Spatial autocorrelation in biology
1. Methodology

ROBERT R. SOKAL, F.M.L.S.

Department of Ecology and Evolution, State University of New York at Stony Brook, Stony Brook, New York 11794, U.S.A.

AND

NEAL L. ODEN

Department of Ecology and Evolutionary Biology, The University of Michigan, Ann Arbor, Michigan 48109, U.S.A.

Spatial autocorrelation analysis tests whether the observed value of a nominal, ordinal, or interval variable at one locality is independent of values of the variable at neighbouring localities. The computation of autocorrelation coefficients for nominal, ordinal, and for interval data is illustrated, together with appropriate significance tests. The method is extended to include the computation of correlograms for spatial autocorrelation. These show the autocorrelation coefficient as a function of distance between pairs of localities being considered, and summarize the patterns of geographic variation exhibited by the response surface of any given variable.

Autocorrelation analysis is applied to microgeographic variation of allozyme frequencies in the snail *Helix aspersa*. Differences in variational patterns in two city blocks are interpreted.

The inferences that can be drawn from correlograms are discussed and illustrated with the aid of some artificially generated patterns. Computational formulae, expected values and standard errors are furnished in two appendices.

KEY WORDS: — spatial autocorrelation — geographic variation analysis — spatial variation patterns — *Helix aspersa*

CONTENTS

INTRODUCTION

In modern work in biological systematics and ecology one frequently encounters situations in which a variable is mapped onto a geographic space. As examples we may cite maps of the geographic variation of morphological characters or of gene frequencies, or distributional patterns of individual genotypes, or species abundances. In investigating these data it is important to discover whether the observed value of the variable at one locality is dependent on the values at neighbouring localities. If such dependence exists, the variable is said to exhibit spatial autocorrelation. In general, if high values of the variate at one locality are associated with high values at neighbouring localities, the spatial autocorrelation is positive; whereas when high and low values alternate, the spatial autocorrelation is negative.

Although the concept of spatial autocorrelation is not new to evolutionary and ecological research, there have been only sporadic attempts at formal treatment (Matern, 1960; Whittle, 1954). In recent years analytical geographers have begun to attack this problem with vigour and rigour and a recent monograph (Cliff & Ord, 1973) has summed up papers in this field for the past two decades. Much of this work published in technical geographical journals is unfamiliar to biologists. We have found it of considerable value and interest in our own research and hope to bring it to the attention of our colleagues by means of this article and its sequel (Sokal & Oden, 1978).

We shall first review the relatively elementary statistics needed to compute spatial autocorrelations for data coded in various ways and shall illustrate these computations with two small, completely worked out examples to enable readers to familiarize themselves with the method. Next, we report the application of spatial autocorrelation analysis to a larger, actual problem in evolutionary biology, microgeographic variation in the snail, *Helix aspersa*. In conjunction with this example, we extend the method of autocorrelation to analysis of the variation surface by means of correlograms. Finally, we examine the inferences concerning spatial variation patterns to be drawn from various configurations of spatial autocorrelations and correlograms, illustrating our findings with the *Helix* example. Formulas for computing autocorrelations and testing their significance are furnished in two appendices.

STATISTICAL INTRODUCTION

General considerations

Most biosystematic applications of spatial autocorrelation analysis will be based on a sample of localities from an area—or, more formally, a sample of

points in the plane. These points may be regularly distributed as in a grid or lattice resulting from systematic sampling studies typical for ecology. Alternatively, the points may be irregularly distributed over the area as in locality samples typical for geographic variation studies. The nature of the distribution of the points over the area—regular, or following some theoretical distribution—will not concern us directly. For purposes of spatial autocorrelation the distribution of points is considered as given.

Onto this collection of points there are mapped one or more variables, with one value per variable for each point. The variables may be nominal (categorical, e.g., colour morphs, such as light or dark; genotypes, such as ++, +b, or bb; or different species, such as A, B, C . . .), ordinal (ranked, e.g., a set of n localities may be ranked in order of population density for some species occurring there), or interval (continuous, e.g., morphological measurements, gene frequencies). The single value of the variable at each point may be based on a single observation (typical of nominal variables such as the genotype or species identification of an individual found at a point in an area) or it may be a statistic based on a sample of individuals taken from the locality, such as a character mean or a gene frequency based on a locality sample.

To obtain the spatial autocorrelation coefficient of a variable we shall correlate the values of that variable for pairs of points in the plane. However, not all pairs of points will be correlated, only those that are considered neighbours. The criteria for considering any pair of points neighbours (such points are also termed adjacent or connected) are left up to the investigator. In a regular grid some simple forms of connection are: rook's moves (orthogonal connections only; Fig. 1A), bishop's moves (diagonal connections only; Fig. 1B), and queen's moves (both orthogonal and diagonal connections; Fig. 1C). For irregularly distributed points various other types of connections may be established. They are usually some function of geographical proximity, but may be modified by consideration of the biological model, such as direction of gene flow or ecological barriers. Tobler (1975) furnishes a list of various schemes for connecting points. The scheme employed for examples in this paper is Gabriel-connectedness, discussed further on. Another scheme is featured in one of the examples in the companion paper (Sokal & Oden, 1978).

Once a scheme for connecting the points (localities) has been devised, the connections can be represented in two equivalent forms. One of these is a graph (in the graph-theoretical sense; see Busacker & Saaty, 1965), with the points constituting the vertices of the graph and the connections between points
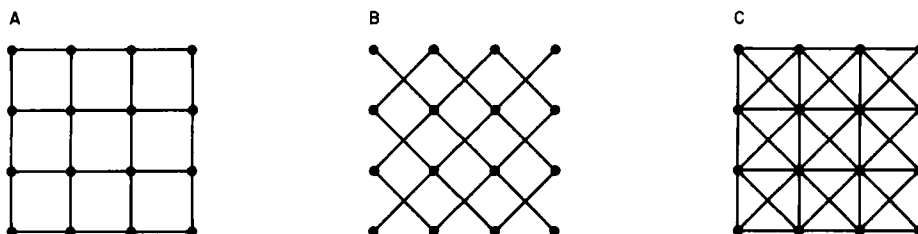

Figure 1. Simple connections in a regular grid. The 16 points or localities in this regular grid are connected by rook's moves in panel A, bishop's moves in panel B, and queen's moves in panel C.

comprising its edges. An example for a regular grid is shown in Fig. 2 and for irregularly-spaced points in Fig. 3. An equivalent representation is a square weight or adjacency matrix W of dimension n (for n localities). Tables 1 and 2 show the adjacency matrices for the graphs in Figs 2 and 3, respectively. In these examples a connection between two localities i and j is shown by setting elements w_{ij} and w_{ji} in the matrix equal to 1. For every pair of localities ij that is not connected $w_{ij} = w_{ji} = 0$. We need not always be restricted to binary weights when computing autocorrelations. Any pair of localities closer in some sense than some other pair can be given greater weight. Thus weights may range from zero to some arbitrarily large positive number, or, equivalently, the weights may be normalized to range from zero (unconnected) to unity (closest neighbours). Also, we have so far implied symmetry in the adjacency matrix by setting $w_{ij} = w_{ji}$; in most work on autocorrelation the distance from i to j is the same as the distance from j to i. Nevertheless, instances will occur in which the distance (in some sense) from i to j is not equal to that from j to i, as in upstream-downstream or uphill-downhill relations and in seed shadows. The formulas cited below handle symmetrical or asymmetrical matrices. Given a set of points or localities, a variable mapped onto these localities, and a graph or matrix representing the interconnections or weights, we can proceed to the computation of the autocorrelation coefficient.

In Appendix A we list computational formulas, expected values, and variances of various statistics of spatial autocorrelation. These formulas have been abstracted with some notational changes from a larger assemblage featured by Cliff & Ord (1973). They are divided into autocorrelation measures for nominal data and those for interval and ranked data.

Nominal data

For nominal data one calculates join counts. A join is a synonym for an edge connecting two points or localities. These joins can be between points that are like or unlike with respect to the categorical variable. Thus if the example is one of a spatial distribution of colour morphs (B for black and W for white), there will be edges connecting like individuals (BB and WW joins), and edges connecting unlike morphs (BW joins). The BB join count is the total number of edges in the study connecting two individuals of morph B. Clearly, the sum of the joins counts for BB, WW and BW will equal the total number of edges in the graph.

Cliff & Ord (1973) furnish expectations and variances for join counts in the nominal case for two hypotheses—sampling with replacement and sampling without replacement. In the first case the model assumes that the value of the variables for each locality is the result of a random sampling from a larger population in which the probability that the variable will assume value r is p_r. That is, the sampling outcomes are independently realized for each locality. In sampling without replacement, the assumption is that the proportion of localities in the study of type r is p_r, but the number of such localities is fixed so that the probability of any one locality being of type r is conditional on the number of localities of type r that have already been sampled. This is the appropriate assumption for most biological cases likely to be analyzed by this method and it is the only one considered here. In most situations we are

unlikely to know the parent distribution of types from which the observed spatial distribution has been sampled.

We shall now feature a sample computation for nominal data.

Example 1. Computation of spatial autocorrelation for a variable in nominal scale, regularly-spaced localities, binary weights, sampling without replacement.

The data are nine squares in a checkerboard pattern, as shown in Fig. 2. For this example the squares are connected by rook's moves; that is, each square is connected to all squares directly above, below, or to either side of it. The graph illustrating this relationship is shown in Fig. 2 and comprises 12 edges. The adjacency matrix describing this graph is presented in Table 1. Next to each row in the matrix is presented the type (colour) of the variable at the corresponding data point.

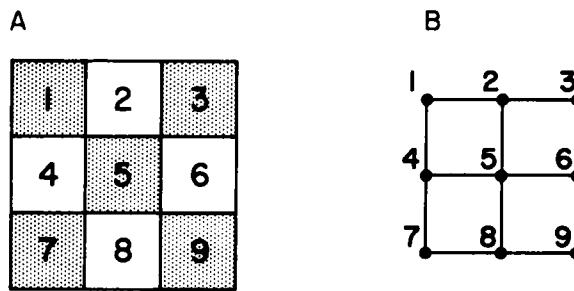A   B

Figure 2. The checkerboard of Example 1 connected by rook's moves. The regular 3 x 3 checkerboard pattern consists of five black and four white squares, numbered as shown in panel A. The squares, represented as vertices in a graph, are connected by rook's moves in panel B.

Table 1. Adjacency matrix for Example 1

Locality number	Type	Locality number								
		1	2	3	4	5	6	7	8	9
1	B	X	1	0	1	0	0	0	0	0
2	W	1	X	1	0	1	0	0	0	0
3	B	0	1	X	0	0	1	0	0	0
4	W	1	0	0	X	1	0	1	0	0
5	B	0	1	0	1	X	1	0	1	0
6	W	0	0	1	0	1	X	0	0	1
7	B	0	0	0	1	0	0	X	1	0
8	W	0	0	0	0	1	0	1	X	1
9	B	0	0	0	0	0	1	0	1	X

The total number of BB, BW, and WW joins may in this case be read directly from the checkerboard, but in more complex situations or if describing the data to a computer, these quantities may be calculated from the formula given in Appendix A. The number of BB joins equals $\frac{1}{2}\sum_{ij} w_{ij}(BB)_{ij}$. The variable $(BB)_{ij}$ equals 1 whenever localities i and j are both B, and zero otherwise. Thus for localities 1 and 2, $(BB)_{12} = 0$, but for localities 1 and 3, $(BB)_{13} = 1$. However, it can be seen that in this example all instances where $(BB)_{ij} = 1$ are paired with values of $w_{ij} = 0$ from the adjacency matrix. Thus the summation

equals zero; there are no BB joins in these data. The values for part of the example are shown in the table below.

Localities i	j	Weight w_{ij}	$(BB)_{ij}$	$(BW)_{ij}$
1	2	1	0	1
1	3	0	1	0
1	4	1	0	1
1	5	0	1	0
1	6	0	0	1
1	7	0	1	0
1	8	0	0	1
1	9	0	1	0
2	1	1	0	1
2	3	1	0	1
⋮	⋮	⋮	⋮	⋮
9	7	0	1	0
9	8	1	0	1
			$\sum_{ij} w_{ij}(BB)_{ij} = 0$	$\sum_{ij} w_{ij}(BW)_{ij} = 24$

The number of BW joins equals $\frac{1}{2}\sum_{ij} w_{ij}(BW)_{ij}$ where the variable $(BW)_{ij}$ equals 1 when locality i is B and j is W or when i is W and j is B. The variable is zero otherwise. Thus $(BW)_{ij} = 1$ for localities 1 and 2, but $(BW)_{ij} = 0$ for localities 1 and 3. Values for part of the example are shown in the table above. Summing $w_{ij}(BW)_{ij}$ over all locality pairs ij $(i \neq j)$ we obtain 24. Therefore, the number of BW joins is 12. Proceeding similarly we can obtain the number of WW joins, which is zero.

The formula for the expected value μ'_1 in Appendix A requires the quantity W, the sum of the matrix weights, which in this example with binary symmetrical weights equals the number of ones in the matrix, i.e., 24, or twice the number of edges connecting the graph. The total numbers of the different types of squares on the board is

$$n_{\text{B}} = 5 \qquad n_{\text{W}} = 4 \qquad n_{\text{total}} = 9.$$

Therefore, the expected number of BB joins is

$$\frac{Wn_{\text{B}}^{(2)}}{2n^{(2)}} = \frac{24(5)(4)}{2(9)(8)} = 3\tfrac{1}{3}.$$

Similarly, the expected number of WW joins = 2. The expected number of BW joins equals

$$\frac{Wn_{\text{W}}n_{\text{B}}}{n^{(2)}} = \frac{24(5)(4)}{9(8)} = 6\tfrac{2}{3}.$$

For the computation of the variances two further quantities are needed to evaluate the formula for μ_2 given in Appendix A. The first is $S_1 = \frac{1}{2}\sum_{ij}(w_{ij} + w_{ji})^2$. In this case with binary symmetric weights every edge between a pair of localities contributes $(1 + 1)^2 = 4$ to the sum and S_1 is simply $4 \times$ (number of edges in the graph) = 48. Quantity $S_2 = \sum_{i=1}^{n} (w_{i.} + w_{.i})^2$, where $w_{i.}$ and $w_{.i}$ are the sums of the ith row and ith column of the weight matrix

respectively. Since the weights used here are symmetric, S_2 is four times the sum of the squared marginal totals of either the rows or the columns $= 4(2^2 + 3^2 + 2^2 + 3^2 + 4^2 + 3^2 + 2^2 + 3^2 + 2^2) = 272$. For BB joins

$$\mu_2 = \frac{1}{4}\left[\frac{S_1 n_B{}^{(2)}}{n^{(2)}} + \frac{(S_2 - 2S_1)n_B{}^{(3)}}{n^{(3)}} + \frac{(W^2 + S_1 - S_2)n_B{}^{(4)}}{n^{(4)}} - W^2\left(\frac{n_B{}^{(2)}}{n^{(2)}}\right)^2\right]$$

$$= \frac{1}{4}\left[\frac{(48)(5)(4)}{(9)(8)} + \frac{(272 - 2\cdot48)(5)(4)(3)}{(9)(8)(7)} + \frac{(24^2 + 48 - 272)(5)(4)(3)(2)}{(9)(8)(7)(6)}\right.$$

$$\left. -24^2\left(\frac{(5)(4)}{(9)(8)}\right)^2\right] = 0.9759.$$

Similarly, μ_2 for WW is 0.8909. The variance for BW is a different equation given in Appendix A, but involving the same terms, and yields 1.5430. To calculate the standard normal deviate, use the quantities given above in the equation:

$$S.N.D. = (\text{observed} - \mu_1)/\mu_2^{\frac{1}{2}}.$$

For this example, the S.N.D.'s are:

$$BB = -3.4156$$
$$WW = -2.2450$$
$$BW = 3.4564$$

The results are highly significant. The total number of BB and WW joins is far below expected, and if we interpret excess in the total number of like joins as a synonym for positive spatial autocorrelation, this example is significantly negatively autocorrelated, as we would have anticipated for a checkerboard. This is also borne out by the higher than expected BW joins. Note that these significance tests are probably not correct because asymptotic normality cannot be assumed for only nine localities. This small example has been shown only to facilitate learning the computational algorithm.

If we had connected the localities by bishop's moves instead of by rook's moves, we would have obtained strong positive autocorrelation because all localities would have had like-coloured ones as their neighbours. By the queen's moves criterion, autocorrelation should approach 0 in a large checkerboard because each locality will have an equal number of neighbours of both colours. Since graphs based on different criteria imply different hypotheses, the results of autocorrelation analyses of the same data set necessarily will differ with the connection criterion employed.

The significance tests in this instance should probably be carried out by a randomization test by computer since $n < 10$. However, for illustrative purposes we proceed to carry out the test using the correction for small sample sizes given in Appendix B. All symbols for moments given below relate to either BB or BW, depending on the join count being computed.

For the BB joins we first evaluate degrees of freedom

$$\nu = 2(\mu_1')^2/\mu_2 = 2(3\tfrac{1}{3})^2/(0.9759) \approx 23.$$

Since the χ^2 distribution is used as an approximation only, these degrees of freedom have no intrinsic meaning but might be best referred to as

"pseudo-degrees of freedom". For a two-tailed test at the 5% significance level, α in each tail equals 0.025. Since the join count for BB $< \mu'_1$ we are interested in the lower tail and require $\chi^2_{0.975[23]} = 11.688$. Then the 5% critical count for BB joins is

$$BB_{0.05} = \mu_2 \chi^2_{0.975\,[23]} /2\mu'_1 = (0.9759)11.688/2(3\tfrac{1}{3}) = 1.71.$$

Since our observed join count for BB is 0, this is significantly below expectations.

Although our pseudo-degrees of freedom are not large ($\nu = 23$), we evaluate below the formula based on the normal approximation to the χ^2 distribution to illustrate its use.

$$BB_\alpha = \frac{\mu_2 \left[t_{\alpha[\infty]} + \left(\dfrac{4(\mu'_1)^2 - \mu_2}{\mu_2} \right)^{\frac{1}{2}} \right]^2}{4\mu'_1}$$

$$BB_{0.05} = \frac{(0.0759)\left[-1.96 + \left(\dfrac{4(3\tfrac{1}{3})^2 - 0.9759}{0.9759} \right)^{\frac{1}{2}} \right]^2}{4(3\tfrac{1}{3})}$$

$$= 1.63.$$

These corrected critical values compare with the critical value that could have been obtained without the small sample correction, i.e.

$$BB_\alpha = t_{\alpha[\infty]}\, \mu_2^{\frac{1}{2}} + \mu_1 ,$$
$$BB_{0.05} = -1.96(0.9759)^{\frac{1}{2}} + 3\tfrac{1}{3} = 1.40.$$

Note that the uncorrected critical value is more conservative than the corrected one. For like joins at small sample sizes, the uncorrected critical point is too far from expectation in the lower tail, and too close to it in the upper tail of the distribution of outcomes.

For the BW joins the test is in the upper tail since the observed number of joins, 12, is greater than $\mu'_1 = 6\tfrac{2}{3}$. The formula given in Appendix B for the upper tail gives $K = 1 - \sqrt{10\alpha} = 1 - \sqrt{10(0.025)} = 0.5$. Therefore

$$BW_\alpha = t_{\alpha[\infty]}\, \mu_2^{\frac{1}{2}} + \mu'_1 - K_\alpha ,$$
$$BW_{0.05} = 1.96(1.5430)^{\frac{1}{2}} + 6\tfrac{2}{3} - 0.5 = 8.60.$$

Clearly, the observed number of unlike joins, 12, is significantly greater than the expected value. The critical value obtained without this correction is

$$BW_\alpha = t_{\alpha[\infty]}\, \mu_2^{\frac{1}{2}} + \mu'_1 ,$$
$$BW_{0.05} = 1.96(1.5430)^{\frac{1}{2}} + 6\tfrac{2}{3} = 9.10.$$

In this instance the correction is more conservative.

Computations for cases of nominal data with more than two types can be carried out similarly. However, there is an additional quantity to be computed—the total number of joins between localities of different types. It is

the sum of the number of rs joins where r varies from 1 to $k - 1$, and s varies from $r + 1$ to k.

Interval data

The autocorrelation coefficients for interval (and ordinal) data are Moran's statistic I and Geary's coefficient c. Although in the computations that we carried out we routinely computed c as well as I under normal as well as randomization assumptions, we have limited our account in this paper as well as the companion paper (Sokal & Oden, 1978) to I, Moran's coefficient tested for significance under the randomization assumption. While empirically results computed by both coefficients are similar, they are not identical. Both are analogous to correlation coefficients in that the numerator is an estimator of the covariance and the denominator an estimator of the variance of the observations. However, the numerator of I (see Appendix A) is a product-moment term analogous to the conventional correlation coefficient, while that of c (see Cliff & Ord, 1973: 12) is a squared term as in a distance coefficient (Geary's coefficient is thus structurally related to the well known Durbin-Watson statistic; see Kendall, 1973). Thus I will be strongly affected by marked, joint departures of two neighbouring values from their mean, while c will depend on the absolute differences between neighbouring variates. Moran's coefficient I varies from -1 to $+1$, the expected value approaching zero for large sample size in the absence of autocorrelation (as in the conventional product-moment correlation coefficient). Geary's c ranges from 0 (maximal positive autocorrelation) to a positive value for high negative autocorrelation. Its expectation in the absence of autocorrelation is 1.

The standard errors of either statistic can be calculated under two assumptions. The n values (for n localities) can be considered a sample from an infinite, normally distributed population for that variable. Alternatively, no assumption need be made about the distribution, in which case the autocorrelation is tested against the values the autocorrelation coefficient could take given all possible permutations of the observed variates over the localities. This latter randomization assumption is more reasonable for most systematic applications and it is the only one featured in this exposition. In our experience standard errors do not differ greatly for these two assumptions.

Next we turn to a sample computation for spatial autocorrelation in interval data.

Example 2. Computation of spatial autocorrelation for a variable in interval scale, irregularly-spaced localities, binary weights.

A continuous variable, forewing length of the aphid *Pemphigus populitrans-versus* for the year 1961, was mapped onto eight localities in Douglas and Leavenworth Counties, Kansas (Sokal, Heryford & Kishpaugh, 1971). The localities are indicated in Fig. 3.

Note that there is no necessary association between interval measure and irregular spacing. Nominal data can be irregularly spaced and interval data can be spaced regularly, and examples showing these combinations are featured in the companion paper (Sokal & Oden, 1978).

To compute spatial autocorrelation we need to indicate connections among the localities. For this example, and for several others to be featured below, we
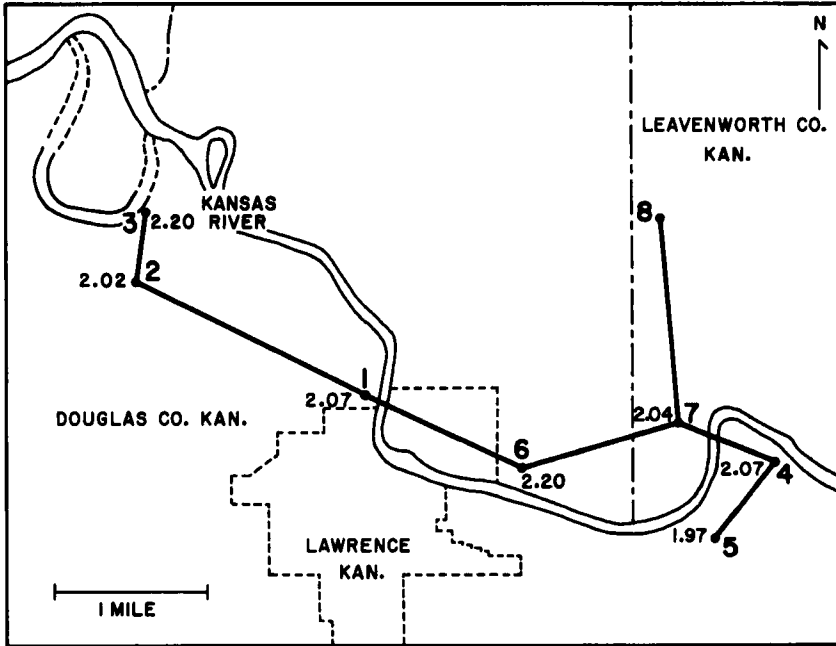
Figure 3. Map to illustrate geographical connections among the eight localities of Example 2. The map gives locations for eight collection sites in Kansas. The collection sites are identified by code numbers; means (in mm) for the variable forewing length are shown next to the locality code numbers. The localities are connected by a Gabriel network.

Table 2. Locality means and adjacency matrix for Example 2

Locality number	Mean forewing length (mm) x_i	z_i	Locality number:	1	2	3	4	5	6	7	8
1	2.07	0.0025		X	1	0	0	0	1	0	0
2	2.02	−0.0475		1	X	1	0	0	0	0	0
3	2.20	0.1325		0	1	X	0	0	0	0	0
4	2.07	0.0025		0	0	0	X	1	0	1	0
5	1.97	−0.0975		0	0	0	1	X	0	0	0
6	2.20	−0.1325		1	0	0	0	0	X	1	0
7	2.04	−0.0275		0	0	0	1	0	1	X	1
8	1.97	−0.0975		0	0	0	0	0	0	1	X

have chosen to employ a Gabriel-connected graph (Gabriel & Sokal, 1969) to indicate the pattern of geographical interactions. In such a graph any two localities, A and B are considered contiguous (connected) if no other locality lies on or within the circle whose diameter is the line AB. Thus A and B are connected when $d^2_{AB} < d^2_{AC} + d^2_{BC}$, where d^2_{AB} is the squared geographic distance between localities A and B, and C is any third locality. The rationale for using this criterion in geographic variation analysis is discussed by Gabriel & Sokal (1969). The resulting graph, often called a Gabriel-connected graph or Gabriel network, is planar. The Gabriel network for the localities of the aphid study is indicated in Fig. 3. The Gabriel criterion is just one of various possible criteria for connecting geographical localities.

The means and the adjacency matrix for the Gabriel-graph of Fig. 3, are given in Table 2. Since no further weighting was applied to the edges, the adjacency matrix is a binary weight matrix. The Moran coefficient I is given in Appendix A as

$$I = n\sum_{ij}w_{ij}z_iz_j/W \sum_{i=1}^{n} z_i^2 .$$

To compute this expression we break it down into parts. For the numerator we need all products of pairs of localities. Thus

Localities		Weight	z_iz_j	$w_{ij}z_iz_j$
i	j	w_{ij}		
1	2	1	(0.0025)(−0.0475)	−0.00011875
1	3	0	(0.0025) (0.1325)	0
1	4	0	(0.0025) (0.0025)	0
1	5	0	(0.0025)(−0.0975)	0
1	6	1	(0.0025) (0.1325)	0.00033125
1	7	0	(0.0025)(−0.0275)	0
1	8	0	(0.0025)(−0.0975)	0
2	1	1	(−0.0475) (0.0025)	−0.00011875
2	3	1	(−0.0475) (0.1325)	−0.00629375
⋮	⋮	⋮	⋮	⋮
8	6	0	(−0.0975) (0.1325)	0
8	7	1	(−0.0975)(−0.0275)	0.00268125

$$\sum_{ij}w_{ij}z_iz_j = -0.0147125$$

Since most weights are zero very few multiplications are actually carried out. The weighted sum of the products $\sum_{ij}w_{ij}z_iz_j = -0.0147125$.

The quantity W, the sum of the matrix weights, in this example with binary symmetrical weights, equals the number of ones in the adjacency matrix, i.e., 14, or twice the number of edges connecting the graph.

The sum of squares of the deviates

$$\sum_{i=1}^{n} z_i^2 = 0.057150.$$

Therefore $I = 8(-0.0147125)/14(0.057150) = -0.14711$.

The expected value of I, $\mu_1' = -\dfrac{1}{(8-1)} = -0.14286$.

For the computation of the variance, three further terms are needed to evaluate the formula for μ_2 given in Appendix A. The first of these is $S_1 = \frac{1}{2}\sum_{ij}(w_{ij} + w_{ji})^2$. In this case with binary symmetric weights every edge between a pair of localities i and j contributes $(1 + 1)^2 = 4$ to the sum and S_1 is simply $4 \times$ (number of edges in the graph) $= 28$. The second term is $S_2 = \sum_{i=1}^{n} (w_i. + w_{.i})^2$, where $w_i.$ and $w_{.i}$ are the sums of the ith row and ith column of the weight matrix, respectively. For this expression the marginal totals of the weight matrix should be computed. In this simple example row and column sums are the same because the matrix is symmetric

[thus $S_2 = \sum_{ij}(2w_{i\cdot})^2 = 4\sum_{ij}w_{i\cdot}^2$] and the $w_{i\cdot}$'s can be evaluated in one's head. Thus

$$S_2 = 4(2^2 + 2^2 + 1^2 + 2^2 + 1^2 + 2^2 + 3^2 + 1^2) = 4 \times 28 = 112.$$

The third term is

$$b_2 = n\sum_{i=1}^{n} z^4 / \left(\sum_{i=1}^{n} z^2\right)^2 = 8(0.0008028441)/(0.05715)^2 = 1.966476303.$$

Therefore

$$\mu_2 = \frac{n[(n^2 - 3n + 3)S_1 - nS_2 + 3W^2] - b_2[(n^2 - n)S_1 - 2nS_2 + 6W^2]}{(n-1)^{(3)}W^2}$$

$$- \frac{1}{(n-1)^2}$$

$$= \frac{8[(8^2 - 24 + 3)28 - 8(112) + 3(14)^2] - (1.966476)[(8^2 - 8)28 - 2(8)(112) + 6(14)^2]}{(8-1)(8-2)(8-3)(14)^2}$$

$$- \frac{1}{(8-1)^2}$$

$$= \frac{5295.914559}{41160} - \frac{1}{49} = 0.108258.$$

The standard error $\mu_2^{\frac{1}{2}} = 0.32903$.

Obviously, the difference from expectation is not significant: $(I - \mu_1)/\mu_2^{\frac{1}{2}}$ = $[-0.14711 - (-0.14286)]/0.32903 = 0.01292$. Note that asymptotic normality could not be assumed for only eight localities. This small example has been shown only to facilitate learning the computational algorithm.

In this example the proper significance test would also be by means of a randomization test because $n < 10$. However, for illustrative purposes we proceed to carry out the test using the correction for small sample sizes given in Appendix B.

Compute A, the number of edges in the connecting graph. As we saw earlier this is $\frac{1}{2}W$ in this example with binary weights. Therefore $A = 7$. We compute

$$R = 2(\sqrt{n} - 1)/\sqrt{n} = 2(1.82843)/(2.82843) = 1.2929,$$

$$Q = 2(\sqrt{n} - 1)(2\sqrt{n} - 1)/n = 2(1.82843)(4.65685)/8.0 = 2.1287,$$

$$A/n = 0.875.$$

Since it is not true that $R < A/n < Q$, we set $k_\alpha = 1$. This is equivalent to computing the critical value I_α without the correction.

$$I_\alpha = t_{\alpha[\infty]}\mu_2^{\frac{1}{2}} - k_\alpha(n-1)^{-1},$$

$$I_{0\cdot05} = -1.96(0.32903) - 1(8-1)^{-1} = -0.7877.$$

Since the observed I is only -0.1471, it clearly is not significant.

Connections and weights

It should be obvious from our analysis of the checkerboard in Example 1 that the manner in which we connect the localities will importantly affect the nature of the autocorrelation obtained. To the skilled investigator this will prove an asset rather than a liability; various types of connectivity may be defined to test for various types of structure. Let us take a hypothetical example, a mapping of the population change over a given time period in an insect pest population in a field. If the hypothesis is that the insects have moved from the centre of the field to the periphery, then decrease in population size at the centre should match increase in population size at the periphery. By connecting only those field plots in the centre of the field (as defined by the investigator) with those peripheral plots to which insects might have plausibly migrated, but not connecting central localities to other central localities or peripheral localities to other peripheral ones, an adjacency matrix is created on which the population change can be examined for autocorrelation. If the insects at the centre have indeed migrated to the periphery over the period of the study, negative autocorrelation of current population levels would result on the network defined above.

As mentioned earlier, weights need not be binary (1 for connections, 0 for lack of connection), but can be expressed as a function of the distance between localities or some other measure of ease of movement from one locality to the other. The formulas in Appendix A are given in terms of this general case where the elements w_{ij} are the weights of the weight matrix W.

Another way to use weights to test for structure in a spatial data set might be to weight all connections between points by the cosines of the angles the connections make with an hypothesized test direction. By computing auto-correlations for different test directions, one might discover underlying directional mechanisms in the data. This approach is analogous to the investigation of rook's moves, bishop's moves and queen's moves in the regular grid.

The instances of autocorrelation discussed so far have been restricted to pairs of points adjacent in the connection network, but we are not limited to such pairs. Autocorrelations can be computed for pairs of localities an arbitrary number of edge lengths or geographic distance units apart. One could then show how autocorrelation would change as a function of distance between points. A graph showing such a relationship is a correlogram. The construction of correlograms is discussed further in the next section.

AN EXAMPLE: HELIX ASPERSA

This example illustrates what may be the most frequent application of these techniques—interval data for irregularly-spaced points in an area. The variables studied here are gene frequencies rather than morphometric variables. When only two alleles are found at any one locus, obviously only one gene frequency need be analyzed. In the case of multiple alleles, one less than the number of alleles is studied. In the data presented below we consistently omitted the most infrequent allele. The data set is from a study describing microgeographic variation in allozyme frequencies in the European brown snail, *Helix aspersa.*
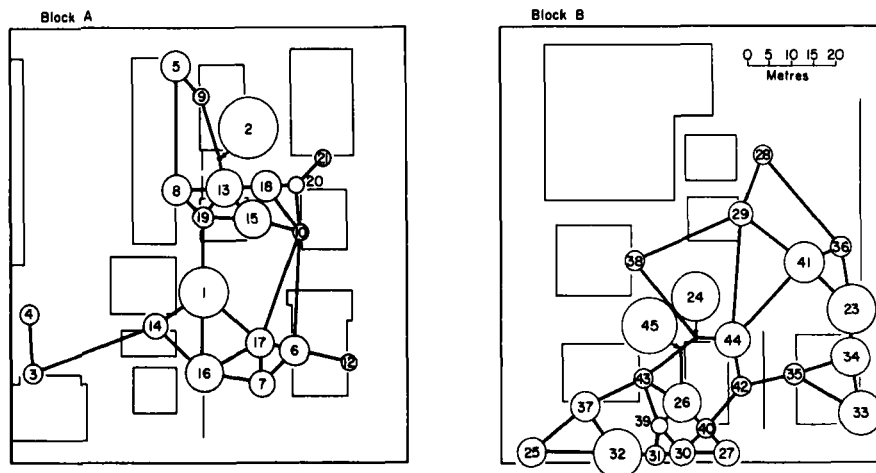
Figure 4. Location and size of colonies of *Helix aspersa* on adjacent city blocks in Bryan, Texas. These maps have been modified from Selander & Kaufman (1975). Twenty colonies are shown in Block A, 23 in Block B. Colonies are designated by numbers 1-21 in Block A, 23-45 in Block B. There is no colony numbered 11. The circle size is proportional to colony size. Lines connecting the colonies are a Gabriel-connected graph. In Block A an edge was added connecting localities 5 and 8 because of continuity of habitat along the building.

Selander & Kaufman (1975) report exhaustive collections of these snails from shrubbery in two adjacent city blocks in Bryan, Texas. The blocks were approximately 80 x 100 m and contained 20 and 23 snail colonies, respectively. The allozyme variation was analyzed for five polymorphic enzymes and allozyme frequencies computed for each population.

The colonies are mapped in Fig. 4 and were joined to form a Gabriel-connected graph, as shown in the figure. Actually the graph is not truly Gabriel-connected since the edge between colonies 5 and 8 has been added, although it violates the Gabriel conditions. This was done because continuity of habitat along the building suggested the possibility of gene flow. This reiterates the point that the methods described here are not predicated on a specific type of graph. Maps showing the geographic variation of the allozyme frequencies are featured in Selander & Kaufman (1975) and are not reproduced here.

Figure 5 displays plots of Moran's I as a function of interpoint distance. These plots are called correlograms (Kendall, 1973; Box & Jenkins, 1971). Since correlograms furnish information on the behaviour of the autocorrelation coefficients as interpoint distance increases, they are used in time series analysis, where interpoint distance indicates time intervals, to furnish suggestions as to the nature of the process generating the series. Thus if a variable X_t at time t is a function only of its value X_{t-1} at time $t-1$ plus a random contribution, its correlogram will be different from that of a variable that is a function of its previous value at still earlier time periods. Similar inferences may be made in spatial autocorrelations where the dependence of a variable is in two dimensions.

We have constructed correlograms by two different definitions of distance. In one of these (correlograms not shown) the distance between two points is the least number of edges sufficient to connect the points. Thus, in Fig. 4, colonies 5 and 19 were considered neighbours in distance class 2, and in no
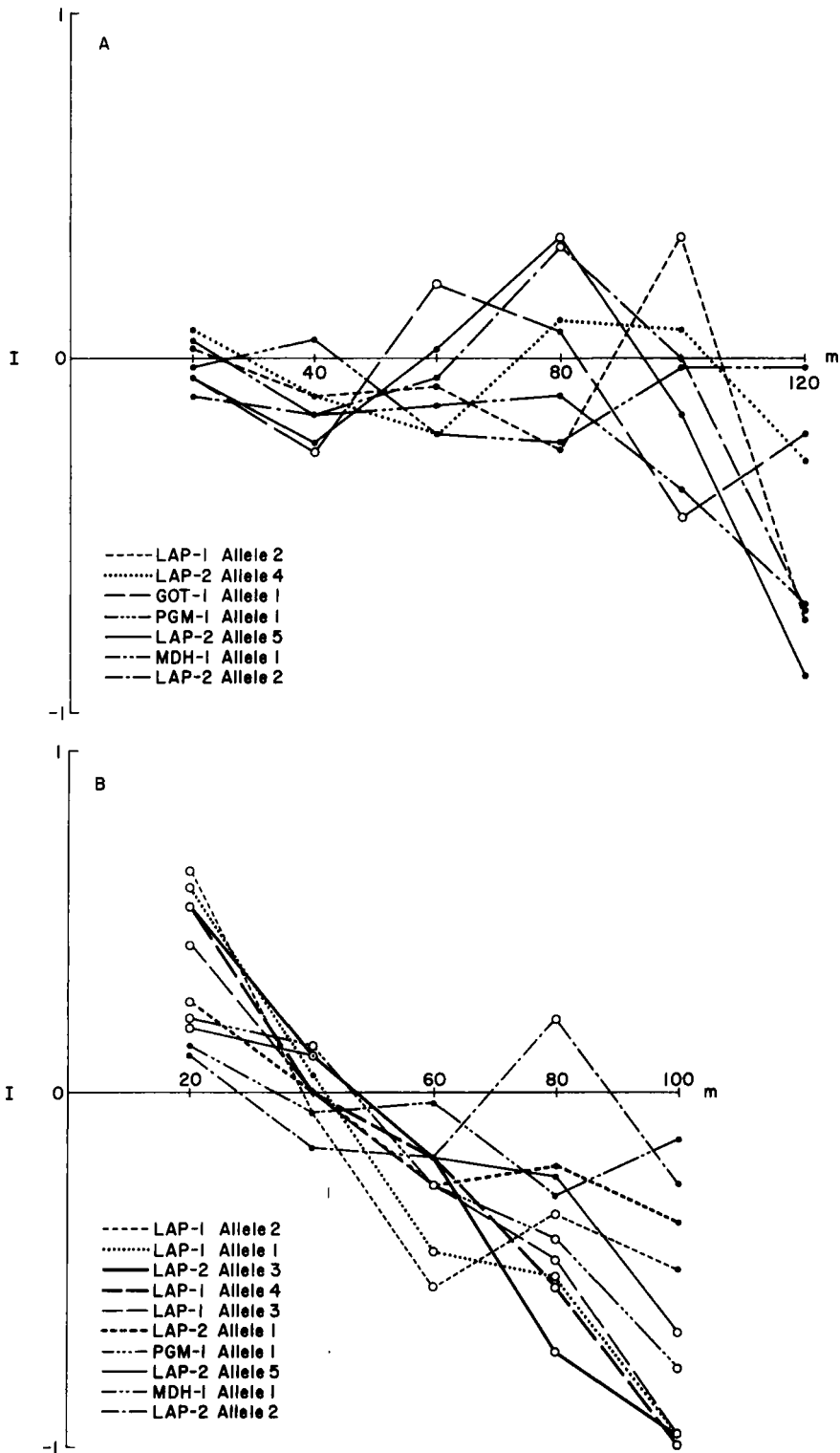
Figure 5. Distance-corrected correlograms for geographic variation patterns of allozyme frequencies of the snail *Helix aspersa* in two blocks in Bryan, Texas. A, Block A; B, Block B. Abscissa: shows distance in metres. Here, unlike Fig. 7, distances indicated are upper limits of the distance class rather than class marks. Ordinate: Moran's autocorrelation coefficient I. Significant autocorrelation coefficients ($P < 0.05$) are indicated by open circles. Based on data from Selander & Kaufman (1975).

other distance class, even though longer paths between 5 and 19 exist, as for example 5-9-2-13-19. The smallest number of edges sufficient to connect the most distant pair of points is called the diameter or solution time of the graph which in the case of Block A is 7. Thus a correlogram can be constructed for which the autocorrelation coefficient is calculated for each distance class i, $i = 1, \ldots, 7$ by considering as neighbours only those point pairs whose interpoint distance is i. In other words, a new weight matrix was constructed for each class i in which weights of 1 were assigned to point pairs which were neighbours at distance i whereas weights of 0 were assigned to point pairs which were not. If any point was not i edges distant from at least one other point, it was excluded from the weight matrix for that distance class. By contrast, the correlograms shown in Fig. 5 are what we have termed "distance-corrected". In such correlograms basically the same criteria are followed for determining which localities are neighbours, except that the actual lengths along graph edges are considered. Distance classes with intervals are erected, and, as before, for each distance class only those point pairs are considered neighbours whose interpoint distances (shortest connecting paths as measured along the graph edges) fall within the distance class. The shortest path distances computed for all pairs of colonies in each block of the *Helix* study were grouped into distance classes with an interval of 20 m. The class marks indicated in the figure are upper bounds (i.e., all distances up to 20 m, or those between 20 and 40 m), rather than the more conventional midpoints.

Distance-corrected correlograms are appropriate in those biological situations where it is desirable to retain the Gabriel criterion (see Gabriel & Sokal, 1969) and yet avoid the space distortion inherent in correlograms of the first type discussed above. A final point is that the two methods may yield different shortest distance paths. Thus in Fig. 4 the distance between localities 5 and 13 along the three edges 5-9, 9-2, and 2-13 is shorter by the distance-corrected method than along the two edges 5-8 and 8-13. It has been our experience with several data sets that correlograms based on binary connection matrices and those that are distance-corrected while differing in detail, show generally comparable profiles. This is not true when the geographical lengths of the edges are very variable.

The distance-corrected correlograms in Fig. 5 show marked differences between the two blocks and among some of the allozyme frequencies. Several loci in Block B show a consistent pattern of significant positive autocorrelations up to 20 m, and of significant negative autocorrelation in the 60 and 80 m classes. Not too much reliance should be placed on the last distance class(es) since the autocorrelation coefficient for these classes is generally based on very few point pairs. Altogether, out of 50 autocorrelation coefficients 28 are significant. By contrast in Block A only 6 out of 42 autocorrelation coefficients are significant. While autocorrelations in Block B tended to decrease with increasing distance for most loci, such a trend could not clearly be established for Block A. We may conclude that, on the whole, neighbouring populations in Block B are more similar to each other, and distant populations more different, than such populations are in Block A. Some allozyme frequencies in each block show no significant autocorrelation whatsoever.

Although both blocks were shown to be significantly heterogeneous for most loci by Selander & Kaufman (1975), the two blocks differ considerably in

patterns when tested by the method of Royaltey, Astrachan & Sokal (1975). Of seven patterns in Block A only one was significant, while seven out of ten of the allozyme patterns in Block B were found significant by the same test (Selander & Kaufman, 1975).

In Block A the heterogeneous means show little significant spatial structure, and what little there is, is not similar for various allozyme frequencies. Because there is not much pattern a stochastic interpretation is favoured, probably several immigration events leading to founder effects. A selectionist interpretation—i.e., different sites differ in various selective agents that affect the various allozyme frequencies—is improbable in view of the parallel variation for these same frequencies in Block B. That block looks very different. All significantly heterogeneous allozyme frequencies in this block have similar and significant patterns as shown by the Royaltey, Astrachan and Sokal test and by the correlograms (compare Fig. 5A and B). One might therefore suppose a common selective agent on a gradient. In view of the diversity of loci studied, this seems unlikely, and it would be more probable to assume a migratory event followed by diffusion which would give rise to a situation mimicking an adaptive cline. This would also explain the discrepancy between the correlogram patterns for Block A and Block B, since if these loci are subject to strict genetic selective control, one would have to postulate important differences in the microclimatic and microenvironmental substructuring of Blocks A and B for which Selander & Kaufman offer no basis. An alternative explanation for the differences between the blocks might be that there were substantial differences in time since colonization. If Block A had been colonized early, but Block B more recently, the pattern in A could be interpreted as selection due to diverse microhabitats, whereas that for B could be diffusion between two diverse immigrant stocks, before selection had patterned the samples to the microhabitats in B. However, we have no evidence for such a sequence of events in these adjacent city blocks.

STRUCTURAL IMPLICATIONS

Autocorrelation and surface structure

The meaning of strong positive or negative autocorrelation at the lowest distance class is fairly straightforward. Positive autocorrelation indicates that values at adjacent localities are similar, while negative spatial autocorrelation signifies marked differences between adjacent pairs. While the determination of these phenomena and of their origin is of obvious interest to biologists, systematists are concerned with the larger pattern of variation exhibited by the variable over the entire surface studied. Autocorrelations at the lowest distance class do not necessarily furnish the basis for inferences about the entire structure. An example will make this clear. In a regular gradient such as a cline (see Fig. 6A) there will be strong positive autocorrelation at short distances, since near neighbours resemble each other. But we could also expect high positive spatial autocorrelation for the shortest distance class if the study area is divided into two large contiguous patches each comprising numerous localities, with homogeneous high values assigned to the localities in patch 1, while homogeneous low values are assigned to those of patch 2 (Fig. 6B). In
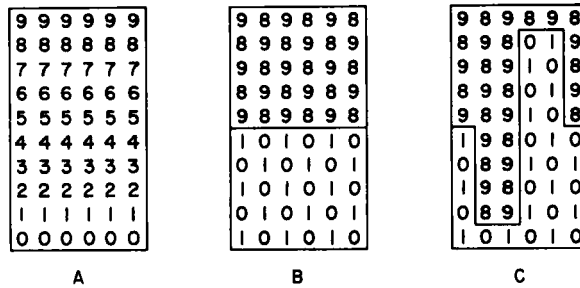
```
9 9 9 9 9 9      9 8 9 8 9 8      9 8 9 8 9 8
8 8 8 8 8 8      8 9 8 9 8 9      8 9 8 0 1 9
7 7 7 7 7 7      9 8 9 8 9 8      9 8 9 1 0 8
6 6 6 6 6 6      8 9 8 9 8 9      8 9 8 0 1 9
5 5 5 5 5 5      9 8 9 8 9 8      9 8 9 1 0 8
4 4 4 4 4 4      1 0 1 0 1 0      1 9 8 0 1 0
3 3 3 3 3 3      0 1 0 1 0 1      0 8 9 1 0 1
2 2 2 2 2 2      1 0 1 0 1 0      1 9 8 0 1 0
1 1 1 1 1 1      0 1 0 1 0 1      0 8 9 1 0 1
0 0 0 0 0 0      1 0 1 0 1 0      1 0 1 0 1 0

        A                B                C
```

Figure 6. Schematic diagram to represent the effect of shapes and homogeneity of patches. A. A cline. B. Two homogeneous patches with a relatively short boundary between them. C. Two homogeneous interdigitating patches. For the consequences of these shapes in terms of correlograms, see text.

such a case the number of short-distance joins or edges within one patch will considerably exceed the number of such joins between patches, the latter number being limited to the number of ways a locality in patch 1 can be connected to a locality in patch 2 using only distances in the lowest distance class. Only when the boundary between the patches is lengthened, as in a sinuous or interdigitating boundary (see Fig. 6C), will the number of interpatch joins with their negative contribution to the numerator of the autocorrelation coefficient be sufficiently large to cancel the positive contribution of the within-patch joins. The cline and the two homogeneous patches are generally acknowledged as radically different patterns that need to be distinguished. Therefore, we need a means of providing an overall summary of the entire distribution of the variable over the space studied. In this paper we have employed correlograms for such summaries. While correlograms do not exhaust the possibilities for useful summarization, they seem to be a fruitful initial approach.

Artificial patterns

For further insights into the nature of the inferences that can be made from correlogram profiles we resorted to the five artificial patterns generated by Royaltey, Astrachan & Sokal (1975). These patterns had been mapped on a Gabriel network for 53 localities in eastern North America shown in Fig. 7. The patterns were generated by deliberately assigning the ranks from 1 through 53 to the vertices of this graph in such a manner as to achieve the desired effect. The cline (shown in Royaltey *et al.*, 1975: fig. 2) was produced by introducing a gradual transition from high values in the north to low values in the south. The depression features a central core of low ranks in southern Indiana (not Ohio, as incorrectly stated by Royaltey *et al.*, 1975), gradually increasing toward the periphery of the area. In the double depression a low area in southern Indiana and Ohio is separated from another low area in southern Illinois and Missouri by three high ranked localities. The ranks increase in magnitude toward the periphery of the area. The intrusion is an area of high rank in the south-west with a sharp abutment bordering low rank localities, while in the crazy quilt pattern high ranks are surrounded by low ones, and vice versa, in such a way that a traveller traversing the edges from one portion of the

graph to the other would alternatingly encounter high and low ranks. None of the patterns described above is optimal. They were arrived at by pencil and paper, and in no case was more than half an hour spent on generating a pattern. More clear-cut patterns could have been produced by suitable computer algorithms. Yet the ones generated are probably as marked as any likely to be encountered in nature in most instances. Royaltey *et al.* (1975) developed a method for testing the significance of departures from randomness in geographic variation patterns, and were able to show that these artificial patterns differed highly significantly from random ones.

So as to be able to relate findings on these patterns to the other work described in this paper, we preferred to carry out the autocorrelation not on ranks but on an interval variable. To do so we ranked the means of an interval variable studied for these 53 localities (stem mother head width from the
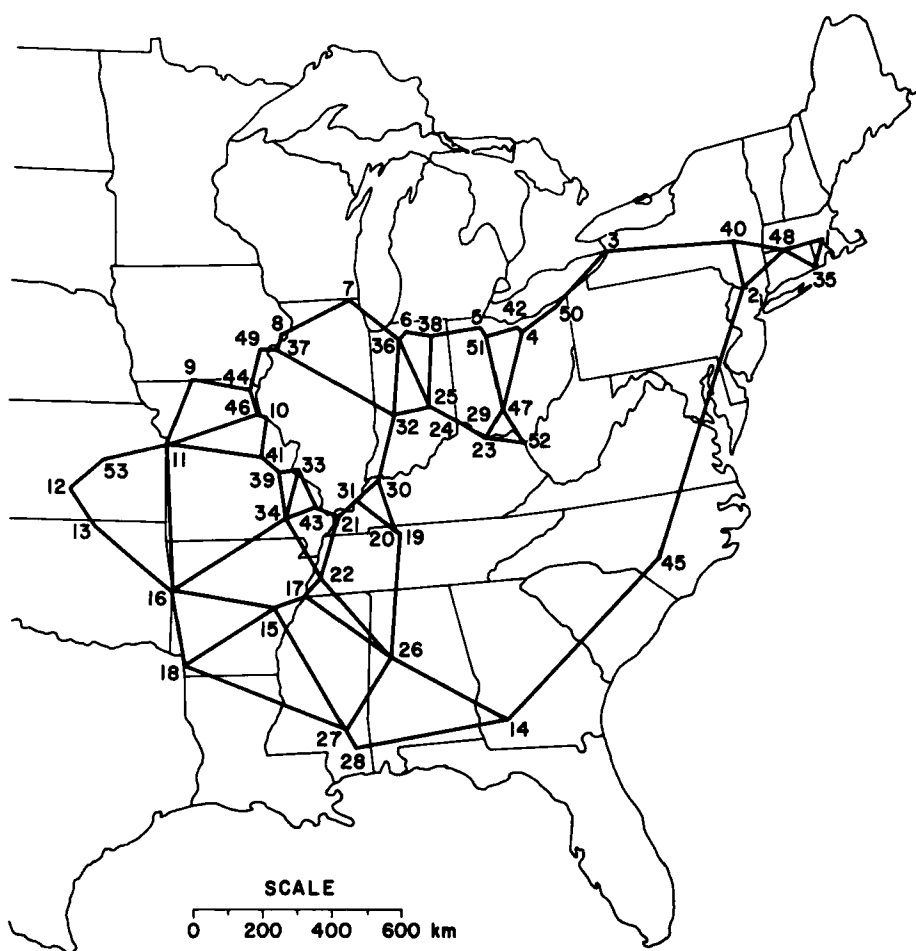
Figure 7. A Gabriel-connected graph of 53 localities in the eastern United States. The numbers at the vertices of the graph are code numbers for the 53 localities. This graph was originally generated in connection with a study of the aphid *Pemphigus populitransversus,* and for this reason the edge between localities 45 and 52 was omitted since gene flow across the Appalachian mountain range is impossible for these organisms.

revisited populations of the aphid *Pemphigus populitransversus*—Sokal & Riska, in prep.). The interval scale values (in mm) were then assigned to localities so that their ranks matched the ranks given those localities for the five artificial patterns. In this way a known variable was artificially reshuffled spatially to create a cline, a depression, a crazy quilt, and so on. The resulting correlograms of the patterns are shown in Fig. 8. The correlograms shown are distance-corrected, although during the generation of the artificial patterns the edges of the Gabriel network were considered to be of equal length, and geographic distance was thus not taken into consideration in forming the patterns. The correlograms based on equal edge lengths were also computed, yielding, not surprisingly, somewhat better definition. Nevertheless, the differences among these patterns are sufficiently marked for the distance-corrected correlograms to yield characteristic profiles. Again the upper bounds of distance classes are indicated in the figure.

The correlogram for the cline shows an almost regular monotonic decline from significant positive autocorrelation at 200 to 800 km to significant negative autocorrelation from 1400 to 2800 km. The interpretation here is quite simple, because of the particular configuration of the study area (see Fig. 7). For distances of 700 km and more it is difficult to remain in a region
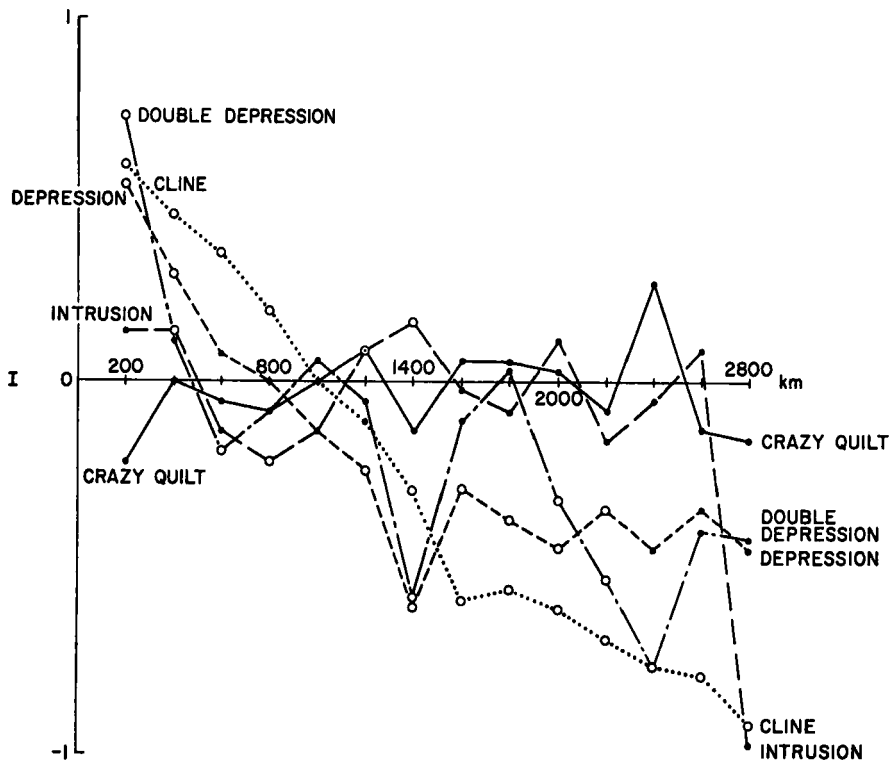
Figure 8. Correlograms based on the map and Gabriel network of Fig. 7 for five artificially-generated variables intended to simulate distinct patterns of geographic variation. (See text for details on the patterns.) Abscissa: distance in kilometres. Ordinate: Moran's autocorrelation coefficient *I*. Significant autocorrelation coefficients (*P* ≤ 0.05) are indicated by open circles.

where the variables are more or less uniform. Areas of equal value for the variable run roughly east to west, while the cline is north-south. Many of the longer distances are diagonal or north-south. In fact, the highest distances (over 2000 km) are necessarily diagonal given the shape of this network. They are, therefore, inevitably negatively correlated. It is interesting to compare the correlogram for the cline based on equal edge lengths (not shown) with the results just discussed. In this correlogram negative autocorrelations at distance class 9 give way to nonsignificant autocorrelations near expectation for the highest distance classes. This is so because of the quite unequal length of the Gabriel edges in different parts of the graph. One can traverse the area from north-east to south-west using less edges than in an east-west transect straight across the northern boundary (see Fig. 7). Thus some like and unlike localities are the same number of edges apart and cancel each other at the higher distances.

The depression, which is nothing but a circular cline, shows a correlogram quite similar to that of the regular cline, with the lowest negative autocorrelation at 1400 km, the approximate radius of the depression. Beyond 1600 km an appreciable number of the comparisons will be between localities on opposite sides of the periphery of the area, which in the depression pattern are quite similar to each other and cancel other comparisons between divergent locality pairs, one from the centre, the other from the extreme periphery. In the double depression the pattern is on the whole quite similar to that of the depression. The intrusion pattern shows positive autocorrelation at distances up to 400 km because of the homogeneity within the compact south-western region. There is a significant negative autocorrelation at 600 km. Travel over that distance will in most cases lead from one region to another one differing greatly in mean.

The crazy quilt pattern surprisingly yields no significant autocorrelations at all. Although it starts out with a negative value of I, it does not show the expected alternation between high and low values as distance increases, because of weak correspondence between number of edges and geographic distance for different portions of the graph. Had the crazy quilt generating mechanism been distance-dependent in the first place, i.e. had the means been unlike each other every 200 km, then a tendency for regular oscillations between positive and negative autocorrelations would have been evident in the correlogram. The correlogram based on equal edge lengths (not shown) yields a highly significant negative autocorrelation between localities one edge length apart, while yielding nonsignificant autocorrelations for all greater edge length. This latter phenomenon can be explained by the irregularity of the graph. At distances greater than one edge length the relation between two means cannot be predicted on the average. In fact, in an irregular grid such as that of Fig. 7, it is generally impossible to alternate high and low values perfectly.

For obvious reasons, geographic variation patterns of actual data are less readily interpreted than the artificial patterns just discussed. We shall consider two examples from the snail *Helix aspersa* illustrated earlier. Distance-corrected correlograms for the allozyme frequencies of Lap-1 allele 1 and Mdh-1 in Block B are shown in Fig. 5B. While Lap-1 allele 1 shows positive autocorrelation at 20 m and significant negative autocorrelation at 60, 80 and 100 m, no significant autocorrelation is exhibited by Mdh-1. How are these differences in
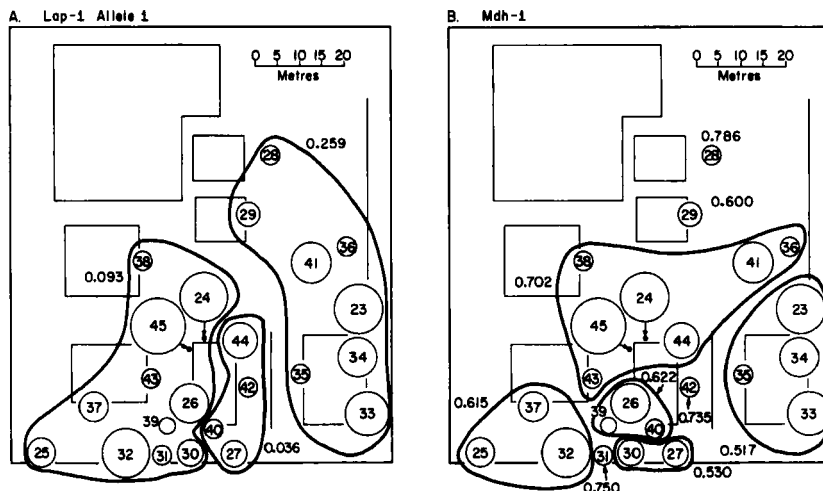
Figure 9. Maps of two allozyme frequencies, Lap-1 allele 1, and Mdh-1, in Block B. Statistically homogeneous (maximally acceptable) and geographically-connected localities defined by the method of Sokal & Riska (in prep.) are indicated by heavy lines surrounding the colonies. Single colonies not surrounded by heavy lines would reject as statistically heterogeneous if any neighbouring locality were added to them or would cause heterogeneity by being added to any existing cluster of localities. The maps are the same as shown in Fig. 4. Numbers next to clusters or single localities indicate mean or single allozyme frequencies, respectively.

correlogram profiles reflected in terms of differences of the geographic variation patterns? The characters have been mapped in Fig. 9 by applying a technique developed by Sokal & Riska (in prep.). The variables mapped onto the localities are subjected to single-linkage clustering (Sneath & Sokal, 1973) based on distances between the variables (in this case the gene frequencies) with the restriction that members of a cluster must be geographically connected—in our case Gabriel-connected. As the clustering proceeds agglomeratively, the partition imposed on the set of localities is tested for statistical homogeneity within clusters and for heterogeneity among clusters by a simultaneous test procedure—in this case by the G-test for homogeneity of proportions (Gabriel & Sokal, 1969; Sokal & Rohlf, 1969). In the solution illustrated in Fig. 9 each subset of means is maximally acceptable—addition of the variable for any other contiguous locality would result in a significantly heterogeneous subset. The method thus results in a unique partition of the set of localities into geographically connected and statistically acceptable (homogeneous) parts.

For Lap-1 allele 1 the allozyme frequencies are grouped into three "patches" that indicate a high in the east and north and a decline to the south-west. The geographic variation pattern of Mdh-1 looks quite dissimilar—four patches with two or more colonies and four singleton colonies without any clear restriction for high or low frequencies to any one geographic area. Two pairs of clusters (25, 32, 37) with (26, 39, 40), and (27, 30) with (23, 33, 34, 35) do not differ appreciably in mean frequency of Mdh-1. They apparently became established separately through the clustering of pairs of colonies similar in gene frequency, but became too heterogeneous to permit their coalescing, since each is a maximally acceptable subset. Furthermore, there are a number of unclustered

single colonies which, were they to join neighbouring means or parts, would cause these to be heterogeneous. Thus, including the singletons, there are actually nine parts in the Mdh-1 map in Fig. 9.

The reasons for the differences in correlograms are now apparent. Because of the singletons, the patches in Mdh-1 are somewhat smaller. Far fewer of the gene frequencies up to 20 m apart are *within* patches, hence similar, for Mdh-1 than for Lap-1 allele 1. Also, because the heterogeneous singletons in Mdh-1 are interspersed among the patches, an appreciable proportion of colony pairs up to 20 m apart show marked differences in gene frequencies, neutralizing the contribution toward positive autocorrelation from pairs of colonies within homogeneous patches. However, for Lap-1 allele 1 far fewer neighbouring colonies (up to 20 m apart) are dissimilar. While the negative autocorrelation of Lap-1 allele 1 at distances of 60 through 100 m arises because pairs of colonies at these distances are predominantly members of patches differing considerably in mean gene frequency, this is not so for Mdh-1 which has patches with similar means. At any distance class in Mdh-1, apparently, the contribution to positive autocorrelation by pairs of like means is counteracted by unlike means in the same distance class.

We have carefully examined maps partitioned in the manner of Sokal & Riska (in preparation) for all the *Helix* data reported by Selander & Kaufman (1975), for the artificial patterns, and for other studies, and we were able to explain the correspondence between the maps and their correlograms in the manner just indicated. The correlogram profiles will be affected by the nature and complexity of the response surface of the variable over the geographic plane. In the case of a sloping surface, the result depends on whether the high or low point is located at the margin of the area or somewhere within its confines. The former case leads to a monotonically declining correlogram, the latter case to some recursion at the higher distances with the details depending on the configuration of the surface. When there are areas of homogeneity (patches), the correlogram will reflect the distribution of patch sizes and the relative levels of the means pertaining to each patch. As we have seen, patches with considerable internal heterogeneity will not give clearly defined correlograms, whereas homogeneous patches will do so. The spatial arrangement of patches will also determine the correlogram. If high and low patches alternate, the results will be quite different from an arrangement in which the most divergent patches are farthest apart. Attention should be drawn to the swamping of relations in a large study. Clear cut differences, for example, between two divergent and adjacent homogeneous patches may be diluted by more ambiguous relations of other patches elsewhere on a map.

Other approaches

An alternative to thinking of character surfaces in terms of patches is the following:

Imagine throwing a stick of a specific length over and over again at a certain study area, while noting the values of a variate at the points at which the ends of the stick land. Affirming spatial autocorrelation is tantamount to saying that the values at the ends of the stick are usually like each other, or that they are usually different. Suppose a study area is composed of squares arranged in a

checkerboard fashion, with the colours of the squares alternating from black to white. Brief reflection will reveal that the stick length likely to yield the most negative autocorrelation is that which is roughly the length of one square, while a stick two squares long will yield positive autocorrelation, and so on. This shows that in this case the expected correlogram alternates from positive to negative, with wavelength equal to average patch diameter. This approach is related to Pielou's (1974) concept of "grain".

Another approach to investigating the complex nature of the surface is that of Royaltey *et al.* (1975), examining the distribution of edge lengths in the Gabriel graph. Although we are unable to show it analytically we believe the two approaches to be closely related. A rejection of the null hypothesis of random pattern by the method of Royaltey, Astrachan and Sokal is generally associated with significant autocorrelation at some distance class. Of 18 patterns in the geographic variation study of *Helix aspersa* analysed by Selander & Kaufman (1975) and in this paper, only 4 lacked significant patterns by either criterion, 7 were significant by the Royaltey-Astrachan-Sokal test, as well as in their correlograms, 1 was significant by the former and not by the latter, while 6 were significant by autocorrelation only. Corresponding figures for a study of 36 morphometric variables in the aphid *Pemphigus populitrans- versus* (Sokal & Riska, unpubl.) are 3, 21, 1, and 11. Thus autocorrelation seems to be a finer tool for picking up patterns, whereas the Royaltey- Astrachan-Sokal test seems to be a more conservative test for departure from randomness.

More refined methods for summarizing the pattern of the geographic variation on a surface may be developed in time. The techniques of Matheron (1971) may well be one fruitful approach. In the meantime, correlograms do furnish an informative and heuristic summation of the pattern. As this paper went to press, we learned of an application of spatial autocorrelation analysis to marine ecological data by Jumars, Thistle & Jones (1977).

SUMMARY

Spatial autocorrelation analysis is a technique that can be applied to nominal, ordinal, or interval variables mapped onto a geographic area. It tests whether the observed value of a variable at one locality is independent of values of the variable at neighbouring localities. The methods of spatial autocorrela- tion analysis recently summarized by Cliff & Ord (1973) are explained in some detail. The computation of two autocorrelation coefficients, join counts for nominal data, and Moran's coefficient I for interval data, is illustrated, as is the computation of significance tests. The method is extended to include the computations of correlograms for spatial autocorrelation. These diagrams show the autocorrelation coefficient as a function of distance between pairs of localities being considered. They summarize the patterns of geographic variation as exhibited by the response surface of any given variable.

The above methods of autocorrelation analysis are applied to a study of microgeographic variation of allozyme frequencies in the snail *Helix aspersa.* Differences in the distributions of the gene frequencies are found in two city

blocks where these snails were collected, and an attempt is made to interpret the nature and origin of these differences.

The inferences that can be drawn from autocorrelation profiles about the structure of response surfaces are discussed in some detail and illustrated by analyses of correlograms of artificially generated variation patterns. Size, homogeneity, and arrangement of patches are reflected in correlogram profiles. Computational formulae, expected values and standard errors are furnished in two appendices.

8

REFERENCES

BOX, G. E. & JENKINS, G. M., 1970. *Time and Series Analysis.* San Francisco: Holden-Day.

BUSACKER, R. G. & SAATY, T. L., 1965. *Finite Graphs and Networks:* 294 pp. New York: McGraw-Hill.

CLIFF, A. D. & ORD, J. K., 1971. Evaluating the percentage points of a spatial autocorrelation coefficient. *Geographical Analysis, 3:* 51-62.

CLIFF, A. D. & ORD, J. K., 1973. *Spatial Autocorrelation:* 175 pp. London: Pion.

GABRIEL, K. R. & SOKAL, R. R., 1969. A new statistical approach to geographic variation analysis. *Systematic Zoology, 18:* 259-270.

JUMARS, P. A., THISTLE, D. & JONES, M. L., 1977. Detecting two-dimensional spatial structure in biological data. *Oecologia, 28:* 109-123.

KENDALL, M. G., 1973. *Time-Series:* 197 pp. London: Griffin.

MATERN, P., 1960. Spatial variations; stochastic models and their application to some problems in forest surveys and other sampling investigations. *Matter Meddelanden fran Statens Skogsforskingsinstitut, 49:* 1-144.

MATHERON, G., 1971. The theory of regionalized variables and its applications. *Cahiers Centre Morphologie Mathématique Fontainebleau, No. 5:* 211 pp.

PIELOU, E. C., 1974. *Population and Community Ecology: Principles and Methods.* New York: Gordon & Breach.

ROYALTEY, H. H., ASTRACHAN, E. & SOKAL, R. R., 1975. Tests for patterns in geographic variation. *Geographical Analysis, 7:* 369-395.

SELANDER, R. K. & KAUFMAN, D. W., 1975. Genetic structure of populations of the brown snail (*Helix aspersa*). I. Microgeographic variation. *Evolution, 29:* 385-401.

SNEATH, P. H. A. & SOKAL, R. R., 1973. *Numerical Taxonomy:* 573 pp. San Francisco: W. H. Freeman.

SOKAL, R. R., HERYFORD, N. N. & KISHPAUGH, J. R. L., 1971. Changes in microgeographic variation patterns of *Pemphigus populitransversus* over a six-year span. *Evolution, 25:* 584-590.

SOKAL, R. R. & ODEN, N. L., 1978. Spatial autocorrelation in biology. 2. Some biological implications and four applications of evolutionary and ecological interest. *Biological Journal of the Linnean Society, 10:* 229-249.

SOKAL, R. R. & ROHLF, F. J., 1969. *Biometry:* 776 pp. San Francisco: W. H. Freeman.

TOBLER, W. R., 1975. Linear operation applied to areal data. In J. C. Davis & M. J. McCullogh (Eds), *Display and Analysis of Spatial Data:* 378 pp. London: John Wiley.

WHITTLE, P., 1954. On stationary processes in the plane. *Biometrika, 41:* 434-449.

APPENDIX A

Computational formulas, expected values and variances of autocorrelation statistics

Nominal data (for $k \geqslant 2$ classes or types)

    Join counts

        For localities of the same type:

$$\text{Number of } rr \text{ joins} = \tfrac{1}{2}\sum_{ij}w_{ij}(rr)_{ij}$$

        For localities of different types:

$$\text{Number of } rs \text{ joins} = \tfrac{1}{2}\sum_{ij}w_{ij}(rs)_{ij}$$

    Total number of joins between localities of different types:

        Sum of joins between all possible pairs rs where

$$r = 1 \ldots k-1, \text{ and } s = r+1, \ldots k \text{ for } k \text{ types} =$$

$$\tfrac{1}{2}\sum_{r=1}^{k-1}\sum_{s=r+1}^{k}\sum_{ij}w_{ij}(rs)_{ij}$$

Expected values and variances: (under assumption of sampling without replacement)

Joins between localities of the same type:

$$\mu_1' = \frac{W n_r^{(2)}}{2 n^{(2)}},$$

$$\mu_2 = \frac{1}{4}\left[\frac{S_1 n_r^{(2)}}{n^{(2)}} + \frac{(S_2 - 2S_1) n_r^{(3)}}{n^{(3)}} + \frac{(W^2 + S_1 - S_2) n_r^{(4)}}{n^{(4)}} - W^2 \left(\frac{n_r^{(2)}}{n^{(2)}}\right)^2\right].$$

Joins between localities of two different types:

$$\mu_1' = \frac{W n_r n_s}{n^{(2)}},$$

$$\mu_2 = \frac{1}{4}\left[\frac{2S_1 n_r n_s}{n^{(2)}} + \frac{(S_2 - 2S_1) n_r n_s (n_r + n_s - 2)}{n^{(3)}}\right.$$
$$\left. + \frac{4(W^2 + S_1 - S_2) n_r^{(2)} n_s^{(2)}}{n^{(4)}} - 4W^2 \left(\frac{n_r n_s}{n^{(2)}}\right)^2\right].$$

Total number of joins between localities of different types:

$$\mu_1' = W \sum_{r=1}^{k-1} \sum_{s=r+1}^{k} \frac{n_r n_s}{n^{(2)}},$$

$$\mu_2 = \frac{1}{4}\left\{\left[\frac{S_2}{n^{(2)}} - \frac{4(W^2 + S_1 - S_2)(n-1)}{n^{(4)}}\right]\sum_{r=1}^{k-1}\sum_{s=r+1}^{k} n_r n_s\right.$$

$$+ \left[\frac{2S_1 - 5S_2}{n^{(3)}} + \frac{12(W^2 + S_1 - S_2)}{n^{(4)}} + \frac{8W^2}{n^{(3)}(n-1)}\right]\sum_{r=1}^{k-2}\sum_{s=r+1}^{k-1}\sum_{t=s+1}^{k} n_r n_s n_t$$

$$+ 4\left[\frac{(S_1 - S_2)}{n^{(4)}} + \frac{2W^2(2n-3)}{n^{(2)} n^{(4)}}\right]\sum_{r=1}^{k-1}\sum_{s=r+1}^{k} n_r^2 n_s^2$$

$$- 8\left[\frac{(S_1 - S_2)}{n^{(4)}} + \frac{2W^2(2n-3)}{n^{(2)} n^{(4)}}\right]\sum_{r=1}^{k-3}\sum_{s=r+1}^{k-2}\sum_{t=s+1}^{k-1}\sum_{u=t+1}^{k} n_r n_s n_t n_u\right\}.$$

Interval data (under the randomization assumption)

The generalized Moran statistic is

$$I = \frac{n \sum_{ij} w_{ij} z_i z_j}{W \sum_{i=1}^{n} z_i^2},$$

$$\mu_1' = -(n-1)^{-1},$$

$$\mu_2 = \frac{n[(n^2 - 3n + 3)S_1 - nS_2 + 3W^2] - b_2[(n^2 - n)S_1 - 2nS_2 + 6W^2]}{(n-1)^{(3)} W^2} - (n-1)^{-2}.$$

Ranked data

$$I_R = \frac{n \sum\limits_{ij} w_{ij}(R_i - \bar{R})(R_j - \bar{R})}{W \sum\limits_{i=1}^{n} (R_i - \bar{R})^2}$$

$$\mu_1' = -(n-1)^{-1},$$

$$\mu_2 = \frac{1}{5W^2(n-1)^2(n+1)} [n(n-1)(5n+6)S_1 - (5n+7)(nS_2 - 3W^2)] - (n-1)^{-2}.$$

Explanation of symbols:

Nominal data: w_{ij} = weight given to join (edge) between localities i and j. w_{ij} need not equal w_{ji}.

$\sum\limits_{ij}$ = summation over all i (1 . . . n) and j (1 . . . n), i ≠ j. That is, summation of the full $n \times n$ matrix *except for diagonal entries.*

$(rr)_{ij}$ will equal 1 when localities i and j are both type r, and will equal zero otherwise.

$(rs)_{ij}$ will equal 1 when locality i is r and j is s, or when i is s and j is r. It will equal zero otherwise.

μ_1' = expected value (first moment)

μ_2 = variance (second moment about the mean)

$W = \sum\limits_{ij} w_{ij}$, the sum of the matrix of weights (except for the diagonal entries, if any).

$$S_1 = \tfrac{1}{2} \sum\limits_{ij} (w_{ij} + w_{ji})^2$$

$$S_2 = \sum\limits_{i=1}^{n} (w_{i.} + w_{.i})^2$$ where $w_{i.}$ and $w_{.i}$ are sums of the ith row and ith column of the weight matrix, respectively.

n = number of localities in study.

$n^{(2)} = n(n-1)$,

$n^{(3)} = n(n-1)(n-2)$, and so forth.

n_r, n_s, n_t, n_u = number of localities of types r, s, and t or u, there being k types in all.

Interval data: As above and $z_i = X_i - \bar{X}$ where X_i is the value of variable X for locality i and $\bar{X}$ is the mean of X for all localities.

$$b_2 = \sum\limits_{i=1}^{n} z^4 / \left(\sum\limits_{i=1}^{n} z^2 \right)^2 , \text{ a sample coefficient of kurtosis.}$$

Ranked data:

R_i = rank of locality i for the variable under study.

$\bar{R}$ = mean of the ranks = $(n + 1)/2$.

Note: The formulas are extracted in slightly modified form from a more extensive list in Cliff & Ord (1973).

APPENDIX B

Small-sample corrections

The statistics discussed in Appendix A are asymptotically normal, but deviate from normality at small sample sizes. As a result of Monte Carlo simulations, Cliff & Ord (1971, 1973) suggest ad hoc small-sample corrections to be used when $10 \leqslant n \leqslant 50$. Several general points are relevant to these corrections:

(1) They are not recommended when one locality figures in more than about 30% of the joins.

(2) For $n \leqslant 8$, Cliff & Ord recommend explicitly working out the distribution of the statistic under random permutations.

(3) The corrections are only valid in the tails of the distribution. That is, they are not appropriate for $\alpha > 0.10$ in each tail.

(4) The use of general rather than binary weights speeds the approach to normality.

Nominal data

Binary nominal data are viewed by Cliff & Ord as the "worst case" for small samples, since the distribution is necessarily concentrated at a few points, and is thus non-normal. These authors have not simulated cases with $k \geqslant 3$ types (colours), but the correction for like types presumably remains the same.

In the formulas that follow symbols for moments (μ'_1 and μ_2) will always relate to the specified combination of types; the symbols have therefore not been subscripted for these combinations.

Localities of the same type: (under assumption of sampling without replacement)

This correction is appropriate when $0.2 \leqslant n_r/n \leqslant 0.8$ and works better when n_r/n is above 0.5. Thus the common, rather than the rare types, should be tested. First compute degrees of freedom as

$$\nu = 2(\mu'_1)^2/\mu_2 .$$

The critical join count for rr joins is

$$rr_\alpha = \mu_2 \chi^2_{[\nu]}/2\mu'_1 .$$

For the lower tail employ $\chi^2_{1-\frac{\alpha}{2}[\nu]}$, for the upper tail $\chi^2_{\frac{\alpha}{2}[\nu]}$.

When ν is large ($\nu > 100$) the following expression employing the normal approximation to the χ^2 distribution is equivalent

$$rr_\alpha = \frac{\mu_2 [t_{\alpha[\infty]} + \left(\dfrac{4(\mu'_1)^2 - \mu_2}{\mu_2}\right)^{\frac{1}{2}}]^2}{4\mu'_1} .$$

Localities of different types: (this approximation is for the case of $k = 2$ types only)

Testing the upper tail, set $k_\alpha = 1 - \sqrt{10\alpha}$.

8*

Testing the lower tail:

in sampling with replacement, set $k_\alpha = 2[1 - \sqrt{10\alpha}]$,

in sampling without replacement, set $k_\alpha = 3[1 - \sqrt{10\alpha}]$.

Then compute the critical value for the number of rs joins

$$rs_\alpha = t_{\alpha[\infty]}\mu_2^{\frac{1}{2}} + \mu_1' - k_\alpha.$$

Interval data (for Moran's coefficient I)

Compute A, the number of edges in the connecting graph. In a symmetrical binary weight matrix this equals $\frac{1}{2}W$.

Compute $R = 2(\sqrt{n} - 1)/\sqrt{n}$,

$\qquad Q = 2(\sqrt{n} - 1)(2\sqrt{n} - 1)/n$.

If $R < A/n \leqslant Q$ set $k_\alpha = (10\alpha)^{\frac{1}{2}}$ for one-tailed testing, or $k_\alpha = (5\alpha)^{\frac{1}{2}}$ for two-tailed testing.

Otherwise set $k_\alpha = 1$.

Compute critical value of I as follows

$$I_\alpha = t_{\alpha[\infty]}\mu_2^{\frac{1}{2}} - k_\alpha(n - 1)^{-1}.$$

Note that when $k_\alpha = 1$ the above equation evaluates the critical value assuming a normal distribution of the statistic.

Notes: Except for symbols defined above, the symbols employed in this table are explained at the end of Appendix A. We have used $t_{\alpha[\infty]}$ above to indicate a normal deviate or Student's distribution for infinite degrees of freedom which amounts to the same thing. The formulas given here are those recommended by Cliff & Ord (1973) for the most common cases. Readers carrying out tests on small data sets and those obtaining results with borderline significance are urged to consult the original reference for various comments and suggestions for alternative tests.