Population and Community Ecology for Insect Management and Conservation, Baumgärtner et al. (eds)
© 1998 Balkema, Rotterdam, ISBN 90 5410 930 0

Testing for correlation in the presence of spatial autocorrelation in insect count data

A. M. Liebhold
Northeastern Forest Experiment Station, USDA Forest Service, Morgantown, W.Va., USA
A. A. Sharov
Department of Entomology, Virginia Polytechnic Institute and State University Blacksburg, Va., USA

ABSTRACT: Spatial autocorrelation occurs when values of a variable sampled at nearby locations are more similar than those sampled at locations more distant from each other. Spatial autocorrelation can occur at multiple spatial scales or vary with spatial orientation (cardinal direction). It is common in ecological data. In landscape ecology research, scientists often are interested in testing the statistical association between two variables that have been sampled at multiple locations in space. The presence of spatial autocorrelation in one or both variables may violate the assumption of independence among samples and thereby inflate the degrees of freedom in the traditional test of significance of a Pearson correlation coeeficient. In this paper we used geostatistical simulation to illustrate this phenomenon and suggest strategies for overcoming this problem.

INTRODUCTION

In ecology, few generalizations are universal, yet nearly all data fulfill the generalization that values from samples taken near each other tend to be more similar than those taken farther apart. This tendency is termed spatial autocorrelation or spatial dependence (Cliff and Ord 1973; Rossi et al. 1992; Liebhold et al. 1993).

Despite this universal tendency, ecologists often assume that spatial autocorrelation does not exist. In many if not all studies, samples are replicated through space. The implicit assumption in this design is that separation of samples in space yields independent observations. This assumption is applied widely in ecological studies in which the association between two or more variables is tested. This assumption of independence is true in some cases, but in others, samples may not be separated by adequate distance such that spatial autocorrelation is negligible. Although this problem has been recognized by several authors (Student 1914; Bivand 1980; Clif and Ord 1981; Clifford et al. 1989), most ecologists apparently are unaware of it.

The study of landscape ecology has grown considerably over the last 20 years. Using a landscape approach, ecologists often address traditional ecological problems but focus on patterns and processes over large spatial scales. The adoption of this technique has been facilitated greatly by the advent of geographical information systems (GIS). A GIS is a computer program designed to input, output, and manipulate data that are referenced spatially. Often, landscape ecologists use a GIS to manipulate large data sets that form a matrix of cells (e.g., digital elevation models, remotely sensed vegetation data, gridded climatalogical data. While these exhaustive data are useful for examining landscape-level relationships, statistical problems may arise due to extensive autocorrelation in these and other data. The presence of spatial autocorrelation often violates the assumption of independence that is implicit in many statistical analyses.

In this paper we introduce geostatistical methods for quantifying spatial autocorrelation and examine how the presence of spatial autocorrelation affects error probabilities for tests of association/correlation. We also present an approach for testing for correlation between two spatially autocorrelated variables.

GEOSTATISTICAL MEASURES OF SPATIAL AUTOCORRELATION

Much effort has been invested in characterizing spatial patterns of insect densities. Most earlier studies have attempted to described spatial patterns with the use of dispersion indices such as $s^2/\bar{x}$ (David and Moore 1954), coefficients of Taylor's Power Law (Taylor 1984), I_d (Morista 1959), Lloyd's Patchiness Index (Lloyd 1967), and Iwao's patchiness regression coefficients (Iwao 1972). These indices focus on the frequency distribution of samples (most quantify the relationship of the sample variance to the mean) but ignore the spatial location of samples. This property produces certain undesirable effects: (*a*) these indices often fail to differentiate among different spatial patterns (Jumars et al. 1977), and (*b*) their descriptions of spatial pattern may be highly dependent on the size of sample units (Sawyer 1989).

By contrast, geostatistics is a branch of applied statistics in which both the values and locations of samples are used to describe and model spatial patterns (Isaaks and Srivastava 1989; Liebhold et al. 1996). These methods originally were developed for geological applications but recently there has been considerable interest in their application to ecological problems.

Let $z(x)$ represent the value of a variable at location x and let $z(x+h)$ represent the value of the same variable some h distance, or lag, away. Consider the set of all possible combinations of samples that are separated by h. One way to express the similarity or dissimilarity between the paired values is to plot them in a scattergram of $z(x)$ vs. $z(x+h)$, known as plot is known as an h-scattergram. If the difference between all the $z(x)$ and $z(x+h)$ is small, then the scatter of points will be close to the 45° line and the variable is autocorrelated. Alternatively, the larger the difference between the pairs, the more diffuse the scatter of points around the 45° or $z(x) = z(x+h)$ line. When h is small, the scatter of points will, on average, be "tighter"; when h is large, autocorrelation is less and the scatter typically is more diffuse.

The h-scattergrams can be useful models of the degree of similarity or dissimilarity between samples separated by a common distance, but they are not practical. Because too many h-scattergrams would be required to adequately characterize the spatial similarity for all samples and for all h, a meaningful summary of these h-scattergrams is required. The most familiar tool in geostatistics is the "semi-variogram" or simply "variogram." A variogram summarizes all h-scattergrams for all possible pairings of data for all significant h:

$$\hat{\gamma}(h) = \frac{\sum_{i=1}^{N(h)} [z(x_i) - z(x_i + h)]^2}{2N(h).} \qquad (1)$$

where $\hat{\gamma}(h)$ is the estimated variogram value for lag h and $N(h)$ is the number of pairs of points separated by h. Variograms plot the $\hat{\gamma}(h)$ as a function of distance h (Fig. 1). The variogram values can be

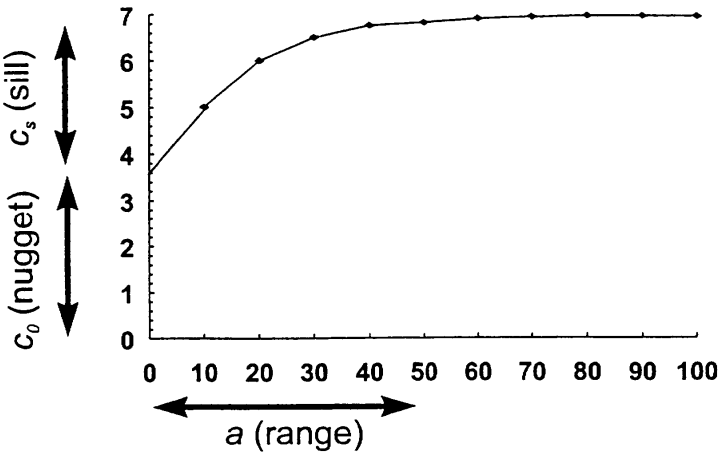

Figure 1. Typical variogram.

computed as averages over all directions, in which case the lag measure is scalar, or specific to a particular direction, in which case the lag measure is a vector.

Typically, when variogram values are plotted for all h, the values are small for low values of h; they increase with increasing distance, and usually level off or become constant after some distance (Fig. 1). Constant variogram values imply that the variance between values does not change with distance. Small values at short lags are indicative of data that are autocorrelated or spatially continuous. Large values, indicate that the paired samples are dissimilar and more spatially discontinuous.

If a variogram displays a leveling-off behavior, then the variogram value at which the plotted points level off is known as the "sill" (Fig. 1). The value of the sill usually is equivalent to the traditional sample variance. The distance at which the variogram values level off is known as the "range." The range designates the average distance within which the samples remain correlated spatially. Variograms that do not demonstrate a leveling off imply that the range is beyond the maximum lag distance analyzed.

Notice in (1) that, strictly speaking, $\hat{\gamma}(0) = 0$ since there is no variability between a sample and itself. In practice, however, when a variogram's scatter of points are extrapolated to lag zero, they often appear to intercept the ordinate at a value that is greater than zero. The variogram value at which the model appears to intercept the ordinate is known as the "nugget." A nugget represents two often co-occurring sources of variability. One source derives from spatial variability at a scale smaller than the minimum lag distance, so it cannot be modeled with the present sampling scheme. The other genesis of a nugget is experimental error, sometimes referred to as the "human nugget." Interpretations from variograms depend on the size of the nugget because the difference between the nugget and the sill (if there is one) represents the proportion of the total sample variance that can be modeled as spatial variability.

Variograms sometimes appear totally flat. These models have a complete lack of structure and their values are nearly identical to the sample variance over all h. In geostatistics these models are known as "pure nugget" variograms. Such variograms represent an absence of spatial dependence at the scale sampled. In this instance, the sample variance adequately summarizes the data's variability.

For entomological and environmental variables the similarity or dissimilarity between locations is rarely uniform with direction. For instance, one might expect, a priori, that a natural population of some insect would more often display different densities with different directions due to migration patterns or some environmental cue. When variograms computed for specific directions show different behaviors, the data are said to be "anisotropic."

Sample variograms often are not as regular as that shown in Figure 1. However, for other statistical procedures (e.g., kriging), it is necessary to fit a model to $\gamma(h)$ as a function of h to obtain values of $\gamma(h)$ for all values of h. Several forms of models used for variogram modeling; we describe only the exponential model of the form:

$$\gamma(h) = c_0 + c_s\left(1 - e^{-3h/a}\right) \tag{2}$$

where c_0 is the nugget effect, c_s is the sill, and a is the range.

HOW DOES SPATIAL AUTOCORRELATION AFFECT TESTS FOR SIGNIFICANCE OF CORRELATION?

The most widely used method for testing the correlation between two variables (a, b) is the computation of the Pearson product-moment correlation coefficient:

$$r_{ab} = \frac{s_{ab}}{s_a s_b} \tag{3}$$

where s_{ab} is the covariance between a and b and s_a is the standard deviation of a. We then test the null hypothesis that a and b are not correlated, $r_{ab} = 0$, by assuming that n independent samples of a and b are sampled from a bivariate normal distribution. It follows that the standard deviation of the correlation coefficient is:

$$s_r = \sqrt{\frac{1-r^2}{n-2}} \tag{4}$$

The null hypothesis is then tested as a t-test with n-2 degrees of freedom,

$$t_s = \frac{|r|}{s_r} \qquad (5)$$

$$= |r|\sqrt{\frac{n-2}{1-r^2}}$$

Critical values of r for a given sample size (n) and error probability (α) can be found in most books of statistical tables.

Spatial autocorrelation in a and b violates the assumption of independence among samples. How does this affect the distribution of r? To answer this question we simulated samples of a and b where a and b were independent of each other but spatially correlated to varying degrees. Sample values were simulated using a technique called "sequential unconditional Gaussian simulation", which generates spatially distributed values that correspond to a specified variogram. Details of this technique are found in Borgman et al. (1984),and Deutsch and Journel (1992). With this procedure, the first step is to define a random path that visits each node of the grid to be simulated. The value at the first node is simply drawn from a normal distribution. At each subsequent node on the path, the variogram model is used to determine the mean and variance of the conditional cumulative distribution function (CCDF) at that location. A value is then drawn randomly from a Gaussian distribution with this mean and variance and added to the data set. The algorithm then proceeds to the next node where the process is repeated until all nodes are simulated.

We generated a 50 x 50 grid (2,500 nodes) of values for all simulations. Simulations were computed using the SGSIM procedure in the GSLIB software package (Deutsch and Journel 1992). This procedure simulates data in Gaussian space (mean = 0, standard deviation = 1) that honors a specific variogram model of sill = 1.0. We assumed an exponential variogram model for all simulations. One thousand simulated grids were generated for each of the following five combinations of variogram parameters ① c_0 = 0, a = 0.2, ② c_0 = 0, a = 2.0, ③ c_0 = 0.5, a = 0.2, ④ c_0 = 0.5, a = 2.0, and , ⑤ c_0 = 1.0 (Fig. 2), where c_0 is the nugget effect, expressed as a fraction of the sill, c_s is 1 - c_0, and a is the range, expressed as a fraction of the simulation domain length (50 cells). When c_0 = 1.0, the variogram was pure nugget effect and simulated values had no autocorrelation structure.

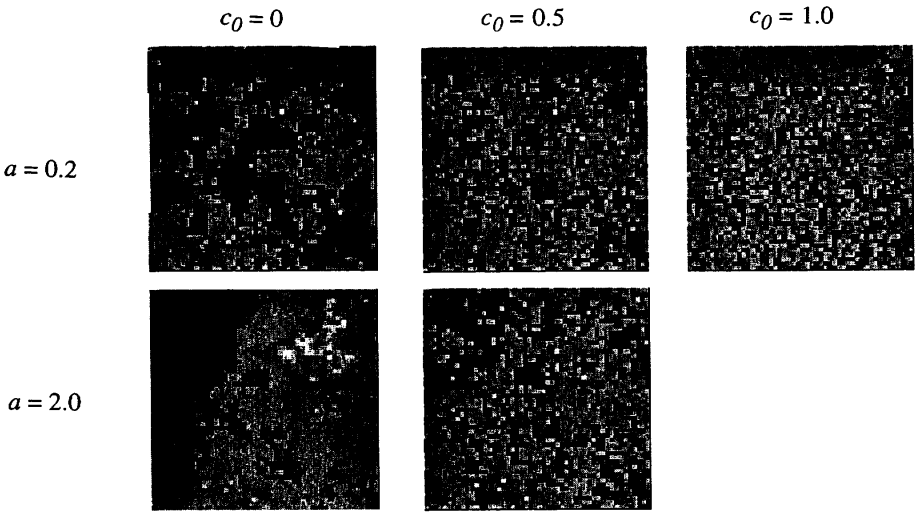

Figure 2. Examples of 50 x 50 grids of values simulated using various variogram parameters.

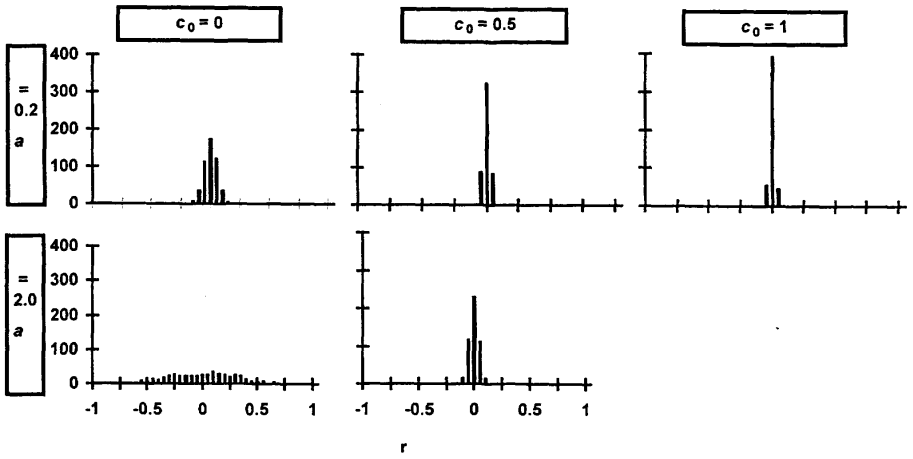

Figure 3. Frequency distribution histograms of correlation coefficients between simulated variables with identical variograms.

Table 1. Standard deviation of simulated correlation coefficients (all 2,500 points.)

Variogram Model Parameters	c_0= 0 a = 0.2	c_0= 0 a = 2.0	c_0= 0.5 a = 0.2	c_0= 0.5 a = 2.0	c_0= 1.0
c_0= 0 a = 0.2	0.0560	0.0913	0.0354	0.0463	0.0199
c_0= 0 a = 2.0		0.2933	0.0480	0.1006	0.0210
c_0= 0.5 a = 0.2			0.0267	0.0296	0.0200
c_0= 0.5 a = 2.0				0.0438	0.0197
c_0= 1.0					0.0199

Correlation coefficients of values were then computed between all combinations of these data sets. For example, 500 simulated grids for $c_0 = 0$, $a = 0.2$ were paired with 500 grids for $c_0 = 0.5$, $a = 0.2$. For each pair of simulations, values at each of the 2,500 grid nodes were paired and a correlation coefficient was calculated. Thus, obtained a frequency distribution of 500 correlation coefficients for each combination (Fig. 3). The standard deviations for simulated r are given in Table 1. When spatial autocorrelation was absent in either a or b, the frequency distribution of simulated correlation coefficients was the same as when autocorrelation was absent in both and a and b. However, when the nugget effect was less than 1 in both variables, the absolute value of r increased. The standard deviation of r was greatest when c_0 was small and a was large. Similar results were obtained when we used a random subset of 200 points rather than all 2,500 grid nodes (Table 2).

Table 2. Standard deviation of simulated correlation coefficients (200 randomly selected points.)

Variogram Model Parameters	c_0= 0 a = 0.2	c_0= 0 a = 2.0	c_0= 0.5 a = 0.2	c_0= 0.5 a = 2.0	c_0= 1.0
c_0= 0 a = 0.2	0.0834	0.1183	0.0813	0.0851	0.0729
c_0= 0 a = 2.0		0.3030	0.0853	0.1206	0.0739
c_0= 0.5 a = 0.2			0.0774	0.0768	0.0715
c_0= 0.5 a = 2.0				0.0854	0.0704
c_0= 1.0					0.0749

METHODS FOR TESTING FOR CORRELATION WHILE ADJUSTING FOR SPATIAL AUTOCORRELATION

As shown earlier, the use of conventional tests of the significance of correlation will yield incorrect results in the presence of spatial autocorrelation in both variables. The question then is: how do we modify these tests to account for this autocorrelation? One approach is to use unconditional simulation to simulate the spatial autocorrelation in each variable. By simulating uncorrelated variables with the same autocorrelation structure, we can estimate the probability that a greater r would occur by chance.

We illustrate this technique by examining the correlation between counts of gypsy moth, *Lymantria dispar* (L.), egg masses in 0.01 ha plots and elevation in the central Appalachian Mountains. In 1988, 4,822 egg-mass plots were sampled as part of a gypsy moth management program in this area (Reardon 1991; Liebhold et al. 1996). All egg-mass counts within the same 1-km^2 cell were averaged. Data on elevation for these same cells were obtained from U.S. Geological Survey digital elevation models (Elassal and Caruso 1983). The Pearson correlation coefficient between these two variables was calculated as r = 0.0402, n = 3,077. This value was greater than the critical value of 0.0355, so application of an ordinary test would indicate rejection of the null hypothesis of no correlation. Sample variograms were obtained for both the egg-mass and elevation data. The egg-mass variogram was modeled using an isotropic exponential model with $c_0 = 0.26$, $c_s = 0.74$, $a = 7.0$; the elevation data were fit using an anisotropic exponential model with $c_0 = 0$, $c_s = 1.0$, $a_{max} = 33.0$, $a_{min} = 20.0$, with the principal axis of anisotropy falling at a 45° angle (relative to the East - West line).

These variograms were used in the unconditional simulation procedure to generate five hundred 149 x 158 grids of simulated egg-mass and elevation data. Because egg mass data were lacking in many of the grid cells, those cells were deleted from both the simulated egg mass and elevation data. The data were then paired and a correlation coefficient was calculated for each of the 500 simulations. The frequency distribution of these correlation coefficients was examined. The observed value of r = 0.0402 was not greater than the 95 percentile or less than the 5 percentile of this distribution (r = 0.1782 for 1 tail test, r = 0.2177 for two-tailed test), so the null hypothesis of no correlation was not rejected.

One problem with this technique is that it is computationally intensive and requires that the user have access to software for conducting unconditional simulation. Clifford et al. (1989) developed a modified test of association that uses a Pearson correlation coefficient in which the degrees of freedom are adjusted by the spatial autocovariance in both variables. Our preliminary comparison of this method with our simulation technique indicates that they yield similar results. Ultimately, we believe that analytical methods such as these will be applied more widely in these situations.

ACKNOWLEDGMENTS

The authors thank B. Manly, J. Gurevitch, M. Hohn, and R.M. Muzika for conversations that led to this

research. We also thank F. Curriero and T. Jacob for reviewing an earlier draft of this manuscript. This research was funded in part by grants 95-37313-1892 and 95-37302-1905 from the USDA NRICGP.

REFERENCES

Bivand, R. 1980. A Monte Carlo study of correlation coefficient estimation with spatially autocorrelated observations. *Quaestiones Geographicae*. 6: 5-10.

Borgman, L.,M. Taheri, and R. Hagan. 1984. Three-dimensional frequency-domain simulations of geological variables. P. 517-541 in G. Verly et al. (eds.), *Geostatistics for Natural Resources Characterization.* Dordrecht, The Netherlands: Reidel.

Cliff, A.D., and J.K. Ord. 1973. *Spatial Autocorrelation*. London: Pion Press.

Cliff, A.D. and J.K. Ord. 1981. *Spatial Processes*. London: Pion Press.

Clifford, P. S. Richardson, and D. Hémon. 1989. Assessing the significance of the correlation between two spatial processes. *Biometrics* 45: 123-134.

David, F.N. and P.G. Moore. 1954. Notes on contagious distributions in plant populations. *Ann. Bot. Lond. N.S.* 18: 47-53.

Deutsch, C.V., and A.G. Journel. 1992. *GSLIB Geostatistical Software Library and User's Guide.* New York: Oxford University Press.

Elassal, A.A. and V.M. Caruso. 1983. Digital elevation models. *U.S. Geol. Surv. Circ.* 895-B.

Isaaks, E.H., and R.M. Srivastava. 1989. *An introduction to applied geostatistics*. New York: Oxford University Press.

Iwao, S. 1972. Application of the $\hat{m}$-m method to the analysis of spatial patterns by changing the quadrat size. *Res. Popul. Ecol.* 14: 97-128.

Jumars, P.A., D. Thistle, and M.L. Jones. 1977. Detecting two-dimensional spatial structure in biological data. *Oecologia* 28: 109-123.

Liebhold, A., E. Luzader, R. Reardon, A. Bullard, A. Roberts, F.W. Ravlin, S. DeLost, and B. Spears. 1996. Use of a geographical information system to evaluate regional treatment effects in a gypsy moth (Lepidoptera: Lymantriidae) management program. *J. Econ. Entomol.* 89: 1192-1203.

Liebhold, A.M., R.E. Rossi, and W.P. Kemp. 1993. Geostatistics and geographic information systems in applied insect ecology. *Annu. Rev. Entomol.* 38: 303-327.

Lloyd, M. 1967. Mean crowding. *J. Anim. Ecol.* 35: 1-30.

Morista, M. 1959. Measuring of the dispersion of individuals and analysis of the distributional patterns. *Mem. Fac. Sci. Kyushu Univ. E(Biol.)* 2: 215-235.

Reardon, R.C. 1991. Appalachian gypsy-moth integrated pest-management project. *For. Ecol. Manage.* 39: 107--112.

Rossi, R.E., D.J. Mulla, A.G. Journel and E.H. Franz. 1992. Geostatistical tools for modeling and interpreting ecological spatial dependence. *Ecol. Monogr.* 62: 277-314.

Sawyer, A.J. 1989. Inconstancy of Taylor's b: simulated sampling with different quadrat sizes and spatial distributions. *Res. Popul. Ecol.* 31:11-24.

Student. 1914. The elimination of spurious correlation due to position in time or space. *Biometrika* 10: 179-181.

Taylor, L.R. 1984. Assessing and interpreting the spatial distributions of insect populations. *Annu. Rev. Entomol.* 29: 321-357.