute the basic *test statistics* for the procedure.

The most difficult part of any testing procedure for analyzing relationships between point patterns is to formulate a null hypothesis of "no relationship" that can be tested without additional prior knowledge about the underlying processes generating these patterns. Here we adopt a less ambitious approach than that of Lotwick and Silverman. Rather than attempt to simulate full statistical independence between these patterns, we focus on a conditionalized framework in which the set of $n_1 + n_2$ joint locations for all points is assumed to be fixed. In this more restrictive setting, our hypothesis of "no relationship" is simply that pattern X_1 is equally likely to occupy any n_1-element subset of these locations (or equivalently, that pattern X_2 is equally likely to occupy any complementary subset of size n_2). One can easily simulate the sampling distribution of rank correlations, τ, under this hypothesis by determining τ for a number of randomly relabelled pattern pairs. If the observed rank correlation is "higher than expected" under this distribution (so that cell-count rankings are similar for each pattern), then it can be inferred that these points tend to appear together, i.e., that there is significant *attraction* between patterns. Similarly, if the rank correlation is "lower than expected" then it may be inferred that these points tend not to appear together, and hence that there is significant *repulsion* between patterns.

Before specifying this procedure in detail, it is important to emphasize that the procedure itself is *not* a formal test of statistical independence between patterns X_1 and X_2. In particular, the null hypothesis above makes no attempt to preserve the marginal distributions of each pattern. So for example, if it is known that the marginal process generating X_1 tends to produce significantly more clustering than the X_2 process, and if it is desirable to preserve these properties in testing independence between X_1 and X_2, then the present procedure is not appropriate. However, if one is simply interested in whether points in X_1 tend to be closer to (or further from) points in X_2 than would be expected if the patterns were unrelated, then the term "unrelated" can be given useful operational meanings other than "statistical independence". A key advantage of the present approach is that it requires no prior knowledge of how these patterns were generated. Other advantages (and limitations) will be discussed in the applications below. But first we develop this testing procedure in more detail.

3.1. Steps of the Procedure

Given two point patterns, $X_j = \{x_{ij} = (x_{1ij}, x_{2ij}) \in S : i = 1, .., n_j\}$, $j = 1, 2$ in a bounded region of the plane, $S \subset \mathbb{R}^2$, the steps of the testing procedure can be outlined as follows:

(i) First generate a set of n randomly sampled *reference points*, $Z = \{z_i = (z_{1i}, z_{2i}) \in S : i = 1, .., n\}$, as in Diggle-Cox. [In the applications below, n is typically chosen to be roughly the same size as the joint pattern, i.e., $n \approx n_1 + n_2$.]

(ii) Next select a set of *reference distances* (or radii), $R = \{r_1 < r_2 < \cdots < r_k\}$, as in Lotwick-Silverman. [To minimize boundary effects, the maximum distance, r_k, is typically not more than half the maximum interpoint distance between X_1 and X_2.][9]

(iii) For each distance, $r \in R$, and reference point, $z_i \in Z$, count the number of points from pattern X_j within distance r of z_i, i.e., calculate

$$C_{ij}(r) = \sum\nolimits_{t=1}^{n_i} I\left(\|x_{tj} - z_i\| \le r\right) \tag{3.1}$$

and let the n-vector of these cell counts for each pattern $j = 1, 2$ be denoted by $C_j(r) = [C_{ij}(r) : i = 1, .., n]$. [These *cell-count profiles* will form the basis of comparison between patterns X_1 andX_2 (with respect to Z) at each scale r.]

[9] It should be noted at this point that the "non-overlapping" restriction on r for torus wrappings is no longer relevant in the present procedure. Indeed, it will be seen below that this procedure can in principle detect significant attraction and/or repulsion at distances well beyond "half the maximum interpoint distance".

(iv) As in (2.1) above, calculate the *rank correlation* between $C_1(r)$ and $C_2(r)$ and set

$$\tau_{0r} = \tau\left[C_1(r), C_2(r)\right] \quad , \quad r \in R \tag{3.2}$$

(v) To test whether τ_{0r} is significantly positive or negative, simulate N new pattern pairs by randomly switching their labels. To do so, let the *combined pattern*, $X = \{x_1, .., x_{n_1}, .., x_{n_1+n_2}\}$ be defined by $x_i = x_{i1}$ for $i = 1, .., n_1$, and $x_i = x_{i-n_1,2}$ for $i = n_1 + 1, .., n_1 + n_2$, and construct $s = 1, .., N$ random permutations,

$$(1, .., n_1, n_1 + 1, ..n_1 + n_2) \rightarrow (\pi_{s1}, .., \pi_{sn_1}, \pi_{s,n_1+1}, .., \pi_{s,n_1+n_2}) \tag{3.3}$$

of the integers $\{1, .., n_1 + n_2\}$. The corresponding *random-pattern pair*,(X_{s1}, X_{s2}), for each permutation s is then given by

$$\begin{aligned} X_{s1} &= \{x_{\pi_{si}} : i = 1, .., n_1\} \\ X_{s2} &= \{x_{\pi_{si}} : i = n_1 + 1, .., n_1 + n_2\} \end{aligned} \tag{3.4}$$

(vi) Now replace the observed patterns (X_1, X_2) with each pair of randomly permuted patterns (X_{s1}, X_{s2}), and repeat steps (iii) and (iv) to obtain a set of rank-correlation coefficients $\{\tau_{sr} : s = 1, .., N\}$ at each distance $r \in R$.

(vii) Finally, to construct the desired P-values for the "no relationship" test at each distance r , let m_r denote the number of τ-values in $\{\tau_{sr} : s = 1, .., N\}$ that are *greater* than τ_{0r}, and let m_{0r} denote the number that are *equal* to τ_{0r}. The appropriate *attraction P-value* is then defined as follows:[10]

$$P_{att}(r) = \frac{m_r + (m_{0r}/2)}{N + 1} \tag{3.5}$$

and the corresponding *repulsion P-value* is defined by $P_{rep}(r) = 1 - P_{att}(r)$. These P-values can then be plotted as in Figures 2.3 and 2.5 above.

3.2. Test Applications

The ultimate value of any testing procedure depends on how well it behaves in practice. Since the main objective of the present test is to reveal structural differentiation between pattern relationships at different scales, the best way to gage its effectiveness is to apply it to a range of data sets that may exhibit such structural variations. We begin with the examples above, and then consider one application to a larger urban data set.

[10] The value in the numerator has the effect of treating each τ-value tied with τ_{0r} as being "half above" and "half below" τ_{0r} in the overall ranking of τ-values.

3.2.1. The Simulated Model with Small-Scale Repulsion and Large-Scale Attraction

Recall that first pair of patterns above was generated from a simple probabilistic model exhibiting both small-scale repulsion and large-scale attraction. In view of the simplicity of this model, one could in principle attempt to estimate the power of this test to reject the "no relationship" hypothesis at a selected range of r_1 and r_2 values. But such power calculations would tell us little about the ability of this test to capture overall structural variations in pattern relationships at different scales. Hence we choose rather to simply examine the P-value plots obtained for selected cases (as in Figures 2.3 and 2.5 above), and ask whether the attraction and repulsion relationships exhibited seem reasonable. In the present case, we focus on the specific instance of this model ($r_1 = 5$ and $r_2 = 8$) used to illustrate the Diggle-Cox and Lotwick-Silverman tests above. Here we again use $n = 200$ reference points (as section 2.1) and a range of reference distances, r, between 1 and 30 (as in section 2.2). The results for a sample of $N = 999$ random permutations are shown in Figure 3.1 below. As with the Lotwick-Silverman test, we again see significant repulsion at distances around $r_1 = 5$ and attraction and distances beyond $r_2 = 8$.

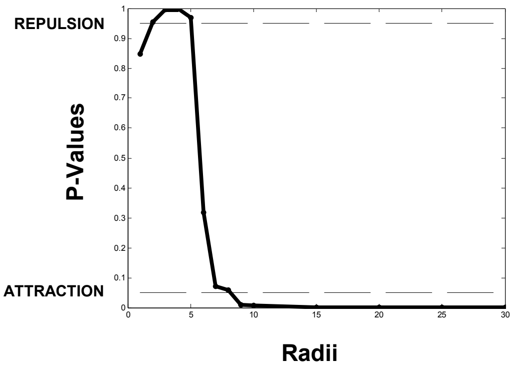

Figure 3.1: P-Values for the Patterns in Figure 2.1

However there are some differences. Notice first that repulsion is not significant at very small distances ($r = 1, 2$) even though there must be significant repulsion at these scales for the model itself. The reason this is not picked up relates to the density of the process simulated. For $n_1 = n_2 = 100$, it turns out that there are no reference points with cell counts above one at these small scales, so that there is no chance for any agreement between cell counts except for zero cell counts. This has the effect of making all cell profiles for random patterns look negatively correlated, so that the negative correlation for the observed pattern is *not significant*. So at very small scales relative to average point spacing, the present test can be misleading.

At the other extreme, notice that unlike Lotwick-Silverman, significant attraction persists out to $r = 30$. Moreover, simulations show that this attraction continues to be significant well beyond the usual upper limit of "half the maximum interpoint distance". The key reason for this is that each point of pattern X_2 in the present model is by construction close to at least one point in X_1 (with distance not exceeding 8). This has the effect of producing a relatively uniform population mix at large scales. But random relabeling of these populations will tend to exhibit some clumps at these scales, so that even for cell counts including most of the points in the combined pattern, the present model continues to exhibit agreement between cell-count profiles that is statistically significant.

3.2.2. The Healthy and Diseased Myrtles Patterns

Recall from Figure 2.4 that there appeared to be some significant clumpiness of diseased trees that was not be picked up by the Lotwick-Silverman test (Figure 2.5). When the present test was applied with $N = 999$ random permutations, the results yielded the P-value plot shown in Figure 3.2 below. Here it is clear that

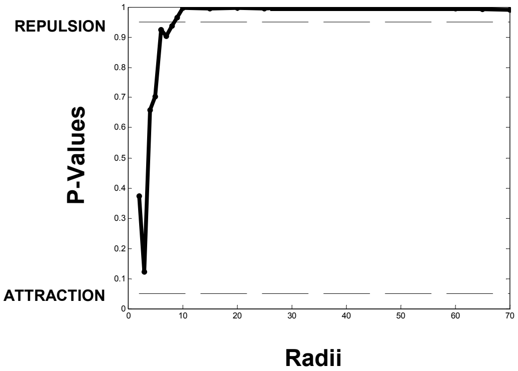

Figure 3.2: P-Value Plot for Myrtles

while the results are similar to Lotwick-Silverman for small distances ($r \leq 5$), they are dramatically different for all larger distances. Here there is now significant repulsion at larger scales which is consistent with the clumpiness observed visually. As in the example above, this repulsion continues to be significant at scales approaching the full width of region S, and again suggests that this testing procedure may be less sensitive to boundary effects than Lotwick-Silverman.

This example also serves to illustrate an important feature of the present conditional testing framework. By fixing the set of point locations, this hypothesis *excludes* all other locations for pattern points. This has some advantages, but also some weaknesses. On the positive side, it avoids the need for any boundary restrictions (such as the "rectangularity" restriction implicit in torus-wrapping schemes). More generally, it assumes only that pattern points in X_1 can occur everywhere that points in X_2 can occur, and visa versa. Suppose for example that the "holes" shown by dashed lines in Figure 3.3 below are rocky areas where no trees can grow. Then by considering only random permutations of "viable" tree locations, this test avoids all potentially non-viable locations such as rocky areas.

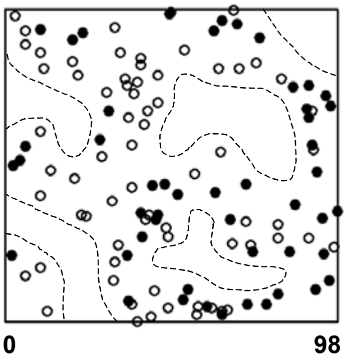

Figure 3.3: Holes in the Myrtle Landscape

On the other hand, if these blank areas turn out to be equally well suited for trees, then their presence surely indicates some degree of attraction between these populations. In the present case, since every diseased myrtle was presumably once a healthy myrtle, there must surely be some degree of attraction between these populations that is not being accounted for. Hence in this type of conditioned framework, it is important to keep in mind the types of relevant locational variation that may be excluded.

3.2.3. Supermarkets and Convenience Stores in Philadelphia

Our final example involves the locations of grocery stores in the city of Philadelphia. The key question here relates to the locational pattern of larger stores (supermarkets) versus smaller (convenience) stores. Since there is no absolute distinction between supermarkets and convenience stores, the present populations were defined simply in terms of floor space. Stores of at least 5000 sq.ft. were classified as "supermarkets" and those of less than 5000 sq.ft. were classified as "convenience stores".[11] In Figure 3.4 below the locations of supermarkets ($n_1 = 174$) are shown by dots, and the locations of convenience stores ($n_2 = 284$) are shown by circles. The enlarged portion shows the denser area of stores near the center of Philadelphia. Here supermarkets are seen to be reasonably well mixed with convenience stores, except for a few noticeable clumps (in West Philadelphia and in Northeast Philadelphia). A random pattern of $n = 400$ reference points was generated, and a test with $N = 999$ random permutations produced the P-value plot shown in Figure 3.5 below, for a set of reference distances up to about one mile. Here there appears to be some significant repulsion between these store types at distances of around 1000 feet and significant attraction at about twice that distance.

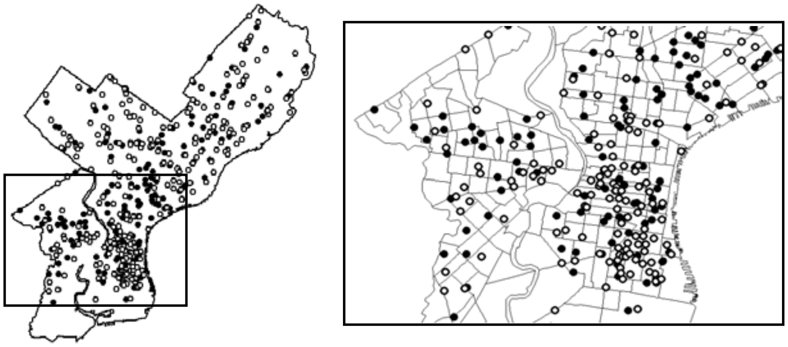

Figure 3.4: Supermarkets and Convenience Stores in Philadelphia

Before attempting to interpret these results, it is important to observe that

[11] In Philadelphia the two major covenience-store chains are Wawa and 7 Eleven. Since the largest of these stores is just below 5000 sq.ft., this size was used as the dividing line.

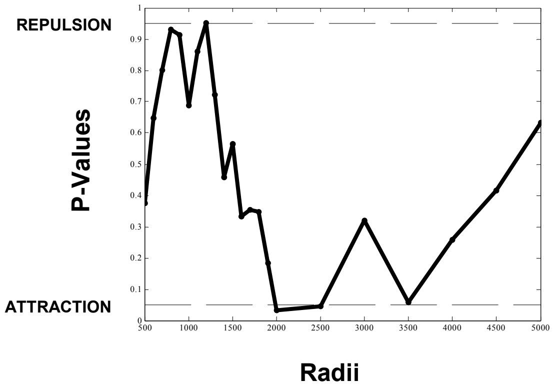

Figure 3.5: P-Value Plot with Purely Random Reference Points

in the present case we actually have additional information about "viable" store locations. In particular, the location of grocery stores is obviously influenced by the *population distribution*. Moreover, since population data is readily available, it is appropriate to consider ways of using this information. While it would in principle be possible to broaden the range of potential store locations by postulating that likely locations are proportional to population density, this approach is hampered by other considerations influencing store locations (such as zoning restrictions and locations of major roads).

However, such restrictions do not apply to the *reference point* distribution. Here one can consider randomly selected "customers" as the relevant reference points, and ask how many supermarkets and/or conveniences stores are within various distances of these customers. This not only yields more meaningful reference points from a behavioral viewpoint, but also has the statistical advantage of generating higher reference-point densities in areas where there are likely to be more stores. This second approach was operationalized by sampling $n = 400$ reference points from a probability distribution proportional to Philadelphia census-tract population densities, as shown in the center of Figure 3.6 below. The sample produced is shown on the right hand side, and the original pure random sample is shown on the left. Notice that the new reference points tend to be relatively less concentrated in those areas with lower population density, as seen for example at the southern tip of Philadelphia where population densities are very low.

The test above was rerun for $N = 999$ random permutations using these new reference points, and produced the P-value plot shown in Figure 3.7 below. Here

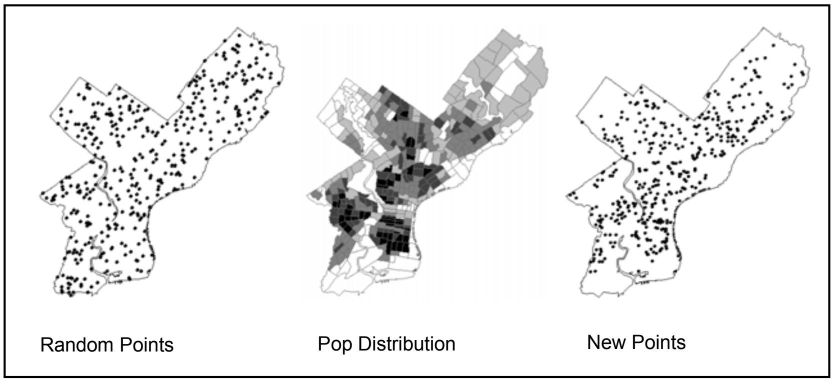

Figure 3.6: Reference Points for Philadelphia

it is apparent that the significant repulsion and attraction seen with respect to the random reference points at distances below 2500 has disappeared. While there are still some strong fluctuations in P-values at these distances, they are far less extreme than in the purely random case. In addition, there now appears to be significant repulsion at a distance of one mile (reflecting the types of clumps observed in Figure 3.4 above). Simulations at larger scales show that in fact *both* tests find significant repulsion in the one-mile to two-mile range. Hence the main purpose of the present example is to point out the dramatic differences that can occur at smaller distances.

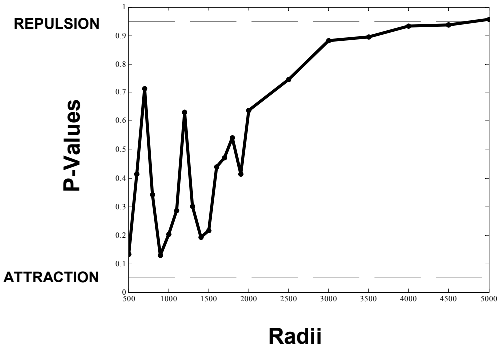

Figure 3.7: P-Values for Population-Generated Reference Points

A key reason for this difference can be seen by again considering those random reference points at the southern tip of Philadelphia, where there is little population and *no stores*. Though the scale is not shown on the map in Figure 3.4, the width of the enlarged box is approximately 10 miles. So it should be clear that there are no stores with 2500 feet of any of these points. More generally, all purely random reference points falling in very low population areas will tend to have zero cell counts for both supermarkets and convenience stores at these distance ranges. Since these null pairs are considered as "ties" in the rank-correlation measure, τ, they are essentially dropped from consideration. In view of the large number of such occurrences at these distance ranges, the rank-correlation value for the purely random case is in fact determined by a much smaller set of reference points than in the population-sensitive case. Statistically this has the effect of increasing the variation of τ, thus tending to produce more extreme values. Hence the significance results for random reference points at these scales are largely an artifact of the "holes" in the store-location pattern, and at the very least, are more difficult to interpret than those for the population-sensitive case.

While it is difficult to draw any general conclusions based on this example alone, it nonetheless serves as a useful "cautionary tale" in constructing tests of this type. The key point here is that it is vital to use as much additional information about the given point patterns as is possible.

4. Concluding Remarks

The purpose of this paper has been to propose a combination of the Diggle-Cox and Lotwick-Silverman tests for analyzing attraction-repulsion relationships between patterns at different scales. While the examples above suggest that this method has promise, they of course offer no definitive conclusions.

It should be noted that a number of variations on this general testing scheme are possible. First, while the use of rank-correlation to compare cell-count profiles has the advantage of being independent of relevant point densities, one could also accomplish this by looking at product-moment correlations, or even mean cell-counts normalized by density estimates, as is typically done in K-function statistics. Moreover, rather than using purely random reference points, one could consider a number of conditional randomization schemes that are meaningful even when no additional information (such as population densities) is available. For example, one could in principle avoid oversampling in empty regions by rejection-sampling schemes in which only those reference points "sufficiently close" to at least one pattern point are kept. Even for the pattern points themselves, one could in principle develop a number of resampling schemes other than random relabeling. For example, when no other reference distributions are available, one could use the empirical distribution defined by the combined point pattern as a basis for bootstrap resampling procedures.

It should also be recognized that attraction-repulsion comparisons are only one type of inter-pattern relationship that can be quantified. For example the use of circular cell counts implicitly assumes that pattern relationships are independent of direction, i.e., are *isotropic* in nature. Hence, as an extension of the present attraction-repulsion comparisons, one could ask whether these differ with respect to direction as well as scale. In addition, there are a host of other relevant comparisons that can be made directly in terms of marginal distributions, such as the relative clustering or dispersion within each pattern. With respect to the present testing procedure, it would be particularly desirable to develop testable null hypotheses that preserve at least some of this marginal information while at the same time avoiding the creation of unintended structure, such as that encountered in the Lotwick-Silverman test.

Finally it should be noted that the analysis above has been developed entirely in terms of *global* attraction and repulsion between point patterns. This naturally raises the question as to whether it is meaningful to consider *local* attraction and repulsion between patterns. One simple answer is that the relevant region S can always be partitioned into subregions, $(S_1, .., S_k)$, and the present analysis carried out in each subregion (provided that they are not too small to allow meaningful rank correlations). But perhaps a more interesting question relates to whether or not it meaningful to define *pointwise* attraction and repulsion, in a manner paralleling standard measures of local versus global clustering and dispersion within a single point pattern. Here the answer appears to be somewhat more problematic. For example, if a given pair of cell counts $[C_{i1}(r), C_{i2}(r)]$ are both zero, then one may ask whether or not this similarity of values indicates "local attraction" at the point $z_i \in Z$. Rather than ponder such philosophical issues, the view taken here is that the properties of "attraction" and "repulsion" are essentially statements about whether points in each pattern *tend* to occur together or not. In so far as such tendencies are only detectable by comparing patterns at many locations, these properties would seem to be global by their very nature.

References

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

[2] Diggle, P.J., (2003) *Statistical Analysis of Spatial Point Patterns, 2nd Edition*,Arnold: London.

[3] Diggle, P.J. and T.F. Cox, (1981) "On sparse sampling methods and tests of independence for multivariate spatial point patterns", *Bulletin of the International Statistical Institute*, 49: 213-229.

[4] Kendall, M.G., (1962) *Rank Correlation Methods*, 3^{rd} edition, Hafner Publishing Company: New York.

[5] Lotwick, H.W. and B.W Silverman, (1982) "Methods for analysing spatial processes of several types of points", *Journal of the Royal Statistical Society, B*, 44: 406-413.

Ripley, B.D.,(1976) "The second-order analysis of stationary point processes", *Journal of Applied Probability*, 13: 255-266.

[6] Ripley, B.D., (1977) "Modelling spatial patterns", *Journal of the Royal Statistical Society,B*, 39: 172-192.