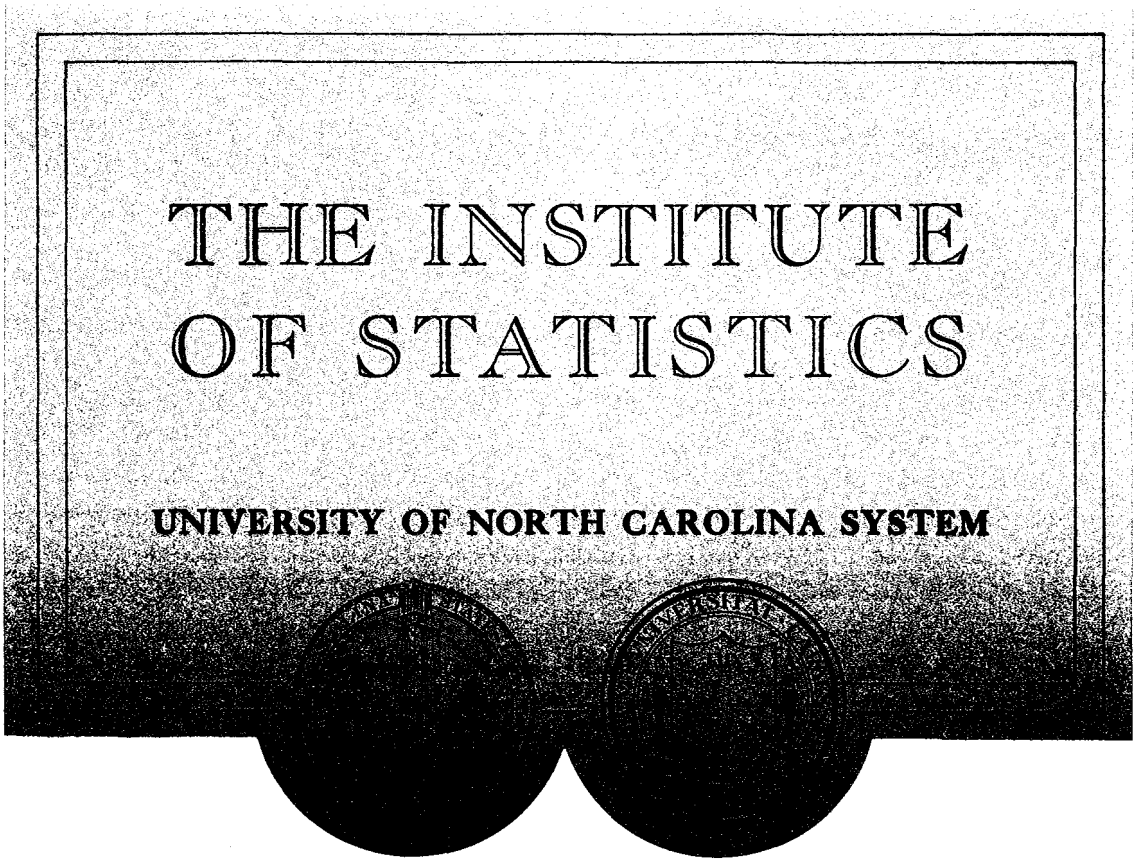

P. M. Dixon

Testing Spatial Independence in Multivariate Point Processes

Mimeo Series # 2215

NORTH CAROLINA STATE UNIVERSITY
Raleigh, North Carolina

Mimeo Series #2215

Savannah River Ecology Laboratory is operated by the University of Georgia for the Department of Energy under contract DE-AC09-76SROO-819.

TITLE: Testing Spatial Independence in Multivariate Point Processes

AUTHOR(S): P. M. Dixon

SUBMITTED TO: Biometrics

UPD 4702/10-89

The University of Georgia

Savannah River Ecology Laboratory

Testing Spatial Independence in Multivariate Point Processes

Philip Dixon

Savannah River Ecology Lab, University of Georgia

Drawer E

Aiken SC 29802-0005 USA

and

Biomathematics Program, Dept. of Statistics,

North Carolina State University, Raleigh NC 27695

November 3, 1991

Summary

Segregation, one aspect of the spatial relationship between two sets of locations, occurs when points are associated with like-type points, rather than being independently associated. Pielou proposed a 1 d.f. chi-square test of spatial independence using a nearest-neighbor contingency table, where each point is classified by its label and the label of its nearest neighbor. Pielou's test is inappropriate if all locations within a study area are mapped. Moments of cell counts in the nearest-neighbor contingency table are derived under the hypothesis that labels are independently assigned to points. These suggest both a new 2 d.f. test of segregation and separate 1 d.f. tests that specifically assess the segregation of each type of point. These results are extended to processes with k types of points. The overall test statistic has k(k-1) d.f.; each of the k type-specific tests has k-1 d.f. In small samples, the proposed test has the appropriate size, unlike the Pielou test. An appropriate odds ratio statistic is proposed as a measure of the degree of segregation. The methods are illustrated with three examples: Pielou's Douglas Fir / Ponderosa Pine data, locations of male, female, and juvenile water tupelo trees, and locations of five tree species in a hardwood swamp.

Key words: Spatial pattern, segregation, nearest neighbor, Pielou test.

1 Introduction

Ecologists are frequently interested in the relationship between two spatial point processes. Two sets of locations may be segregated (individuals occur near like individuals), associated (individuals occur near unlike individuals), or independent. Tests of spatial independence have been applied to locations of nests of different ant species (Harkness and Isham 1983), locations of different plant species (Pielou 1961, Diggle 1983, Whipple 1980), and locations of different sexes of dioecious plants (Bawa and Opler 1977).

Spatial independence in marked point processes is often tested by a nearest-neighbor procedure devised by Pielou (1961). A contingency table is constructed by cross-classifying each individual by its identity and the identity of its nearest neighbor (Table 1). Pielou (1961) proposed testing spatial independence by calculating a 1 d.f. chi-square test of independence, including Yate's correction:

$$\chi_1^2 = \sum_{i,j} \frac{(\mid N_{ij} - EN_{ij} \mid -0.5)^2}{EN_{ij}} \tag{1}$$

where, $EN_{ij} = \frac{N_{i.}N_{.j}}{N}$. $N_{i.}$ and $N_{.j}$ are the row i and column j marginal totals, and N is the total number of observations.

Insert Table 1 near here.

Less frequently used is Pielou's measure of segregation,

$$S = 1 - \frac{N_{AB} + N_{BA}}{EN_{AB} + EN_{BA}} \tag{2}$$

which ranges from -1 to 1 (Pielou 1961). Values of 0 indicates spatially independent populations, 1 indicates extreme segregation (i.e. clumps of type A points well separated from clumps of type B points), and -1 indicates extreme association (i.e. isolated AB pairs) (Pielou 1961).

Another characteristic of nearest neighbor relationships is whether they are symmetrical ($N_{AB} = N_{BA}$ in a population with two types) or unsymmetrical ($N_{AB} \neq N_{BA}$) (Pielou 1961). Pielou proposed a test of symmetry using a 1 d.f. χ^2 statistic

$$\chi_1^2 = \frac{(\mid N_{AB} - N_{BA} \mid -1)^2}{N_{AB} + N_{BA}} \tag{3}$$

Distributions of nearest-neighbor statistics depend on whether they are calculated from completely sampled data, in which all points within an area are mapped, or sparsely sampled data (Brown and Rothery 1978). Pielou demonstrated her tests using completely mapped data, but they have since been applied indiscriminantly to both completely mapped and sparsely sampled data. Application of (1) and (3) to completely mapped data is inappropriate because the events in the contingency table are not independent (de Vos 1973, Meagher and Burdick 1980). Meagher and Burdick demonstrated the problems caused by the lack of independence in completely mapped data and recommended that Monte-Carlo simulation be used to calculate critical values.

In this paper, I reconsider the application of the Pielou tests of spatial independence and symmetry. In section 2, two definitions of spatial independence are presented and used to derive expected cell counts and the variance-covariance matrix of the cell counts. These

expectations lead us to new tests and measures of spatial segregation, presented in section 3. Small sample properties of both old and new tests are assessed by Monte-Carlo simulations, reported in section 4. In section 5, the new tests and measures are demonstrated using Pielou's data on spatial patterns in a Ponderosa Pine/Douglas Fir woodland, data from a study of spatial segregation between sexes of water tupelo trees, and data on locations of five tree species in a swamp.

2 Moments of the cell counts

2.1 Expected cell counts for homogeneous Poisson processes

Expected cell counts for the nearest neighbor contingency table (table 1) can be derived under either of two specifications of independence of locations. The first is that the locations of each type of point are realizations of independent homogeneous planar Poisson processes. In this case, the two types are independent and the univariate process, ignoring the identity of the points, is a homogeneous Poisson process.

Consider a bivariate process with two types of points (A and B). The event that a point is type A and a neighbor is type A is the same as the event that the distance from a type A point to the nearest type A neighbor is less than the distance to the nearest type B neighbor. For independent Poisson processes, the large sample p.d.f. of D_{ij}, the distance

from a point of type i to the nearest point of type j, is (Ripley 1981)

$$f_{ij}(x) = 2\rho_j \pi x e^{-\rho_j \pi x^2}$$

where ρ_j = density of points of type j (number per unit area). D_{ij} and D_{jj} are independent, for independent Poisson processes (Ripley 1981), so the joint density of (D_{ij}, D_{jj}) factorizes. Hence,

$$\begin{aligned} \text{P[neighbor is A} \mid \text{point is A]} &= P[D_{AA} < D_{AB}] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{y} f_{AA}(x) f_{AB}(y)\, dx\, dy \\ &= \frac{\rho_A}{\rho_A + \rho_B} \end{aligned}$$

By standard calculations,

$$\begin{aligned} EN_{AA} &= \{EN\}\ \{\text{P[point is A} \mid \text{N]}\ \}\ \{\text{P[neighbor is A} \mid \text{point is A]}\ \} \\ &= \{R(\rho_A + \rho_B)\}\ \{\frac{\rho_A}{\rho_A + \rho_B}\}\ \{\frac{\rho_A}{\rho_A + \rho_B}\} \\ &= R\,\rho_A \frac{\rho_A}{\rho_A + \rho_B} \end{aligned}$$

where R = area of the region under study and N is the observed number of points in a realization of the process. In general, expected cell counts for large samples from independent Poisson processes are:

$$EN_{ij} = R\,\rho_i \frac{\rho_j}{\rho_i + \rho_j}.$$

If each density is estimated by the obvious estimator, N_i/R, the expected cell counts are:

$$EN_{ij} = N_i N_j / N \tag{4}$$

A Binomial process is a Poisson process conditioned on N_A and N_B, the observed number of type A and type B points. The distributions of distance to nearest neighbor in Binomial processes are slightly different. For distances to points of the same type, e.g. D_{AA}, the p.d.f is: $\frac{2\pi(N_A-1)d_{AA}}{R}\exp(-\pi d_{AA}^2(N_A-1)/R)$. For dissimilar types of points, e.g. D_{BA}, the p.d.f. is: $\frac{2\pi N_A d_{BA}}{R}\exp(-\pi N_A d_{BA}^2/R)$, where N_A/R is the observed density of type A points in a region with area R. Hence, the expected cell counts are:

$$EN_{AA} = N_A\frac{(N_A-1)}{(N-1)} \tag{5}$$

$$EN_{AB} = N_A\frac{N_B}{(N-1)} \tag{6}$$

$$EN_{BA} = N_B\frac{N_A}{(N-1)} \tag{7}$$

$$EN_{BB} = N_B\frac{(N_B-1)}{(N-1)} \tag{8}$$

In general, expected cell counts for a multivariate process are:

$$EN_{ij} = \begin{cases} N_i(N_i-1)/(N-1) & \text{if } i=j \\ N_iN_j/(N-1) & \text{if } i\neq j \end{cases} \tag{9}$$

2.2 Expected cell counts under independent labelling

An alternative definition of spatial independence can be derived by conditioning on the observed locations and considering only the distribution of the labels assigned to locations. Two processes are spatially independent if the observed labels are a realization of a simple random sample, without replacement, from the pool of N_A type A labels and N_B type B labels. Conditional on the locations, each point has a fixed nearest neighbor, and P[neighbor is type i | point is type j] can be derived by a counting argument. Using a type

A point as an example, its nearest neighbor may be assigned one of $N-1$ equally probable labels. $N_A - 1$ of those labels are A, hence P[neighbor is A | point is A] = $(N_A - 1)/(N-1)$. The expected cell counts under randomized labels are the same as those from independent Binomial processes (5 - 8).

Note that the expected cell counts depend only on the numbers of each type of point, i.e. the row marginal totals of the contingency table. This is true for both definitions of spatial independence. Unlike the usual test of independence in a 2x2 table and Pielou's suggested test, the expected cell counts do not depend on the column marginal totals.

2.3 Variances and covariances under independent labelling

Observations in the nearest-neighbor contingency table (Table 1) are not independent, because of reflexive nearest neighbors (Cox 1981) and shared nearest neighbors (Moran 1948). None of the standard sampling models for contingency tables, e.g. multinomial, binomial, or hypergeometric, are appropriate. However, second moments of the cell counts can be derived for the randomized label hypothesis of spatial independence. Conditional on the observed locations, the moments of N_{xy} are special cases of the moments of join count statistics derived by Moran (1948).

For a bivariate process with fixed number of type A and type B individuals, the variance-covariance matrix of the count vector $(N_{AA}, N_{AB}, N_{BA}, N_{BB})'$ has a particularly

simple form (See appendix for details):

$$\sum = \begin{pmatrix} \text{Var } N_{AA} & -\text{Var } N_{AA} & -\text{Cov } N_{AA}N_{BB} & \text{Cov } N_{AA}N_{BB} \\ -\text{Var } N_{AA} & \text{Var } N_{AA} & \text{Cov } N_{AA}N_{BB} & -\text{Cov } N_{AA}N_{BB} \\ -\text{Cov } N_{AA}N_{BB} & \text{Cov } N_{AA}N_{BB} & \text{Var } N_{BB} & -\text{Var } N_{BB} \\ \text{Cov } N_{AA}N_{BB} & -\text{Cov } N_{AA}N_{BB} & -\text{Var } N_{BB} & \text{Var } N_{BB} \end{pmatrix}$$

The elements of this matrix can be calculated from three properties of the set of locations: N, R, and Q, and three probabilities: $P_{xx}, P_{xxx}, P_{xxxx}$.

$$\text{Var } N_{AA} = (N+R)P_{AA}+(2N-2R+Q)P_{AAA}+(N^2-3N-Q+R)P_{AAAA}-N^2P_{AA}^2 \quad (10)$$

$$\text{Var } N_{BB} = (N+R)P_{BB}+(2N-2R+Q)P_{BBB}+(N^2-3N-Q+R)P_{BBBB}-N^2P_{BB}^2 \quad (11)$$

$$\text{Cov } N_{AA}N_{BB} = (N^2-3N-Q+R)P_{AABB}-N^2P_{AA}P_{BB} \quad (12)$$

$P_{xx}, P_{xxx}, P_{xxxx}$, the probabilities that a randomly chosen pair, triple, or quartet of points, respectively, have the indicated labels. N is the number of locations, R is the number of points with reflexive nearest neighbors, and Q is the number of points that share a nearest neighbor. In practice, R and Q may be calculated from the given set of locations, or they can be approximated by their expected values for some spatial process that generates the set of locations. For planar poisson processes, ER is approximately 0.6215 N (Cox 1981) and EQ is approximately 0.6332 N (Cusick and Edwards 1990). Note that Σ for two types of points has rank = 2 and that the variances of the cell counts are all larger than that expected under binomial sampling, $N_i p_j(1-p_j)$.

The variance-covariance matrix for a multivariate process contains other covariances besides Cov $N_{AA}N_{BB}$. Derivations of all the variance and covariance elements are presented in the appendix. If there are k types of labels, Σ has rank k(k-1).

2.4 Asymptotic distribution of the cell counts

Asymptotic normality of N_{AA} has been proven by Cusick and Edwards (1990). By symmetry, N_{BB} also has an asymptotic normal distribution. For bivariate processes, asymptotic normality of N_{AB} and N_{BA} follows from the relationships $N_{AB} = N_A - N_{AA}$ and $N_{BA} = N_B - N_{BB}$. The machinery used by Cusick and Edwards is appropriate for cell counts of the form N_{ii} from multivariate processes, but it is not appropriate for off-diagonal cell counts of the form N_{ij}. However, simulation results suggest that all cell counts from multivariate processes are approximately normally distributed when the marginal totals are large.

3 Tests of spatial independence and symmetry

3.1 Overall test of spatial independence

The asymptotic distribution of the cell counts suggests a test of spatial independence based on the quadratic form

$$C = (\mathbf{N} - \mathrm{E}\mathbf{N})'\mathbf{\Sigma}^{-}(\mathbf{N} - \mathrm{E}\mathbf{N}) \tag{13}$$

where $\mathbf{N} = (N_{AA}, N_{AB}, N_{BA}, N_{BB})'$ and $\mathbf{\Sigma}^{-}$ is a generalized inverse of $\mathbf{\Sigma}$ For bivariate processes, (13) simplifies to

$$\begin{aligned} C \;=\; & \tfrac{1}{1-r^2}\left(\frac{(N_{AA} - EN_{AA})^2}{\mathrm{Var}\ N_{AA}} + \frac{(N_{BB} - EN_{BB})^2}{\mathrm{Var}\ N_{BB}}\right. \\ & \left. - \frac{2r\,(N_{AA} - EN_{AA})(N_{BB} - EN_{BB})}{\sqrt{\mathrm{Var}\ N_{AA}\mathrm{Var}\ N_{BB}}}\right) \end{aligned} \tag{14}$$

where $r = \mathrm{Cov}\ N_{AA}N_{BB}/\sqrt{\mathrm{Var}\ N_{AA}\mathrm{Var}\ N_{BB}}$. C has an asympototic Chi-square distribution with 2 degrees of freedom, under the null hypothesis of spatial independence, because $\mathbf{\Sigma}$ has rank = 2 (Rao 1973).

3.2 Type-specific and Type/Neighbor-specific tests

The form of the test statistic immediately suggests two further tests of spatial independence:

$$C_{AA} \;=\; \frac{N_{AA} - EN_{AA}}{\sqrt{\mathrm{Var}\ N_{AA}}} \tag{15}$$

$$C_{BB} = \frac{N_{BB} - EN_{BB}}{\sqrt{\text{Var } N_{BB}}} \tag{16}$$

Under the null hypothesis of spatial independence, each statistic has an asymptotic N(0,1) distribution. The C_{AA} and C_{BB} statistics have useful ecological interpretations, although they are not independent. For example, C_{AA} assesses whether the neighbors of type A points occur in the proportions expected under random labelling. Patterns of spatial association need not be the same for each type of point (Pielou 1961), so independent labelling of neighbors of one type of point does not imply independent labelling for neighbors of the other type of point. The C_{AA} and C_{BB} statistics provide separate tests of independent labelling for neighbors of each type of point, or they may be combined into C, the overall test of spatial independence.

In processes with more than two types of points, it is possible to define type-specific tests by testing subsets of the entries in the nearest-neighbor contingency table (Table 1). The hypothesis that the neighbors of a particular type of point occur in the same proportion as expected under random labelling can be tested by computing the (k-1) d.f. Chi-square statistic for that row of the contingency table, using the expectations (9), variances, and covariances (Appendix A) of the cells in that row. A more specific hypothesis that an observed cell count is equal to its expectation under random labelling can be tested by a type/neighbor-specific test. These can be computed using a z-statistic like equations (15-16). In both multivariate and bivariate processes, the type-specific and type/neighbor-specific tests are not independent of each other, because of the covariance structure of the cell counts.

3.3 Measures of segregation

Pielou's coefficient of segregation for a bivariate process (2) measures the relative lack of AB and BA point-neighbor pairs. The expected cell counts (5-8) suggest type-specific measures of association based on the log odds-ratios:

$$S_{AA} = \log \frac{N_{AA}/N_{AB}}{(N_A - 1)/N_B} \tag{17}$$

$$S_{BB} = \log \frac{N_{BB}/N_{BA}}{(N_B - 1)/N_A} \tag{18}$$

S_{AA} is the log odds that a type A point has a type A neighbor, relative to the odds that a randomly chosen point is type A. Values of S_{AA} near 0 suggest random labelling of neighbors of type A points. Values larger than 0 indicate that type A points cluster with other type A points, while values less than 0 indicate a lack of type A neighbors around type A points. Note that these odds-ratios are not the usual odds-ratios for 2x2 contingency tables. When there are more than two types of points, a coefficient of segregation can be defined for each combination of type of point and type of neighbor. A point/neighbor-specific coefficient of segregation is:

$$S_{ij} = \log \frac{N_{ij}/(N_i - N_{ij})}{(N_j - 1)/(N - N_j + 1)} \tag{19}$$

S_{ij} is the log odds that a type i point has a type j neighbor, relative to the odds that a randomly chosen point is type j. For bivariate processes, S_{AB} and S_{BA} can be defined, but $S_{AB} = -S_{AA}$ and $S_{BA} = -S_{BB}$.

3.4 Tests and measures of symmetry

Pielou's test of symmetry (3) may be replaced in bivariate processes by:

$$\begin{aligned} z &= \frac{N_{BA} - N_{AB}}{\sqrt{\text{Var}\ (N_{BA} - N_{AB})}} \\ &= \frac{N_{BA} - N_{AB}}{\sqrt{\text{Var}\ N_{AB} + \text{Var}\ N_{AB} - 2\text{Cov}\ N_{AB}N_{BA}}} \end{aligned} \tag{20}$$

By sustituting $N_{AB} = N_A - N_{AA}$ into (20), one can see that the test of $N_{BA} - N_{AB} = 0$ is identical to a test of $N_{BA} + N_{AA} = N_A$, or equivalently $N_{.A} = N_{A.}$. In this form, the symmetry test can be generalized to k-type processes as a test of whether the frequencies that types of points occur as nearest neighbors (the column totals in table 1) are proportional to their frequencies in the population (the row totals). $\mathbf{\Sigma}^*$, the variance-covariance matrix of the column totals, can be calculated from $\mathbf{\Sigma}$, the variance-covariance matrix of the counts. Under the null hypothesis of random labelling, the statistic

$$\chi^2 = (\mathbf{N}_{\cdot\mathbf{i}} - \mathbf{N}_{\mathbf{i}})'\mathbf{\Sigma}^{*-}(\mathbf{N}_{\cdot\mathbf{i}} - \mathbf{N}_{\mathbf{i}}) \tag{21}$$

has an asymptotic χ^2 distribution with k-1 degrees of freedom. In equation (21), $\mathbf{N}_{\cdot\mathbf{i}}$ is the vector of column marginal totals, $\mathbf{N}_{\mathbf{i}}$ is the vector of row marginal totals, and $\mathbf{\Sigma}^{*-}$ is a generalized inverse of $\mathbf{\Sigma}^*$.

4 Small sample properties

In the previous section, it was shown that the variances of the cell counts are larger than that expected by a binomial sampling model. Hence, the Pielou test may reject H_0 too frequently. The empirical size of each test under the null hypothesis of spatial independence was computed by Monte-Carlo simulation. 1900 realizations each of bivariate homogeneous Binomial processes were generated with frequencies, (N_A, N_B), fixed at (25,25), (10,90), (20,80), (30,70), (40,60), (50,50), (25,225), (50,200), (125,125), and (250,250). These frequencies span the range of abundance and relative proportion typically seen in ecological studies. The number of replications was chosen so that the standard error of the rejection proportion was less than 0.5%, assuming that a test rejects at $\alpha =$ 5%. Computations were done using GAUSS version 2.1 on a Northgate 386 PC.

The empirical size of the Pielou test exceeds the nominal size for all the simulated densities (Table 2). In some cases, the empirical size of a nominal 5% Pielou test exceeds 13%. Empirical sizes of tests with Yates' continuity correction are smaller, but can still exceed 10%. The difference between empirical and nominal sizes increases with total density and is usually largest with equal densities of each type of point. In contrast, the 2 d.f. test of independence (14), the 1 d.f. type-specific tests (15), and the 1 d.f. test of symmetry (20) have empirical rejection rates close to nominal rates.

Insert table 2 near here.

Similar results are observed when locations are generated from a heterogenous Poisson

process but labels are independently assigned (Table 3). To generate locations from a heterogeneous process, the relative intensity of a Poisson process was shifted from 0.2 to 0.8 in alternate squares of a 4x4 checkerboard. Such a process can be quickly simulated by transformation of independent U(0,1) pseudo-random numbers. The empirical sizes of the new tests are not significantly different from nominal sizes across a wide range of densities and relative proportions, but the empirical size of Pielou's test (with or without Yates' correction) usually exceed the nominal size.

Insert table 3 near here.

5 Examples

5.1 Pielou's data

Pielou demonstrated her test using data from a completely enumerated stand of Ponderosa Pine (N=160) and Douglas Fir (N=68) trees in British Columbia, Canada (Pielou 1961). Douglas Fir comprises 70% of the trees, but firs are found as the nearest-neighbor of 85% of the firs, and only 56% of the pines (Table 4). Pielou concluded that the two species were spatially segregated, based on a 1 d.f. χ^2 test statistic of 22.02 (with Yates correction). However, her test of symmetry was not significant (with Yates correction: $\chi^2 = 3.213$, 1 d.f., $p = 0.073$; without Yates correction: $\chi^2 = 3.6885$, 1 d.f., $p = 0.055$).

Insert table 4 near here.

To apply the new tests, one needs to know the values of R and Q, the number of trees in reflexive nearest-neighbor pairs and number of trees that are shared neighbors. Pielou reported 67 reflexive pairs, so R = 134. She also reported that 42 trees were nearest-neighbor of 2 other trees, 13 trees were nearest-neighbor to 3 other trees, and no trees were nearest-neighbor to more than 3 trees. From these data, Q can be calculated as $Q = 2(42\binom{2}{2} + 13\binom{3}{2}) = 162$. Using equations (10, 11, and 12), the variance-covariance matrix of the count vector, $(N_{FF}, N_{FP}, N_{PF}, N_{PP})'$, is

$$\begin{bmatrix} 32.68 & -32.68 & -8.72 & 8.72 \\ -32.68 & 32.68 & 8.72 & -8.72 \\ -8.72 & 8.72 & 18.82 & -18.82 \\ 8.72 & -8.72 & -18.82 & 18.82 \end{bmatrix}$$

The expected cell counts are (112.07, 47.93, 47.93, 20.07)', calculated using equations (5 - 8).

The χ^2 test statistic for independent labelling of points (equation 14) is 19.67 with 2 d.f, which is significant at $p \leq 0.0001$. Each species is associated with neighbors of its own type (Log odds-ratio = 0.41 for firs-firs, and 0.28 for pines-pines, equations 17, 18). Using the z tests of species-specific association (equations 15 and 16), the hypothesis that each species is independently associated is rejected. Finally, we re-examine the symmetry in the nearest-neighbor relationships. First, we calculate $\mathrm{Var}(N_{FF} + N_{PF})$ = 32.68 + 18.82 + 2(-8.72) = 34.05. Then, using (20), z = (137+38 - 160)/$\sqrt{34.05}$ = 2.57 (p = 0.010). Nearest-neighbor relationships are not symmetrical; a nearest neighbor is more likely to be

a fir than a pine.

5.2 Sex of water tupelo

Data from Shea *et al.* (in prep.) provide an application to a point process with three types of labels. They studied the small-scale spatial patterning of water tupelo (*Nyssa aquatica*) trees in a riverine swamp along the Savannah River (South Carolina, USA). Tupelo is the most abundant tree in this stand, constituting 85% of the individuals larger than 2.5cm dbh. Tupelo is a dioecious tree; some individuals produce flowers with only male parts, while other individuals mostly produce flowers with only female parts. Locations of all trees in 50m x 50m quadrats were mapped. By inspection of flowers in spring, each tupelo was classified as male, female, or juvenile (did not flower). We will analyze data from one plot (plot 2) in which juveniles were abundant.

Plot 2 contained 257 tupelos (121 male, 104 female, and 32 juvenile). Observed frequencies of point/nearest-neighbor types are presented in Table 5. For each sex, the relative frequencies of each type of nearest-neighbor are close to their frequencies in the population, except for juveniles that are nearest neighbors to juveniles. Hence, most log odds-ratios (Table 6), measuring sex-specific segregration, are close to zero. However, a neighbors of a juvenile tree is more significantly more likely to be a juvenile, relative to the proportion of juveniles in the population.

Insert table 6 near here.

Under independent labelling, the vector of expected cell counts $(N_{MM}, N_{MF}, N_{MJ}, N_{FM}, N_{FF}, N_{FJ}, N_{JM}, N_{JF}, N_{JJ})'$ is (56.72, 49.16, 15.12, 49.16, 41.84, 13.00, 15.12, 13.00, 3.88)'. In this stand, R and Q are 164 and 162, respectively, so the variance-covariance matrix of the cell count vector, calculated using equations (23 - 35) is:

$$\begin{bmatrix}
35.15 & -26.88 & -8.27 & -12.29 & 9.38 & 2.91 & -3.78 & 2.91 & 0.87 \\
-26.88 & 30.11 & -3.23 & 13.29 & -11.88 & -1.41 & -1.00 & 0.25 & 0.75 \\
-8.27 & -3.23 & 11.50 & -1.00 & 2.50 & -1.50 & 4.78 & -3.16 & -1.62 \\
-12.29 & 13.29 & -1.00 & 28.04 & -24.40 & -3.64 & 0.66 & -1.41 & 0.75 \\
9.38 & -11.88 & 2.50 & -24.40 & 30.85 & -6.45 & 2.50 & -3.14 & 0.64 \\
2.91 & -1.41 & -1.50 & -3.64 & -6.45 & 10.10 & -3.16 & 4.55 & -1.39 \\
-3.78 & -1.00 & 4.78 & 0.66 & 2.50 & -3.16 & 8.17 & -5.39 & -2.78 \\
2.91 & 0.25 & -3.16 & -1.41 & -3.14 & 4.55 & -5.39 & 7.78 & -2.39 \\
0.87 & 0.75 & -1.62 & 0.75 & 0.64 & -1.39 & -2.78 & -2.39 & 5.17
\end{bmatrix}$$

An overall test of spatial indepdence can be constructed from the expected moments of the cell counts by using equation (13). The resulting χ^2 test statistic is 11.24 on 6 d.f, p = 0.081 (Table 7). This overall test statistic can be broken either into 2 d.f. components for each species (Table 7) or into z-scores for each cell of the contingency table. Inspection of these statistics suggests that these data provide slight evidence for segregation of juveniles, especially a clustering of juveniles adjacent to other juveniles (table 6), but no evidence for segregation of males and females. Finally, the symmetry of the nearest-neighbor relationships can be tested by comparing the frequencies with which species are nearest neighbors (the column sums of table 5) to the frequencies of species in the population. For

this plot, the nearest-neighbor relationships are not significantly asymmetric (Table 7). If there is any segregation of male and female individuals, as has been found in other dioecious species, this test is not sufficiently powerful to detect it.

Insert table 7 near here.

5.3 Swamp tree species

The final example comes from a study of the spatial patterns of tree species in the Savannah River Swamp (Good and Whipple 1982). All trees (stems > 4.5cm dbh) in a 200m x 50m plot were mapped and identified to species. Four species were common (Table 8). The remaining 60 stems, comprising 8 species, were lumped into the category "other species." Observed frequencies of point/nearest-neighbor pairs are given in Table 8. In the entire 1 ha mapped area, there were 734 trees, 454 reflexive nearest neighbors (R), and 472 joint nearest neighbors (Q).

Insert table 8 near here.

There is strong evidence of segregation for four of the tree species. The overall test of segregation is highly significant (table 9), and four of the five species-specific tests are highly significant (table 9). The one exception is bald cypress; species occur as nearest-neighbors of bald cypress in similar proportions to their proportion in the stand. The log odds ratios for each of the other four species to themselves as nearest-neighbors are large and positive (table 8). For these four species, their nearest-neighbors are more likely

to be the same species. The observed segregation of species may have a variety of ecological causes, including different microhabitat requirements for each species, patches with different disturbance histories, and the tendency of these species to produce multiple stems from a single rootstock. The lack of spatial segregation in cypress may be due to logging, which was concentrated on cypress (Sharitz, Irwin and Christy 1974)), or it may represent some other difference between cypress and the other tree species. Tests and measures of spatial segregation can indentify and describe patterns but they can not identify the processes creating the observed patterns.

Insert table 9 near here.

6 Acknowledgements

Research and manuscript preparation were supported by contract DE-AC09-76SROO-819 between the U.S. Department of Energy and the University of Georgia's Savannah River Ecology Laboratory. I thank Gene Schupp and an anonymous reviewer for comments on the manuscript.

7 References

Bawa, K.S. and Opler, P.A. 1977. Spatial relationships between staminate and pistillate plants of dioecious tropical forest trees. Evolution 31:64-68.

Brown, D. and Rothery, P. (1978). Randomness and local regularity of points in a plane. Biometrika 65:115-122.

Cox, T. 1981. Reflexive nearest neighbors. Biometrics 37:367-370.

Cusick, J. and Edwards, R. 1990. Spatial clustering for inhomogeneous populations (with discussion). Journal of the Royal Statistical Society, Series B 52:73-104.

Good, B.J. and Whipple, S.A. 1982. Tree spatial patterns: South Carolina bottomland and swamp forests. Bulletin of the Torrey Botanical Club 109:529-536.

Harkness, R.D. and Isham, V. (1983). A bivariate spatial point pattern of ants' nests. Applied Statistics 32:293-303.

Meagher, T.R. and Burdick, D.S. 1980. The use of nearest neighbor frequency analysis in studies of association. Ecology 61:1253-1255.

Moran, P.A.P. 1948. The interpretation of statistical maps. Journal of the Royal Statistical Society, Series B 10:243-251.

Pielou, E.C. 1961. Segregation and symmetry in two-species populations as studied by nearest-neighbour relationships. Journal of Ecology 49:255-269.

Rao, C.R. 1973. *Linear Statistical Inference and Its Applications.* 2nd edition. New York: Wiley.

Ripley, B.D. 1981. *Spatial Statistics.* New York: Wiley.

Sharitz, R.R., Irwin, J.E. and Christy, E.J. 1974. Vegetation of swamps receiving reactor effluent. Oikos 25:7-13.

Shea, M.M, Dixon, P.M., and Sharitz, R.R., in prep., Size differences, sex ratio and spatial distribution of male and female *Nyssa aquatica* (water tupelo).

Vos, S. de. 1973. The use of nearest neighbor methods. Tijdschrift voor Economische en Sociale Geographie 64:307-319.

Whipple, S.A. 1980. Population dispersion patterns of trees in a southern Louisiana hardwood forest. Bulletin of the Torrey Botanical Club 107:71-76.

Appendix: Variance-covariance matrix of the counts

First, I will derive the variances and covariances of the counts from a multivariate process with any number of types of points. Simplifications for bivariate processes are discussed later. Consider first Var N_{AA} and define:

X_{ij} = I(point i is type A and point j is type A)

w_{ij} = I(point j is the nearest neighbor of point i)

Then, $N_{AA} = \sum_{ij} w_{ij} X_{ij}$. Following Moran (1948),

$$\begin{aligned} EN_{AA} &= NEX_{ij} \\ EN_{AA}^2 &= E\sum_{ij} w_{ij}X_{ij} \sum_{kl} w_{kl}X_{kl} \\ &= \sum_{i\neq j} w_{ij}EX_{ij} + \sum_{i\neq j} w_{ij}w_{ji}EX_{ij} + \sum_{i\neq j\neq k} w_{ij}w_{ik}EX_{ij}X_{ik} + \sum_{i\neq j\neq k} w_{ij}w_{ki}EX_{ij}X_{ki} + \\ &\quad \sum_{i\neq j\neq k} w_{ij}w_{jk}EX_{ij}X_{jk} + \sum_{i\neq j\neq k} w_{ij}w_{kj}EX_{ij}X_{kj} + \sum_{i\neq j\neq k\neq l} w_{ij}w_{kl}EX_{ij}X_{kl} \end{aligned} \quad (22)$$

The expectations of the random variables and their products can be easily calculated. Define P_{xx}, P_{xxx}, and P_{xxxx} as the probability that two, three, or four points, respectively, will have the indicated labels. Then, sampling without replacement and conditioning on N_A, the observed number of type A points, the expectations are:

$$EX_{ij} = EX_{ij}^2 = EX_{ij}X_{ji} = \frac{N_A(N_A-1)}{N(N-1)} = P_{AA}$$

$$EX_{ij}X_{ki} = EX_{ij}X_{jk} = EX_{ij}X_{kj} = \frac{N_A(N_A-1)(N_A-2)}{N(N-1)(N-2)} = P_{AAA}$$

$$EX_{ij}X_{kl} = \frac{N_A(N_A-1)(N_A-2)(N_A-3)}{N(N-1)(N-2)(N-3)} = P_{AAAA}$$

W, the matrix of join statistics, with elements w_{ij}, is determined by the fixed pattern of nearest neighbors. W is not symmetrical except in unusual circumstances, but the properties of nearest neighbors allow the general expression (22) to be simplified.

$\sum_{i \neq j} w_{ij}^2$	$=$	N	because every point has exactly 1 nearest neighbor.
$\sum_{i \neq j \neq k} w_{ij} w_{ik}$	$=$	0	because no point has 2 nearest neighbors.
$\sum_{i \neq j} w_{ij} w_{ji}$	$=$	R	the number of reflexive nearest neighbors.
$\sum_{i \neq j \neq k} w_{ij} w_{ki}$	$=$	$N - \sum_{i \neq j} w_{ij} w_{ji}$	$= N - R.$
$\sum_{i \neq j \neq k} w_{ij} w_{jk}$	$=$	$N - \sum_{i \neq j} w_{ij} w_{ji}$	$= N - R.$
$\sum_{i \neq j \neq k} w_{ij} w_{kj}$	$=$	Q	the number of points that share a neighbor.
$\sum_{i \neq j \neq k \neq l} w_{ij} w_{kl}$	$=$	$N^2 - 3N - Q + R$	by difference.

Substituting into (22) gives:

$$EN_{AA}^2 = (N+R)P_{AA} + (2N - 2R + Q)P_{AAA} + (N^2 - 3N - Q + R)P_{AAAA}$$

Hence,

$$\text{Var } N_{AA} = (N+R)P_{AA} + (2N-2R+Q)P_{AAA} + (N^2-3N-Q+R)P_{AAAA} - N^2 P_{AA}^2 \quad (23)$$

Similar calculations lead to:

$$\text{Var } N_{BB} = (N+R)P_{BB} + (2N-2R+Q)P_{BBB} + (N^2-3N-Q+R)P_{BBBB} - (NP_{BB})^2 \quad (24)$$

$$\text{Var } N_{AB} = NP_{AB} + QP_{AAB} + (N^2 - 3N - Q + R)P_{AABB} - (NP_{AB})^2 \tag{25}$$

$$\text{Var } N_{BA} = NP_{AB} + QP_{ABB} + (N^2 - 3N - Q + R)P_{AABB} - (NP_{AB})^2 \tag{26}$$

To calculate Cov N_{AA} N_{BB}, define X_{ij} = I(points i and j are type A) and Y_{kl} = I(points k and l are type B). Then,

$$EX_{ij}Y_{ij} = EX_{ij}Y_{ji} = EX_{ij}Y_{ki} = EX_{ij}Y_{jk} = EX_{ij}Y_{kj} = 0$$

because one point can not be type A and type B simultaneously, and

$$EX_{ij}Y_{kl} = \frac{N_A(N_A - 1)(N_b)(N_b - 1)}{N(N-1)(N-2)(N-3)} = P_{AABB}$$

Hence,

$$\text{Cov } N_{AA}N_{BB} = (N^2 - 3N - Q + R)P_{AABB} - N^2P_{AA}P_{BB} \tag{27}$$

Similar calculations lead to the other covariance elements for multivariate processes:

$$\text{Cov } N_{AA}N_{AB} = (N - R)P_{AAB} + (N^2 - 3N - Q + R)P_{AAAB} - N^2P_{AA}P_{AB} \tag{28}$$

$$\text{Cov } N_{AA}N_{BA} = (N - R + Q)P_{AAB} + (N^2 - 3N - Q + R)P_{AAAB} - N^2P_{AA}P_{AB} \tag{29}$$

$$\text{Cov } N_{AA}N_{BC} = (N^2 - 3N - Q + R)P_{AABC} - N^2P_{AA}P_{BC} \tag{30}$$

$$\text{Cov } N_{AB}N_{AC} = (N^2 - 3N - Q + R)P_{AABC} - N^2P_{AB}P_{AC} \tag{31}$$

$$\text{Cov } N_{AB}N_{BA} = RP_{AB} + (N - R)(P_{AAB} + P_{ABB}) + (N^2 - 3N - Q + R)P_{AABB}$$

$$-N^2P_{AB}P_{BA} \tag{32}$$

$$\text{Cov } N_{AB}N_{BC} = (N-R)P_{ABC} + (N^2-3N-Q+R)P_{ABBC} - N^2P_{AB}P_{BC} \tag{33}$$

$$\text{Cov } N_{AB}N_{CB} = QP_{ABC} + (N^2-3N-Q+R)P_{ABBC} - N^2P_{AB}P_{BC} \tag{34}$$

$$\text{Cov } N_{AB}N_{CD} = (N^2-3N-Q+R)P_{ABCD} - N^2P_{AB}P_{CD} \tag{35}$$

where P_{xx}, P_{xxx}, and P_{xxxx} are the probabilities that a pair, triple, quartuple, respectively, of points have the indicated labels and R and Q are the numbers of points with reflexive nearest neighbors and shared nearest neighbors, respectively.

With only two types of points, the variance-covariance matrix of $(N_{AA}, N_{AB}, N_{BA}, N_{BB})'$ has a particularly simple form.

$$\sum = \begin{pmatrix} \text{Var } N_{AA} & -\text{Var } N_{AA} & -\text{Cov } N_{AA}N_{BB} & \text{Cov } N_{AA}N_{BB} \\ -\text{Var } N_{AA} & \text{Var } N_{AA} & \text{Cov } N_{AA}N_{BB} & -\text{Cov } N_{AA}N_{BB} \\ -\text{Cov } N_{AA}N_{BB} & \text{Cov } N_{AA}N_{BB} & \text{Var } N_{BB} & -\text{Var } N_{BB} \\ \text{Cov } N_{AA}N_{BB} & -\text{Cov } N_{AA}N_{BB} & -\text{Var } N_{BB} & \text{Var } N_{BB} \end{pmatrix}$$

This can be derived by substitution of $N_B = N - N_A$ into (25), (26), (28), (32) and (29), or by noting that $N_{AB} = N_A - N_{AA}$, $N_{BA} = N_B - N_{BB}$, and N_A and N_B are constants. The rank of $\sum = 2$, since Cov $N_{AA}N_{BB} \neq$ Var N_{AA} except in degenerate cases (e.g. $N_A = 0$ or $N_B = 0$).

Table 1: 2 x 2 contingency table used to test spatial association. Locations are classified by their label and the label of their nearest neighbor.

		Nearest neighbor A	B	Total
Label	A	N_{AA}	N_{AB}	$N_{A.}$
of point	B	N_{BA}	N_{BB}	$N_{B.}$
Total		$N_{.A}$	$N_{.B}$	N

Table 2: Rejection percentages for tests of spatial segregation under complete spatial randomness. Data from 1900 simulations of independent poisson processes with specified densities (N_A, N_B). Rejection percentages reported for $\alpha = 0.05$ and $\alpha = 0.01$ tests. Standard errors are approximately 0.5% for $\alpha = 0.05$ tests and 0.2% for $\alpha = 0.01$ tests.

Density	Pielou test		Pielou w/Yates		2 d.f. χ^2		1 d.f. χ^2 for A		Symmetry	
(N_A, N_B)	w/o Yates		Equ. (1)		Equ. (14)		Equ. (15)		Equ. (20)	
	α :0.05	0.01	α:0.05	0.01	α:0.05	0.01	α:0.05	0.01	α:0.05	0.01
(10,40)	10.7	3.3	4.1	1.1	5.0	1.2	5.9	1.5	5.6	0.9
(25,25)	14.3	4.3	9.6	2.5	5.2	0.6	5.0	1.0	5.5	1.0
(10,90)	7.7	3.6	3.7	1.9	5.3	1.7	3.2	3.2	5.0	0.7
(20,80)	13.7	4.2	7.5	2.3	5.2	1.2	4.6	1.0	4.6	0.7
(30,70)	12.3	4.4	9.4	2.9	5.1	1.0	5.2	0.8	4.6	1.1
(40,60)	12.6	4.0	8.8	2.5	4.5	0.7	4.9	1.0	5.7	1.0
(50,50)	12.8	4.6	9.8	2.9	4.3	0.6	3.8	0.8	4.8	1.0
(25,225)	8.7	3.7	4.8	2.2	5.1	1.5	5.2	1.2	4.7	0.6
(50,200)	11.5	3.8	8.8	2.7	4.4	1.2	5.5	1.0	3.8	0.9
(125,125)	12.9	4.3	10.7	3.8	5.4	0.9	5.1	0.8	4.9	0.9
(250,250)	13.6	4.3	11.8	3.3	4.2	0.7	5.0	0.9	4.0	0.7

Table 3: Rejection percentages under clustered locations but independent assignment of labels. Data from 1900 simulations of independent heterogeneous poisson processes with specified densities (ρ_A, ρ_B). Rejection percentages reported for $\alpha = 0.05$ and $\alpha = 0.01$ tests. Standard errors are approximately 0.5% for $\alpha = 0.05$ tests and 0.2% for $\alpha = 0.01$ tests.

Density (N_A, N_B)	Pielou test w/o Yates		Pielou w/Yates Equ. (1)		2 d.f. χ^2 Equ. (14)		1 d.f. χ^2 for A Equ. (15)		Symmetry Equ. (20)	
	α :0.05	0.01	α:0.05	0.01	α:0.05	0.01	α:0.05	0.01	α:0.05	0.01
(10,40)	9.7	3.4	4.4	1.2	4.7	0.8	4.8	1.6	4.2	0.8
(25,25)	14.5	4.2	9.3	2.2	4.5	0.8	4.7	1.0	4.8	0.9
(10,90)	7.0	3.8	4.0	2.2	5.3	1.4	3.2	3.2	4.9	0.6
(20,80)	11.5	2.9	6.3	1.6	4.2	0.8	4.0	0.6	4.2	0.6
(30,70)	13.3	4.1	8.6	2.7	5.1	1.0	5.8	0.7	5.0	0.8
(40,60)	11.1	3.7	7.9	2.7	5.0	0.6	4.6	0.5	5.4	0.8
(50,50)	14.4	5.1	9.8	3.8	5.6	0.8	5.2	0.9	5.4	0.9
(25,225)	7.3	2.9	3.9	1.8	4.6	1.2	5.0	0.7	4.8	0.6
(50,200)	12.2	3.7	8.8	2.2	4.0	0.9	4.2	1.2	4.0	0.5
(125,125)	12.1	3.8	9.5	2.9	5.9	1.0	5.0	1.2	5.6	1.2
(250,250)	14.0	4.3	11.8	3.5	4.7	0.9	4.5	0.9	5.2	1.1

Table 4: Nearest neighbor statistics for a completely mapped forest of Ponderosa Pine and Douglas Fir. Percents in the density column are percentages of each species in the population. Percents in the rest of the table are percentages of that type of neighbor, relative to the row total. Data from Pielou (1961).

	To Fir:	To Pine:	Density
From:			
Douglas Fir	137 (85%)	23 (14%)	160 (70%)
Ponderosa Pine	38 (56%)	30 (44%)	68 (30%)

Table 5: Nearest neighbor statistics for male, female, and juvenile water tupelo. Percents in the density column are percentages of each species in the population. Percents in the rest of the table are percentages of that type of neighbor, relative to the row total. Data from Shea et al., in prep. Plot 2.

	To male:	To female:	To juvenile:	Density
From male	63 (52%)	39 (32%)	19 (16%)	121 (47%)
From female	46 (44%)	46 (44%)	12 (12%)	104 (40%)
From juvenile	15 (47%)	8 (25%)	9 (28%)	32 (12%)

Table 6: Log odds-ratios measuring association among sexes of water tupelo. **Boldface** values are significantly different from 0, using species-specific z-tests. Data from Shea et al., in prep. Plot 2.

	To male:	To female:	To juvenile:
From male	0.09	-0.16	0.12
From female	-0.05	0.07	-0.04
From juvenile	0.00	-0.31	**0.45**

Table 7: Results of χ^2 tests for nearest neighbor data from Shea et al.

Test	d.f.	χ^2	P[$\geq \chi^2$]
Overall Segregation	6	11.24	0.081
From male trees	2	4.12	0.13
From female trees	2	0.56	0.76
From juveniles	2	6.12	0.047
Symmetry test	2	5.12	0.077

Table 8: Frequencies of nearest neighbors and log odds ratios for 5 tree species in a southeastern swamp forest. Log odds ratios are in (). **Boldface** values are species/neighbor combinations that are significantly different from that expected under random labelling, using z tests.

To: / From	Ash	W. Tupelo	S. Tupelo	Cypress	Other	Density
Carolina Ash	**82 (0.62)**	**23 (-0.38)**	**23 (-0.35)**	22 (0.03)	**6 (-0.35)**	156
Water Tupelo	**29 (-0.24)**	**112 (0.42)**	**40 (-0.23)**	20 (-0.18)	14 (-0.11)	215
Swamp Tupelo	**26 (-0.27)**	**38 (-0.26)**	**117 (0.54)**	**16 (-0.26)**	**8 (-0.34)**	205
Bald Cypress	**29 (0.19)**	**19 (-0.24)**	29 (0.03)	14 (0.04)	7 (-0.06)	98
Other Species	**5 (-0.47)**	**7 (-0.50)**	**8 (-0.40)**	7 (-0.07)	**33 (1.14)**	60

Table 9: Results of χ^2 tests for tree species data.

Test	d.f.	χ^2	$P[\geq \chi^2]$
Overall Segregation	20	275.6	< 0.0001
From Carolina Ash	4	70.9	< 0.0001
From Water Tupelo	4	41.2	< 0.0001
From Swamp Tupelo	4	65.1	< 0.0001
From Bald Cypress	4	7.1	0.13
From Other Species	4	117.5	< 0.0001
Symmetry	4	12.56	0.014