Nearest neighbor methods

PHILIP M. DIXON

March 5, 2012

Nearest neighbor (NN) methods include at least six different groups of statistical methods. All have in common the idea that some aspect of the similarity between a point and its NN can be used to make useful inferences. In some cases, the similarity is the distance between the point and its NN; in others, the appropriate similarity is based on other identifying characteristics of the points. I will discuss in detail NN methods for studying spatial point processes and analyzing field experiments because both are commonly used in biology and environmetrics. I will very briefly discuss NN designs for field experiments. I will not discuss NN estimates of probability density functions (pdfs) [24], NN methods for discrimination or classification [61], or NN linkage (i.e. simple linkage) in hierarchical clustering [32], pp. 73–84. Although these last three methods have been applied to environmetric data, they are much more general.

1 Nearest Neighbor Methods for Spatial Point Processes

Spatial point process data describe the locations of 'interesting' events (*see* **Point processes, spatial**) and (possibly) some information about each event. Some examples include locations of tree trunks [57], locations of bird nests [11], locations of pottery shards, and locations of cancer cases [21]. I will focus on the most common case where the location is recorded in two dimensions (x, y). Similar techniques can be used for three-dimensional data (e.g. locations of galaxies in space) or one-dimensional data (e.g. nesting sites along a coastline or along a riverbank). Usually, the locations of all events in a defined area are observed (completely mapped data), but occasionally only a subset of locations is observed (sparsely sampled data). Univariate point process data include only the locations of the events; marked point process data include additional information about the event at each location [41], pp. 294-296. For example, the species may be recorded for each tree, some cultural identification may be recorded for each pottery shard, and nest success or nest failure may be recorded for each bird nest.

Location or marked location data can be used to answer many different sorts of questions. The scientific context for a question depends on the area of application, but the questions can be grouped into general categories. One very common category of question concerns the spatial pattern of the observations. Are the locations spatially clustered? Do they tend to be regularly distributed, or are they random (i.e a realization of a homogeneous Poisson process)? A second common set of questions concerns the relationships between different types of events in a marked point process. Do two different species of tree tend to occur together? Are locations of cancer cases more clustered than a random subset of a control group (*see* **Disease mapping**)? A third set of questions deals with the density (number

of events per unit area). What is the average density of trees in an area? What does a map of density look like? Methods to answer each of these types of question are discussed in the following sections.

Theoretical treatments of NN methods for spatial point patterns can be found in [19], [25] and [41]. Applications of NN methods can be found in many articles and books, including [41], [58] and [65].

2 Describing and Testing Spatial Patterns Using Completely Mapped Data

Describing and testing spatial patterns of locations has a long history. Historically, the primary concern was with the question of randomness [1, 2, 16, 52]. Are locations randomly distributed throughout the study area (i.e. are the locations a realization of a Poisson process with homogeneous intensity), or do the locations indicate some structure (i.e. clustering or repulsion between locations)? Because of the many connotations of randomness and the importance of a homogeneous Poisson process as a benchmark, it is commonly called *complete spatial randomness* (CSR).

In this section I will describe NN tests based on completely mapped data. Locations of all events are recorded in an arbitrary study region. Often, the study region is square or rectangular, but this is not a requirement. Tests for the less common case of sampled data are described in the next section.

2.1 Tests Based on Mean Nearest Neighbor Distance

The distances between NNs provide information about the pattern of points. Define W as the distance from a randomly chosen event to the nearest other event in a homogeneous Poisson process with intensity (expected number of points per unit area; *see* **Poisson intensity**) of ρ. The pdf and cumulative distribution function (cdf) of W are

$$g(w) = 2\rho\pi w \exp(-\rho\pi w^2) \tag{1}$$

$$G(w) = 1 - \exp(-\rho\pi w^2) \tag{2}$$

so the mean and the variance of W are

$$\mathrm{E}(W) = \frac{1}{2\rho^{1/2}} \tag{3}$$

and

$$\mathrm{var}(W) = \frac{4-\pi}{4\pi\rho} \tag{4}$$

Based on these moments evaluated at the observed density, $\hat{\rho}$, Clark and Evans [16] proposed using the standardized mean to test CSR. The Clark-Evans statistic [16] is

$$Z_{\mathrm{C}E} = \frac{\bar{w} - \mathrm{E}[W|\hat{\rho}]}{(N^{-1}\mathrm{var}[W|\hat{\rho}])^{1/2}}. \tag{5}$$

Their proposed test was to compare $Z_{\mathrm{C}E}$ to a standard normal distribution.

This test and the many users of it ignore two problems: nonindependence and edge effects. In a completely mapped area, many of the distances between NNs are correlated. That correlation is 1.0 when two points, A and B, are reflexive NNs, i.e. when B is the NN of A, and A is the NN of B [17]. Other authors have called these 'isolated NNs' [56] or 'mutual NNs' [62]. This problem is not restricted to a few points. When points exhibit CSR in two dimensions, approximately 62.15% of the points are reflexive NNs [17]. Their effect is to inflate the variance of the mean NN distance.

Edge effects arise because the distribution of W (2) assumes an unbounded area, but the observed NN distances are calculated from points in a defined study area. When a point is near the edge of the study area, it is possible that the true NN is a point just outside the study area, not a more distant point that happens to be in the study area. Edge effects lead to overestimation (positive bias) of the mean NN distance. Edge effects can be practically important; neglecting them can alter conclusions about the spatial pattern [10].

Edge effects may be minimized by including a buffer area that surrounds the primary study area [16]. NN distances are calculated only for points in the primary study area, but locations in the buffer area are available as potential NNs. With a sufficiently large buffer area, this approach can eliminate edge effects, but it is wasteful since an appropriately large buffer area may contain many locations. The effects of both issues can be elminated by using Monte-Carlo simulation to establish study-specific critical values (*see* **Simulation and Monte Carlo methods**) or by deriving approximations for the edge-corrected moments by curve-fitting to an extensive set of simulation results [31].

One difficulty with tests based on the mean NN distance is that the mean is just a single summary of the pattern. Two point patterns may have the same mean NN distance, but one exhibits CSR and the other does not. One such pattern would have a few patches of clustered points and an appropriate number of widely scattered individuals. The points in the clusters have small NN distances, but the widely scattered individuals have large NN distances. With the appropriate mix of clustered and scattered points, the mean NN distance could be exactly that given by (3).

2.2 Distribution of Nearest Neighbor Distances

An alternative is to consider the entire distribution, $G(w)$, of NN distance [25]. CSR can be tested by comparing the observed distribution function of NN distances, $\widehat{G}(w)$, to the theoretical cdf (2). A variety of test statistics have been suggested, including Kolmogorov–Smirnov type statistics: $\sup_w |\widehat{G}(w) - G(w)|$, Cramer–von Mises type statistics: $\int_w [\widehat{G}(w) - G(w)]^2$, or Anderson–Darling type statistics: $\int_w [\widehat{G}(w) - G(w)]^2 / [G(w)(1 - G(w))]$. The Kolmogorov–Smirnov statistic seems to be the most commonly used. The usual critical values for the one-sample Kolmogorov–Smirnov test are not appropriate here because of nonindependence of NN distances, especially for reflexive NNs, as discussed above. Instead, a Monte Carlo test (*see* **Simulation and Monte Carlo methods**) must be used [28].

Edge-corrected estimators of $G(w)$ have been devised; some of them are summarized in [19], pp. 613–614, 637–638 and [41], pp. 208-212. The edge-corrected cdf can also be estimated by a Kaplan–Meier estimator [3]. Edge corrections reduce the bias in the estimator, but they increase the sampling variance [36].

Although edge-corrected estimators are needed if the observed distribution function is to be compared with the theoretical distribution function under CSR (2) they may not be needed for a Monte Carlo test of CSR. A more

powerful test of CSR is to use a nonedge-corrected estimator and compare the biased estimate of $\widehat{G}(w)$ to the biased mean, $\bar{G}(w)$, under CSR [36]. The biased mean $\bar{G}(w)$ and pointwise prediction intervals are computed by simulation. Values of $\widehat{G}(w)$ above the simulated mean indicate clustering of points (an excess of short distances to NNs). Values of $\widehat{G}(w)$ below the simulated mean indicate regularity (few to no points with short distances to NNs). Tests of departure from CSR at a specific distance, w, are based on quantiles of $G(w)$ estimated by simulation.

The Monte Carlo approach is not limited to testing CSR. It can be used to evaluate the fit of any process that can be simulated. For example, a set of locations might be compared with a Poisson cluster process or a Strauss process [25, 41].

NN methods have been extended in a variety of ways. Tests can be based on other functions of the NN distances (e.g. squared [12] or smallest [63] NN distances), but such tests have not been widely used. Distances to second NN, or the third NN, or perhaps an even further neighbor have been suggested as a way to look at patterns on a larger scale. Finally, the distance between a randomly chosen point and the nearest event also provides information about the spatial pattern [25, 41].

2.3 Point–Event Distances and the $J(x)$ function

The point–event distribution, $F(x)$, considers the distance between a randomly chosen location (not the location of an event) and the nearest event. This can be estimated by choosing m locations in the study area and computing the distance from each location to its NN. As with $G(w)$, edge effects complicate estimation of the cdf. An edge-corrected estimator is

$$\hat{F}_{\mathrm{R}}(x) = \frac{\text{number of } (x_{\mathrm{i}} \leq x, d_{\mathrm{i}} > x)}{\text{number of } (d_{\mathrm{i}} > x)} \tag{6}$$

where x_{i} is the distance between a point and its neighboring event, and d_{i} is the distance between a point and its nearest boundary. When the events are a realization of CSR, X, the point–event distance, and W, the NN distance, have the same distribution, so $F(x) = 1 - \exp(-\rho\pi x^2)$. However, the effects of deviations from randomness on $F(x)$ are opposite to those on $G(w)$. Values of $\widehat{F}(x)$ above the expected value indicate regularity. Values below the expected value indicate clustering.

New summary statistics can derived by combining the event-event distance distribution, $G(x)$, and point-event distance distribution, $F(x)$. Van Lieshout and Baddeley [47] proposed the J function,

$$J(x) = \frac{1 - G(x)}{1 - F(x)}, \tag{7}$$

which is defined for all x for which $F(t) < 1$. $J(t) = 1$ for all t if the process is CSR. $J(t) > 1$ indicates inhibition at distance t and $J(t) > 1$ indicates clustering at distance t. $J(t)$ can be estimated without requiring edge correction because the biases arising from edge effects cancel out if the process is CSR [4]. Recently, estimators of $F(x)$, $G(w)$, and $J(x)$ have been developed for inhomogeneous point processes [46]. Maltez-Mauro et al use both the $J(x)$ and Ripley's $K(t)$ function to evaluate spatial patterns of recruitment in a Mediterranean Oak forest [49].

The NN distance distribution, point–event distance distribution, and Ripley's K function provide different insights into the spatial pattern. The NN distribution function, $\widehat{G}(w)$, is slightly more powerful at detecting departures from

CSR in the direction of regularity [25]. The point–event distribution function provides information about the empty space between points. It appears to be the more powerful method for detecting departures in the direction of clustering [25, 41]. Ripley's K function simultaneously examines the spatial pattern at many distance scales and is now the most popular approach for completely mapped data. However, it is possible to construct point processes with the same $G(w)$ but a different Ripley's $K(t)$ function [41]. Conversely, Baddeley and Silverman [5] illustrate two processes with the same Ripley's K function but very different NN distance distributions and point–event distance distributions.

2.4 Are Events Points or Circles?

The distributional theory in the previous sections assumes that events occupy no space. Treating events as points assumes that it is possible for two events (e.g. locations of tree trunks or bird nests) to be an infinitesimally small distance from each other. The point assumption is reasonable when the area of the events is small relative to the spacing between the events. The assumption is likely to be appropriate for bird nests (generally small) or tree trunks (generally low density), but not for ant nests (large size relative to the density of nests in an area). If events are incorrectly assumed to be points, the analysis of the spatial pattern indicates a tendency to regularity because two events do not occur within a small distance (the physical size of the event) of each other.

An approximation to the mean event–event distance, $\bar{W}$, for nonoverlapping circles under CSR is

$$\bar{W} \approx \frac{d + \exp(-\rho\pi d^2)[1 - \Phi(2\rho\pi d)^{1/2}]}{\sqrt{\rho}} \tag{8}$$

where d and ρ are, respectively, the diameter of the circles and the number of circles per unit area [64]. In (8) a low intensity, ρ, and a small and constant diameter, d, are assumed. The distribution, $G(w)$, of NN distances for nonoverlapping circles can be estimated by Monte Carlo simulation by using a sequential inhibition algorithm [25]. The distribution of event–event distances can be complicated when the circles are large or the density is high. For example, it may not be possible to fit the required number of large circles into the study area.

2.5 Algorithms and Computing

The simplest way to compute NN distances is by direct enumeration; that is, computing the distances between all pairs of points, then reporting the smallest distance for each point as the NN distance. For a large number of points, this becomes impractical, and more efficient algorithms have been developed to approach the problem. Possible approaches include subdividing the region into smaller subregions [51], computing the Dirichlet tessellation (also known as the Voronoi tessellation or Thiessen tessellation) and using that to identify NNs, or computing the quadtree, a sorted matrix of locations that simplifies the search for NNs, and using that to identify NNs [35]. Murtagh [51] reviews the properties of these algorithms.

Functions or procedures for NN spatial analysis are included in few statistical programs, but direct enumeration is very simple to program when needed. Packages of functions for spatial point pattern analysis usually include

functions for point–point and point–event analyses. Many of these are R packages such as Splancs, spatial, and spatstat.

2.6 Example: Trees in a Swamp Hardwood Forest

Figure 1 shows the locations of all 630 trees (stems > 11.5 cm diameter at breast height) and the locations of 91 cypress trees in a 1 ha plot of swamp hardwood forest in South Carolina, USA. There are 13 different tree species in this plot, but over 75% of the stems are one of three species: black gum (*Nyssa sylvatica*), water tupelo (*Nyssa aquatica*), or bald cypress (*Taxodium distichum*). Visually, trees seem to be scattered randomly throughout the plot, but cypress trees seem to be clustered in three bands. NN statistics provide a way to test the hypothesis that stems are randomly distributed throughout the plot. I will illustrate tests based on mean NN distance, $G(w)$, and $F(x)$ using the locations of all 630 trees and the locations of the 91 cypress trees.

For all 630 tree locations, the mean NN distance is 1.99 m. If 630 points were randomly distributed in a 200 m × 50 m rectangle, the expected NN distance is 2.034 m, with a standard error (SE) of 0.044 m, using edge-corrected moments [31]. There is no evidence of departure from CSR ($z = -0.973$ with a two-sided P value of 0.33). The effect of the edge corrections is minimal here because the plot is large and the NN distance is small. The uncorrected expected NN distance is 1.992 m, with an SE of 0.042 m, using (3) and (4).

Conclusions using the distribution of NN distances, $\widehat{G}(w)$, are similar. $\widehat{G}(w)$ was estimated without edge corrections, so $\widehat{G}(w)$ must be compared with simulated values, not the theoretical expectation (3). The observed cdf, the theoretical expected value (3), and the average simulated cdf are very similar (Figure 2a), although the observed $\widehat{G}(w)$ is slightly larger than the expected value at short distances. The differences can be seen more clearly if $1 - \exp(-\rho\pi w^2)$ is subtracted from all curves (Figure 2b). Although $\widehat{G}(w)$ is larger than both the theoretical expected value and the average simulated value, it lies within the pointwise 95% confidence limits. None of the three summary statistics (Kolmogorov–Smirnov, Cramer–von Mises, or Anderson–Darling) is significant at $\alpha = 0.05$. For example, the observed Kolmogorov–Smirnov test statistic of 0.044 is less than the simulated 90th percentile, 0.052. The estimated P values for the Kolmogorov–Smirnov, Cramer–von Mises, and Anderson–Darling test statistics are 0.19, 0.31, and 0.40, respectively.

In contrast, the distribution of point–event distances suggests there is some clustering of tree locations. The observed $\widehat{F}(x)$ is below the theoretical and average simulated curves (Figure 3a, b) and outside the pointwise 95% confidence bounds at large distances (Figure 3b). Because $\widehat{F}(x)$ falls below the expected values, distances from randomly chosen points are stochastically greater than expected if events exhibited CSR. The greater than expected abundance of large empty spaces provides evidence of clustering of the events. The P values for the Kolmogorov–Smirnov, Cramer–von Mises, and Anderson–Darling summary statistics range from 0.004 to 0.009. The conclusion of some evidence for clustering of all tree locations matches the conclusion using Ripley's K function.

For the 91 cypress trees, the mean NN distance is 5.08 m, which is slightly smaller than the edge-corrected expected distance of 5.55 m, with an SE of 0.33 m. Using the NN distance, there is no evidence of a nonrandom distribution; the z-statistic is -1.41, with a two-sided P value of 0.16. The effect of the edge corrections is larger when the density of points is smaller. The uncorrected expected NN distance for the 91 cypress trees is 5.24 m, with an SE of 0.29 m.

Figure 1: Marked plot of tree locations in a 50 m × 200 m plot of hardwood swamp in South Carolina, USA. Circles are locations of cypress trees, squares are locations of black gum trees, and dots are locations of any other species

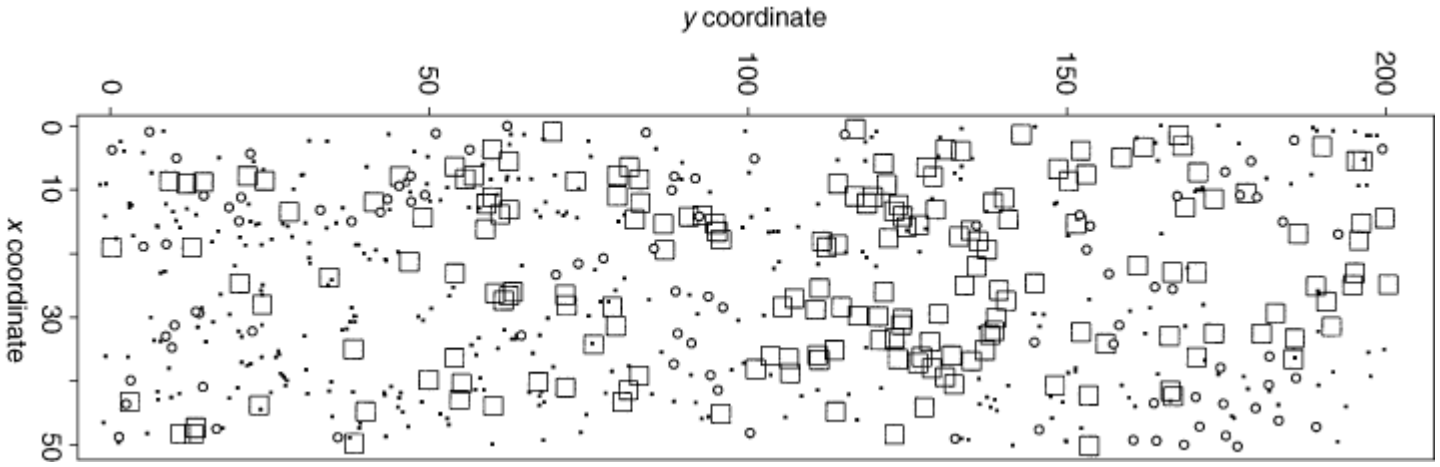


Figure 2: Cdf plots of NN distances for all trees and for cypress trees only, in a 50 m × 200 m plot (see Figure 1). (a) All trees: comparison of observed cdf (solid line), theoretical cdf (dotted line) and average simulated cdf under CSR (dashed line). (b) All trees: comparison of observed cdf (solid line) and average simulated cdf (dashed line). The dotted lines represent the pointwise 0.025 and 0.975 quantiles of the simulated cdf under CSR. For clarity, the theoretical cdf is subtracted from all plotted cdfs. (c) Cypress trees only: comparison of observed cdf (solid line), theoretical cdf (dotted line) and average simulated cdf under CSR (dashed line). (d) Cypress trees only: comparison of observed cdf (solid line) and average simulated cdf (dashed line). The dotted lines represent the pointwise 0.025 and 0.975 quantiles of the simulated cdf under CSR. For clarity, the theoretical cdf is subtracted from all plotted cdfs

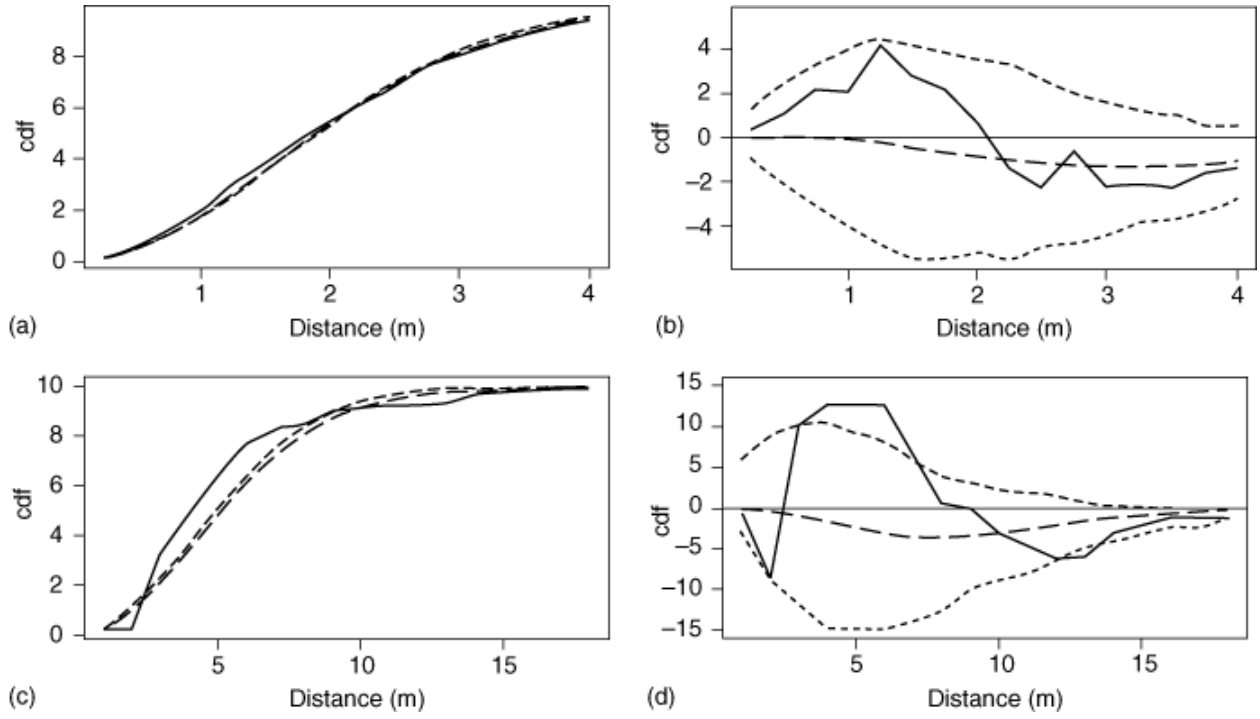
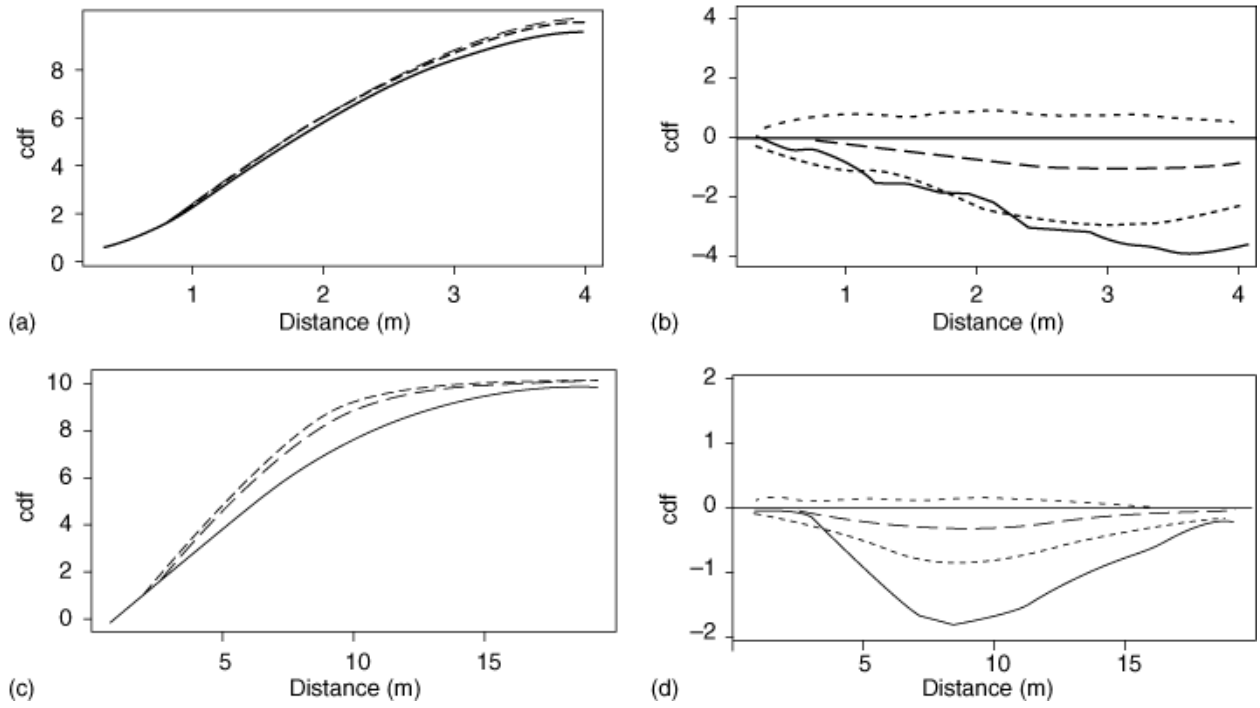


Figure 3: Cdf plots of point–event distances for all trees and for cypress trees only, in a 50 m × 200 m plot (see Figure 1). (a) All trees: comparison of observed cdf (solid line), theoretical cdf (dotted line) and average simulated cdf under CSR (dashed line). (b) All trees: comparison of observed cdf (solid line) and average simulated cdf (dashed line). The dotted lines represent the pointwise 0.025 and 0.975 quantiles of the simulated cdf under CSR. For clarity, the theoretical cdf is subtracted from all plotted cdfs. (c) Cypress trees only: comparison of the observed cdf (solid line), theoretical cdf (dotted line) and average simulated cdf under CSR (dashed line). (d) Cypress trees only: comparison of observed cdf (solid line) and average simulated cdf (dashed line). The dotted lines represent the pointwise 0.025 and 0.975 quantiles of the simulated cdf under CSR. For clarity, the theoretical cdf is subtracted from all plotted cdfs



The distributions of $\widehat{G}(w)$ and $\widehat{F}(x)$ for the 91 cypress trees provide evidence of clustering. There is an unusually large number of NN distances between 3 m and 7 m (Figures 2c and 2d) and $\widehat{G}(w)$ is at or above the pointwise 0.975 quantiles of simulated values (Figure 2d). This excess is consistent with clustering of cypress trees. There are also significant fewer (at least pointwise) NN distances at 13 m. All three summary statistics are significant (Kolmogorov–Smirnov P value = 0.034, Cramer–von Mises P value = 0.007, Anderson–Darling P value = 0.011). The point–event distances are stochastically greater than expected under CSR (Figures 3c and 3d). The observed distribution, $\widehat{F}(x)$, lies outside the pointwise 95% confidence bands for many distances. All three summary statistics are highly significant ($P = 0.001$). The conclusion that cypress trees are strongly clustered matches that using Ripley's K function.

3 Directed Tests

The tests in the previous section are general tests of CSR against an unspecified alternative. Other tests may be more powerful when the alternative is more specific (e.g. events are associated with specific sites, or the density of events increases from east to west). Association between point events and a nonpoint stochastic process can be tested using the NN distance from each event to the nearest part of the second process [7]. However, most directed tests [44, 60, 66] use features other than NN distance.

4 Describing and Testing Spatial Patterns with Use of a Sample of Nearest Neighbor Distances

An alternative to mapping the locations of all events in a study area is to measure NN distances for a random sample of individuals. When NN distances are calculated from a simple random sample of events, the distributional theory for both the mean NN distance and the distribution function is much simpler. Many different tests of CSR have been developed for use with a random sample of point–event, NN, or point–event–event distances. These are summarized and evaluated in [19], pp. 602–614 and [25], pp. 33-40.

The most straightforward way to select a random sample of NN distances is to enumerate all individuals in the statistical population, select a simple random sample of events, and measure the distances from the selected events to their NNs. Enumeration can be avoided by clever use of subregions (described by Byth and Ripley [14]), or by randomly selecting points (not events). The distance from the randomly selected point to the nearest event is a random sample from the distribution of point–event distances, $F(x)$, but the distance from that event to its closest event is not a random sample from $G(w)$ because the point–event and event–event distances are correlated. The distributions of all quantities in the point–event–event sample when events exhibit CSR have been derived [18].

An alternative that is easy to implement in the field is the T-square sample [8], illustrated in [19] and [25], p. 35, a modified point–event–event sample. A point, A, is randomly chosen, and the NN, B, is found. Define X as the distance between A and B. Then, the study area is divided into two half planes by a line through B and perpendicular to AB (hence the name, T-square). Attention is restricted to the half plane that does not contain point A. The distance to the NN, Z, of B in that half plane is measured. When points exhibit CSR, X and Z are independent, and the distribution of $Z/\sqrt{2}$ is the same as the distribution of NN distances, $G(w)$ [8].

4.1 Estimating Density

A random sample of point to NN distances can be used to estimate ρ, the average density of events in the study area. When events exhibit CSR, the maximum likelihood estimate is

$$\hat{\rho} = n \left(\pi \sum_{i=1}^{n} X_i^2 \right)^{-1} \tag{9}$$

where X_i is the distance from a randomly chosen point i to its NN [50]. An unbiased estimate is [59]:

$$\hat{\rho}_{\mathrm{P}} = (n-1) \left(\pi \sum_{i=1}^{n} X_i^2 \right)^{-1} \tag{10}$$

Both estimators are very dependent on the CSR assumption and can be biased if locations are clustered or regularly distributed.

Many other estimators have been proposed. Diggle [25], pp. 39–40 summarizes many of these. Byth [13] evaluated the robustness of many estimators to deviations from CSR. She recommended an estimator, $\hat{\rho}_{\mathrm{T}}$, based on two quantities from T-square sampling: X, the distance from point to nearest event, and Z, the distance to the NN in a half plane:

$$\hat{\rho}_{\mathrm{T}} = n^2 \left[2\sqrt{2} \left(\sum_{i=1}^{n} X_i \right) \left(\sum_{i=1}^{n} Z_i \right) \right]^{-1} \tag{11}$$

5 Nearest Neighbor Methods to Examine Spatial Patterns of More Than One Type of Point

Additional information about an event is often available. For example, tree locations might be marked with the species of tree (a mark with discrete levels) or the size of the tree (a mark with continuous levels). The methods in the previous sections can be used to analyze patterns in all events (ignoring the marks) or subgroups of events (e.g. just species A or just trees larger than 50 cm). However, other interesting questions could be asked about the relationship between the two (or more) sets of locations.

Multivariate spatial point patterns are those where events can be classified into different types, that is, the marks are discrete [19], p. 707. Usually, the number of different types is small; bivariate patterns, with two types of marks, are the most common. Some questions that could be asked about point processes with discrete marks are as follows:

1. Are the processes that generate locations with different marks independent?
2. Are marks randomly assigned to locations? Conditional on the observed locations of superposition of the two marked processes, are the marks independent?

3. Are marks segregated? Are locations with one type of mark surrounded by locations with the same mark?

These questions about the relationships between processes make no assumptions about the marginal pattern of each process. In particular, either process (or the superposition of processes) may be independent, clustered, or regularly distributed. The two general methods to answer these questions are the comparison of distribution functions [39, 48] and the NN contingency table [16, 29, 57]. Other approaches include Ripley's K functions, van Lieshout's J function [45] or parametric point process models.

Define the following multitype extensions of the point–event and NN distances. X_i is the distance from a randomly chosen point to the nearest event with mark i, with cdf $F_i(x)$. W_{ij} is the distance from an event with mark i to the nearest event with mark j, with cdf $G_{ij}(w)$. If the process with mark i is independent of the process with mark j, then

$$F_i(x) = G_{ji}(x) \tag{12}$$
$$F_j(x) = G_{ij}(x) \tag{13}$$

and X_i and X_j are independent [27, 39]. Note that property (12) does not imply property (13), so two tests are needed [39]. One possible test is based on $\sup_x |F_i(x) - G_{ji}(x)|$. Critical values are estimated by Monte-Carlo simulation because of the nonindependence of point–event and event–event distances.

Two different simulation methods could be used in the Monte Carlo test. The choice depends on the null hypothesis. If the null hypothesis is independence between marks (question 1, above), then toroidal shifts or some parametric model should be used to generate the randomization distribution. If the null hypothesis is random assignment of marks conditional on the set of events, then random labeling of events should be used to generate the randomization distribution. In general, these two hypotheses are not equivalent and the sampling distributions are not the same [40].

5.1 Nearest Neighbor Contingency Tables

The NN contingency table focuses on the ecologically important question of segregation [29, 57]. This table describes marks of events and their NNs, not the distance between them (Table 1). In sparsely sampled data, the counts ($N_{\mathrm{A}A}, N_{\mathrm{A}B}, N_{\mathrm{B}A}$, and $N_{\mathrm{B}B}$) are independent Poisson random variables or conditionally independent given the row marginal totals (N_{A} and N_{B}) under the null hypothesis of random labeling [57]. The hypothesis can be tested with a traditional 1 degree of freedom (df) χ^2 test of independence [57] (*see* **Categorical data**).

In completely mapped data, the sampling distribution of the counts is different [17]. If events are randomly labeled, the expected values of the counts depend only on the number of each type of event (N_{A} and N_{B}) and the total number of events, N (Table 2) [29]. The variances and covariances depend on the number of events of each type, the number of reflexive NNs, and the number of shared NNs; for a derivation of the formula, see [29]. The first two moments of the cell counts can be used to test for segregation of type A events [$N_{\mathrm{A}A} > \mathrm{E}(N_{\mathrm{A}A})$], test for segregation of type B events [$N_{\mathrm{B}B} > \mathrm{E}(N_{\mathrm{B}B})$], or construct an omnibus 2 df χ^2 test of random labeling. If the numbers of points are large, then the distributions of test statistics can be adequately approximated by asymptotic

Table 1: Cell counts in an NN contingency table

	Mark of neighbor		
Mark of point	A	B	Total
A	$N_{\mathrm{A}A}$	$N_{\mathrm{A}B}$	N_{A}
B	$N_{\mathrm{B}A}$	$N_{\mathrm{B}B}$	N_{B}
Total	M_{A}	M_{B}	N

N is the number of points in the spatial pattern. N_{A} and N_{B} are the number of type A and type B points. $N_{\mathrm{A}A}$ is the number of type A points with type A NNs. $N_{\mathrm{A}B}$ is the number of type A points with type B NNs. $N_{\mathrm{B}A}$ and $N_{\mathrm{B}B}$ are the number of type B points with type A or type B NNs, respectively. M_{A} and M_{B} are the column sums, i.e. the total number of points with type A neighbors or type B neighbors, respectively

Table 2: Expected cell counts for NN contingency table for completely mapped data

	To:	
From:	A	B
A	$\frac{N_{\mathrm{A}}(N_{\mathrm{A}}-1)}{N-1}$	$\frac{N_{\mathrm{A}}N_{\mathrm{B}}}{N-1}$
B	$\frac{N_{\mathrm{B}}N_{\mathrm{A}}}{N-1}$	$\frac{N_{\mathrm{B}}(N_{\mathrm{B}}-1)}{N-1}$

See Table 1 for definitions of M, N_{A} and N_{B}

normal and χ^2 distributions [15]. If the number of points is small, then the distributions should be determined by Monte Carlo simulation.

Patterns with k marks ($k > 2$) can be analyzed by considering all pairs of marks two at a time (using distance methods or 2×2 NN contingency tables), or by considering the $k \times k$ contingency table. The expected counts and their variances under random labeling follow the same form as those for a 2×2 contingency table, but there are more possible forms for the covariance between two counts. Details are given in [30].

Other approaches that have been suggested for the analysis of multitype point processes include the comparison of bivariate Ripley's K functions [26, 48], empty space methods [48] (comparisons of point–event distance distributions), and mark correlation functions [41].

5.2 Example, Part 2

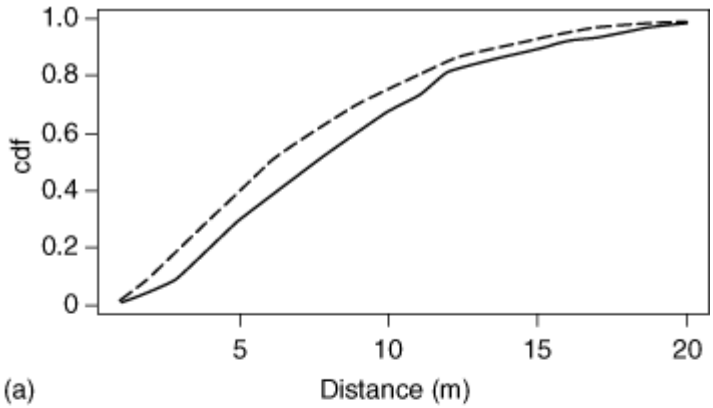
Cypress and black gum are two of the three abundant species in the 1-ha plot of swamp forest considered in Example 1. An interesting ecological question is whether these two species are spatially segregated; that is, do cypress trees tend to be found near other cypress trees and do black gum trees tend to be found near other black gum trees? The marked plot of locations (Figure 1) suggests that cypress trees and black gum occur in different clusters. Confirming this requires an analysis of the bivariate spatial pattern. The three tests that will be illustrated are the comparisons of cdfs (12), (13) [27], the independence of distances [27], and the NN contingency table [29]. The ecological background suggests that random labeling is the more appropriate null hypothesis.

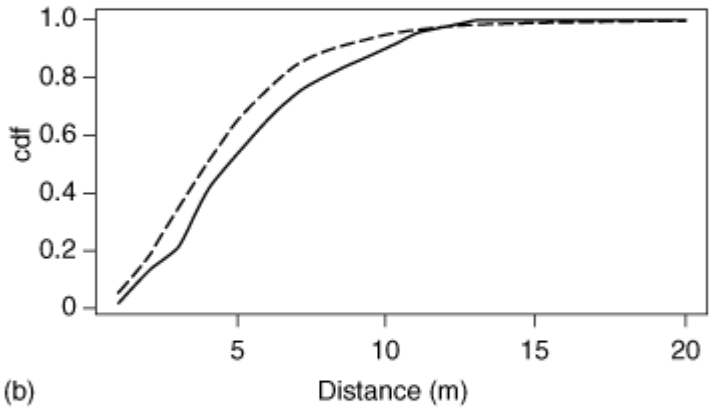
The cdf of distances from randomly chosen points to the nearest black gum tree, $F_{\mathrm{G}}(x)$, and the cdf of point–event distances to the nearest cypress tree, $F_{\mathrm{C}}(x)$, were estimated without edge corrections by using a randomly located grid of points [14]. The cdfs of distances from black gum trees to the nearest cypress tree, $G_{\mathrm{G}C}(x)$, and from cypress trees to the nearest black gum tree, $G_{\mathrm{C}G}(x)$, were also estimated without edge corrections. Both species show the same pattern. Cypress trees are stochastically further from black gum trees than from randomly chosen points (Figure 4a). Also, black gum trees are stochastically farther from cypress trees than from randomly chosen points (Figure 4b).

The observed differences can be compared with those found under random labeling by using the Kolmogorov–Smirnov two-sample statistics, $\sup_x |F_{\mathrm{G}}(x) - G_{\mathrm{C}G}(x)|$ and $\sup_x |F_{\mathrm{C}}(x) - G_{\mathrm{G}C}(x)|$, as the test statistics. The observed maximum differences are not unusually large (P value = 0.109 for black gum and 0.083 for cypress). The distance from a randomly chosen point to the nearest black gum tree, X_{G}, is negatively correlated with the distance from the same point to the nearest cypress tree, X_{C} (Figure 4c; Kendall's $\hat{\tau} = -0.12$, with a one-sided P value, by randomization, of 0.001). This result is consistent with the spatial segregation of the two species. The different results from the three tests are consistent with Diggle and Cox's [27] observation that the correlation test is more powerful than the Kolmogorov–Smirnov test for the sparsely sampled spatial patterns they studied.

The NN contingency table indicates that both species have an excess of NNs of the same species (Table 3). The variances of the cell counts are 38.88 for black gum–black gum and 25.55 for cypress–cypress. The P values can be computed by Monte Carlo randomization or by a normal approximation [29]. In either case, the one-sided P values are small (0.001 or less) for both species.

Figure 4: Cdf of point–event distances for all trees and for cypress trees only. (a) Comparison of cdfs of point to cypress tree distances [$F_{\mathrm{C}}(x)$, dotted line] and black gum tree to cypress tree distances [$G_{\mathrm{G}C}(x)$, solid line]. (b) Comparison of cdfs of point to black gum tree distances [$F_{\mathrm{G}}(x)$, dotted line] and cypress tree to black gum tree distances [$G_{\mathrm{C}G}(x)$, solid line]. (c) Relationship between distances from a randomly chosen point to the nearest cypress tree and nearest black gum tree





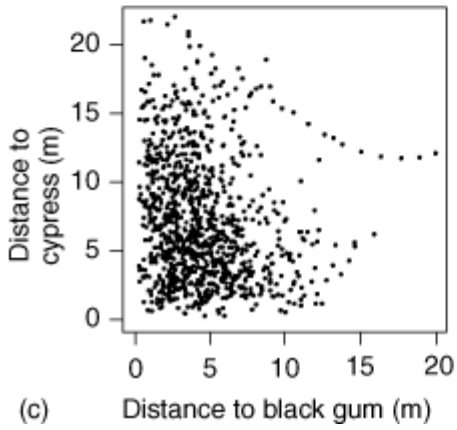


Table 3: Observed counts (N^{O}), expected counts (N^{E}), and z scores for the cypress tree and black gum tree NN contingency table

	Species of neighbor						
	Black gum tree			Cypress tree			
Species of point	N^{O}	N^{E}	z	N^{O}	N^{E}	z	Total
Black gum	149	121.1	4.47	33	60.9	−4.47	182
Cypress	43	60.9	−3.54	48	30.1	3.54	91
Total	192			81			273

6 Nearest Neighbor Methods for Field Experiments

A different set of NN methods can be used to analyze spatially structured field experiments. One of the most common applications is in agronomic variety trials, which commonly include many treatments randomly assigned to small plots arranged in a rectangular lattice. Traditional methods of controlling for between-plot heterogeneity, such as using a randomized complete block (RCB) design, may not be very effective when a large number of treatments forces the blocks to be large. NN methods use information from adjacent plots to adjust for within-block heterogeneity and so provide more precise estimates of treatment means and differences. If there is within-plot heterogeneity on a spatial scale that is larger than a single plot and smaller than the entire block, then yields from adjacent plots will be positively correlated. Information from neighboring plots can be used to reduce or remove the unwanted effect of the spatial heterogeneity and hence improve the estimate of the treatment effect. Data from neighboring plots can also be used to reduce the influence of competition between adjacent plots. Each of these approaches will be briefly discussed below.

Papadakis [53, 54] proposed an analysis of covariance to reduce the effects of small-scale spatial heterogeneity in yields. The value of the covariate for each plot is obtained by averaging residuals from the neighboring plots. The choice of neighboring plots depends on the crop, the plot size and shape, and the spacing between plots. In many row crops, the neighboring plots are defined as the two adjacent plots in the row, except that plots at an end of a row have only one neighbor. In other situations, it may be appropriate to consider four neighbors, which include the two between-row neighbors. If the spatial heterogeneity is such that the effects of within-row and between-row neighbors are quite different, then one could compute separate covariates for within-row and between-row neighbors. Once the covariates are computed, treatment effects are re-estimated using an analysis of covariance. For adjustment in one dimension (e.g. along crop rows), the model would be

$$Y_i = \mu + X_i\tau_i + \beta R_i + \varepsilon_i \tag{14}$$

where Y_i is the observed yield on the ith plot, R_i is the mean residual on neighbors of the ith plot, β is the spatial dependence parameter, μ and τ_i are the parameters in the model for the treatment effects, X_i is the row of the design matrix for the ith plot, and ε_i is the residual variability in yields, which are assumed to be uncorrelated. When β is 0, observations on adjacent plots are independent; larger positive values of $\beta\,(\beta < 1)$ correspond to

increasing spatial correlation between neighboring plots. The observed value of β depends on plot size, plot shape, plot spacing, and the scale of the spatial heterogeneity. Values are often close to 1 when plots are small. In the absence of treatments, and ignoring edge effects, the Papadakis model implies that correlations between plot yields have a first-order autoregressive structure, with $\text{corr}(Y_i, Y_j) = \lambda^{|i-j|}$, where $\beta = 2\lambda/(1+\lambda^2)$. Values of β close to 1 imply that λ is also close to 1.

The autoregressive correlation structure implied by the ad hoc Papadakis model is one example of the random field approach to a spatially designed experiment [6, 19, 71]. Many other models, including the iterated Papadakis method, the Wilkinson NN [69] model, the Besag and Kempton [9] first-order difference models, the Williams model [67] and the Gleeson–Cullis autoregressive integrated moving average (ARIMA) models [20, 38] (*see* **Time series**) correspond to different specifications of a spatial correlation matrix. Computations are handled either by a general purpose residual maximum likelihood (REML) algorithm for linear mixed effects models [e.g. PROC MIXED or PROC GLIMMIX in SAS, lme() in the R nlme package, lmer( ) in the R lme4 package], or by specialized software for a particular model (e.g. TwoD for two-dimensional Gleeson–Cullis models [37]).

The properties of these methods have been extensively discussed over the past 20 years. Dagnelie [22, 23] provides a review and historical summary of the Papadakis model. Wu and Dutilleul [70] use uniformity trial data to compare autoregressive models, difference models, and traditional RCB analyses. The efficiency of a spatial analysis, relative to an RCB design, is usually greater than 1.2 and can be as high as 2 [23]. Wilkinson [69] claimed the Papadakis estimator of treatment effects was biased, but Pearce [55] found generally very small biases in the scenarios he considered. The Papadakis method appears to work best when there are at least three replicates per treatment, many treatments (greater than 10), and strong but patchy spatial heterogeneity [23]. When there is an underlying trend, first-order difference models appear to work well.

Medium-scale spatial heterogeneity usually causes a positive correlation between adjacent plots. When there is competition between plots, neighbors can have a negative effect on the response in adjacent plots [42]. The Papadakis model (14) can be extended to estimate treatment-specific competitive effects. The choice of covariate should be influenced by biological mechanisms. If competition for sunlight is important, a reasonable covariate could be the difference between the mean height of plants in the plot and the mean height on neighboring plots. If disease spread is important, a reasonable covariate could be the mean disease severity on neighboring plots [43]. The coefficient for the covariate [β in (14)] estimates the strength of the competitive relationship.

Experimental design for a study that will use some form of neighbor-adjusted analysis usually focuses on neighbor balance; that is, ensuring that all pairs of treatments occur in adjacent plots equally frequently. Adjacent plots can be defined as only those within the same row (one-dimensional neighbor-balanced designs [68]) or as including both those in the same row and those in the same column (two-dimensional neighbor-balanced designs [34]). The choice will depend on the size, shape, and spacing of the plots and on the biological and physical mechanisms influencing the correlation between plots. Methods for construction of one-dimensional or two-dimensional designs can be found in [33], [34], and [68].

References

[1] Ashby, E. (1936). Statistical ecology, Botanical Review **2**, 221–235.

[2] Ashby, E. (1948). Statistical ecology. II – a reassessment, Botanical Review **14**, 222–234.

[3] Baddeley, A. & Gill, R.D. (1997). Kaplan–Meier estimators of distance distributions for spatial point processes, The Annals of Statistics **25**, 263–292.

[4] Baddeley, A., Kerscher, M., Schladitz, K., & Scott, B.T. (2000). Estimating the J function without edge correction. *Statistica Neerlandica* **54**, 315-328.

[5] Baddeley, A.J. & Silverman, B.W. (1984). A cautionary example on the use of second-order methods for analyzing point patterns, Biometrics **40**, 1089–1093.

[6] Bartlett, M.S. (1978). Nearest neighbour models in the analysis of field experiments (with discussion), Journal of the Royal Statistical Society, Series B **40**, 147–174.

[7] Berman, M. (1986). Testing for spatial association between a point process and another stochastic process, Applied Statistics **35**, 54–62.

[8] Besag, J.E. & Gleaves, J.T. (1973). On the detection of spatial pattern in plant communities, Bulletin of the International Statistical Institute **45**, 153–158.

[9] Besag, J. & Kempton, R. (1986). Statistical analysis of field experiments using neighbouring plots, Biometrics **42**, 231–251.

[10] Brewer, R. & McCann, M.T. (1985). Spacing in acorn woodpeckers, Ecology **66**, 307–308.

[11] Brown (1975). A test of randomness in nest spacing, Wildlife **26**, 102–103.

[12] Brown, D. & Rothery, P. (1978). Randomness and local regularity of points in a plane, Biometrika **65**, 115–122.

[13] Byth, K. (1982). On robust distance-based intensity estimators, Biometrics **38**, 127–135.

[14] Byth, K. & Ripley, B.D. (1980). On sampling spatial patterns by distance methods, Biometrics **36**, 279–284.

[15] Ceyhan, E. (2010). On the use of nearest neighbor contingency tables for testing spatial segregation, Environmental and Ecological Statistics **17**, 247–282.

[16] Clark, P.J. & Evans, F.C. (1954). Distance to nearest neighbor as a measure of spatial relationships in populations, Ecology **35**, 445–453.

[17] Cox, T.F. (1981). Reflexive nearest neighbours, Biometrics **37**, 367–369.

[18] Cox, T.F. & Lewis, T. (1976). A conditioned distance ratio method for analyzing spatial patterns, Biometrika **63**, 483–491.

[19] Cressie, N. (1991). Statistics for Spatial Data, Wiley, New York.

[20] Cullis, B.R. & Gleeson, A.C. (1991). Spatial analysis of field experiments – extension to two dimensions, Biometrics **47**, 1449–1460.

[21] Cuzick, J. & Edwards, R. (1990). Spatial clustering for inhomogeneous populations (with discussion), Journal of the Royal Statistical Society, Series B **52**, 73–104.

[22] Dagnelie, P. (1987). La méthode de Papadakis en expérimentation agronomique: considérations historiques et bibliographiques, Biométrie et Praximétrie **27**, 49–64.

[23] Dagnelie, P. (1989). The method of Papadakis in agricultural estimation: an overview, Biuletyn Oceny Odmian (Cultivar Testing Bulletin) **21**, 111–122.

[24] Devroye, L. & Györfi, L. (1985). Nonparametric Density Estimation. The L_1 View, Wiley, New York.

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

[26] Diggle, P.J. & Chetwynd, A.G. (1991). Second-order analysis of spatial clustering for inhomogeneous populations, Biometrics **47**, 1155–1163.

[27] Diggle, P.J. & Cox, T.F. (1983). Some distance-based tests of independence for sparsely-sampled multivariate spatial point patterns, International Statistical Review **51**, 11–23.

[28] Diggle, P.J., Besag, J. & Gleaves, J.T. (1976). Statistical analysis of spatial point patterns by means of distance methods, Biometrics **32**, 659–667.

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

[30] Dixon, P.M. (2002). Nearest-neighbor contingency table analysis of spatial segregation for several species, Ecoscience **9**, 12–151.

[31] Donnelly, K.P. (1978). Simulations to determine the variance and edge effect of total nearest-neighbour distance, in Simulation Studies in Archaeology, I. Hodder, ed., Cambridge University Press, Cambridge, pp. 91–95.

[32] Everitt, B.S., Landau, S., Leese, M. & Stahl, D. (2011). Cluster Analysis, 5th Edition, Wiley, Chichester.

[33] Federer, W.T. & Basford, K.E. (1991). Competing effects designs and models for two-dimensional field arrangements, Biometrics **47**, 1461–1472.

[34] Freeman, G.H. (1979). Some two-dimensional designs balanced for nearest neighbours, Journal of the Royal Statistical Society, Series B **41**, 88–95.

[35] Friedman, J., Bentley, J.L. & Finkel, R.A. (1977). An algorithm for finding best matches in logarithmic expected time, ACM Transactions on Mathematical Software **3**, 209–226.

[36] Gignoux, J., Duby, C. & Barot, S. (1999). Comparing the performance of Diggle's tests of spatial randomness for small samples with and without edge-effect correction: application to ecological data, Biometrics **55**, 156–164.

[37] Gilmour, A.R. (1992). TwoD: a program to fit a mixed linear model with two dimensional spatial adjustment for local trend, NSW Agriculture Research Center, Tamworth, Australia.

[38] Gleeson, A.C. & Cullis, B.R. (1987). Residual maximum likelihood (REML) estimation of a neighbour model for field experiments, Biometrics **43**, 277–288.

[39] Goodall, D.W. (1965). Plot-less tests of interspecific association, Journal of Ecology **53**, 197–210.

[40] Goreaud, F. & Pélissier, R. (2003). Avoiding misinterpretation of biotic interactions with the intertype K12-function: population independence bs. random labelling hypothesis. *Journal of Vegetation Science* **14**, 681-692.

[41] Illian, J., Penttinen, A., Stoyan, H., & Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Wiley, Chichester.

[42] Kempton, R. (1982). Adjustment for competition between varieties in plant breeding trials, Journal of Agricultural Science **98**, 599–611.

[43] Kempton, R. (1991). Interference in agricultural experiments, *Proceedings of the Second Biometric Society East/Central/Southern African Network Meeting*, Harare, Zimbabwe.

[44] Lawson, A. (1988). On tests for spatial trend in a non-homogeneous Poisson process, Journal of Applied Statistics **15**, 225–234.

[45] van Lieshout, M.N.M. (2006). A J-function for marked point patterns. *Annals of the Institute of Statistical Mechanics* **58**, 235-259.

[46] van Lieshout, M.N.M. (2011). A J-function for inhomogeneous point processes. *Statistica Neerlandica* **65**, 183–201.

[47] van Lieshout, M.N.M. & Baddeley, A. (1996). A nonparametric measure of spatial interaction in point patterns. *Statistica Neerlandica* **50**, 344-361..

[48] Lotwick, H.W. & Silverman, B.W. (1982). Methods for analyzing spatial processes of several types of points, Journal of the Royal Statistical Society, Series B **44**, 406–413.

[49] Maltez-Mauro, S., García L.V., Marañón, T., & Freitas, H. (2007). Recruitment patterns in a Meditteranean Oak forest: a case study showing the important of the spatial component. *Forest Science* **53**, 645–652.

[50] Moore, P.G. (1954). Spacing in plant populations, Ecology **35**, 222–227.

[51] Murtagh, F. (1984). A review of fast techniques for nearest neighbor searching, Compstat **1984**, 143–147.

[52] Neyman, J. & Scott, E.L. (1952). A theory of the spatial distribution of galaxies, Astrophysical Journal **116**, 144–163.

[53] Papadakis, J.S. (1973). Méthode statistique pour des expériences sur champ, Institut d'Amélioration des Plantes a Thessaloniki (Thessalonika, Greece), *Bulletin Scientifique*, No. 23.

[54] Papadakis, J.S. (1984). Advances in the analysis of field experiments, Proceedings of the Academy of Athens **59**, 326–342.

[55] Pearce, S.C. (1998). Field experimentation on rough land: the method of Papadakis reconsidered, Journal of Agricultural Science, Cambridge **131**, 1–11.

[56] Pickard, D.K. (1982). Isolated nearest neighbors, Journal of Applied Probability **19**, 444–449.

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

[58] Pielou, E.C. (1977). Mathematical Ecology, Wiley, New York.

[59] Pollard, J.H. (1971). On distance estimators of density in randomly distributed forestes, Biometrics **27**, 991–1002.

[60] Rathbun, S.L. (1996). Estimation of Poisson intensity using partially observed concomitant variables, Biometrics **52**, 226–242.

[61] Ripley, B.D. (1996). Pattern Recognition and Neural Networks, Cambridge University Press, Cambridge.

[62] Schilling, M.F. (1986). Mutual and shared neighbor probabilities: finite- and infinite-dimensional results, Advances in Applied Probability **18**, 388–405.

[63] Silverman, B. & Brown, T. (1978). Short distance, flat triangles, and Poisson limits, Journal of Applied Probability **15**, 815–825.

[64] Simberloff, D. (1979). Nearest neighbor assessments of spatial configurations of circles rather than points, Ecology **60**, 679–685.

[65] Stoyan, D. & Penttinen, A. (2000). Recent applications of point process methods in forestry statistics, Statistical Science **15**, 61–78.

[66] Waller, L.A., Turnbull, B.W., Clark, L.C. & Nasca, P. (1994). Spatial pattern analyses to detect rare disease clusters, in Case Studies in Biometry, N. Lange et al. eds, Wiley, New York, pp. 3–23.

[67] Williams, E.R. (1986). A neighbour model for field experiments, Biometrika **73**, 279–287.

[68] Williams, R.M. (1952). Experimental designs for serially correlated observations, Biometrika **39**, 151–167.

[69] Wilkinson, G.N., Eckert, S.R., Hancock, T.W. & Mayo, O. (1983). Nearest neighbour (NN) analysis of field experiments (with discussion), Journal of the Royal Statistical Society, Series B **45**, 151–211.

[70] Wu, T. & Dutilleul, P. (1999). Validity and efficiency of neighbor analyses in comparison with classical complete and incomplete block analyses of field experiments, Agronomy Journal **91**, 721–731.

[71] Zimmerman, D.L. & Harville, D.A. (1991). A random field approach to the analysis of field-plot experiments and other spatial experiments, Biometrics **47**, 223–239.