Daniel Wartenberg

Multivariate Spatial Correlation: A Method for Exploratory Geographical Analysis

In this paper, I develop a multivariate extension of the univariate method of spatial autocorrelation analysis that I call multivariate spatial correlation (MSC). By accounting for the spatial dependence of data observations and their multivariate covariance simultaneously, complex interactions among many variables in a geographic context are analyzed. Using a methodological scheme borrowed from the techniques of principal components analysis (PCA) and factor analysis, a strategy for the exploratory analysis of spatial pattern in the multivariate domain is developed.

Spatial autocorrelation analysis is a statistical approach for quantifying the spatial relations among a set of univariate data observations. Since many processes occur in a geographic context, allowance for spatial dependence is essential in the analysis of geographically distributed data (Griffith 1978; Cliff and Ord 1981). Multivariate analysis is an array of statistical methods for quantifying the relations among many variables in a set of observations. Since many processes involve more than one variable, allowance for their dependence on each other is essential in modeling and in understanding their covariance (Morrison 1976). This paper is an attempt to define an analytical technique that accommodates both of these considerations simultaneously, and examines the spatial dependence of multivariate observations.

METHODS

Spatial autocorrelation is defined in terms of univariate data observations. Moran's coefficient I (Moran 1948, 1950; Cliff and Ord 1981), for example, is the weighted sum of the product of separate data observations, centered to the expected value of the observations, standardized to adjust for the variance of the observations, and normalized for the total sum of the weights. The following

Contribution No. 539 in Ecology and Evolution from the State University of New York, Stony Brook. This work is part of a doctoral dissertation submitted to the Department of Ecology and Evolution, State University of New York, Stony Brook. The author thanks Drs. R. R. Sokal, F. J. Rohlf, J. D. Thomson, and R. C. Grimson for comments on this manuscript and Drs. N. Oden and R. W. Setzer for many hours of helpful discussions. A reviewer pointed out the problem of negative eigenvalues and provided helpful guidance. B. Thomson and D. DiGiovanni assisted with technical aspects of this study. This research was supported by grant GM 2826202 from the National Institute of General Medical Sciences to R. R. Sokal.

Daniel Wartenberg is fellow, Interdisciplinary Programs in Health, Harvard School of Public Health.

Geographical Analysis, Vol. 17, No. 4 (October 1985) © 1985 Ohio State University Press
Submitted 11/84. Revised version accepted 4/85.

formula for this coefficient is given:

$$I = \frac{n \sum_{(2)} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{S_0 \sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{1}$$

where

$S_0 = \sum_{(2)} w_{ij}$
w_{ij} = weight for locality pair (i, j)
x_i = observation at locality i
$\bar{x}$ = mean of x_is
$\sum_{(2)} = \sum_{i=1}^{n} \sum_{j=1}^{n}$ for $i \neq j$.

This spatial autocorrelation coefficient is analogous to a Pearson product-moment correlation coefficient, but the terms within the summation in the numerator are each weighted by an interlocality factor, w_{ij}. By algebraic rearrangement,

$$I = \frac{1}{S_0} \sum_{(2)} w_{ij} \frac{z_i^* z_j}{\frac{1}{n} \sum_{i=1}^{n} z_i^2} \tag{2}$$

variables as above and $z_i = x_i - \bar{x}$.

Alternatively, if we standardize z_i prior to analysis and scale the weights w_{ij} to sum to 1.0, then

$$I = \sum_{(2)} w_{ij}^* z_i'^* z_j', \tag{3}$$

where

$$z_i' = \frac{(x_i - \bar{x})}{\sqrt{\left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2\right)}}$$

$$\sum_{(2)} w_{ij} = 1.0.$$

If, instead of using univariate data, we define each observation as a vector of individual observations of m variables, we can similarly define a matrix of coefficients, **M**:

$$\mathbf{M} = \mathbf{Z}^t \mathbf{W} \mathbf{Z}, \tag{4}$$

where

M is an m by m, variable by variable, spatial correlation matrix
Z is an n by m, location by variable, standardized and centered (by variable) data matrix
$\mathbf{Z}^t$ is an m by n, variable by location, standardized and centered (by variable) data matrix, the transpose of **Z**
W is an n by n, locality by locality, weight matrix.

Each coefficient in the matrix **M** is a Mantel-type coefficient (Mantel 1967). That is, each coefficient is a general cross-product statistic among elements of two matrices in which these elements are distances (or similarities) among pairs of objects (Hubert, Golledge, and Costanzo 1981). The distributional properties of each diagonal element of **M** are the same as for univariate autocorrelation values. Indeed, the diagonal values are themselves Moran's I coefficients. Each off-diagonal element is, by analogy, a bivariate cross-correlation coefficient, the spatial correlation of one variable with another variable calculated by summing the values over all pairs of localities, and weighted as in the autocorrelations. One such coefficient exists for each pair of variables. The expected value and variance of these coefficients under a permutational hypothesis have been derived by Mantel (1967), but their distribution is unknown. For large sample sizes, the distribution is often asymptotically normal, but deviations from normality are not unusual. Klauber (1975), developing a multivariate analytic approach similar to MSC, derived expectation and variance equations for the cross-product statistics when more than two samples exist. However, he himself notes the limitations of the use of these statistics given the unusual distribution of the raw data (i.e., reciprocal distance). In addition, all the problems of significance testing in PCA, such as multiple comparisons and tests on successive factors after one is found to be significant, apply equally well to MSC. Given these problems and the fact that the full distributional properties of these coefficients have not been worked out, this approach to assessing significance will not be addressed here any further.

An alternative derivation can be given in which the spatial correlation methodology is thought of as a part of a generalized principal components analysis (Appendix). This approach is relevant for statistical modeling of the covariance structure of the data. However, a discussion of this issue is beyond the scope of this paper and will be explored elsewhere.

This spatial correlation matrix, **M**, which is in quadratic form, can be decomposed into orthogonal components using eigenvector analysis. These components, as in PCA, reflect the distribution of variation, in this case spatially weighted variation, throughout the multivariate data field. All statements made in reference to spatial variance or spatial components use these terms for convenience by analogy with PCA and should not be interpreted in the strict statistical sense. The first component explains the maximum amount of variance that can be explained by a linear combination of the variables. The second component explains the maximum amount of residual variance (i.e., that not explained by the first component) that can be explained by a linear combination of the original variables, while remaining orthogonal to the first component. A third component can be extracted that is orthogonal to the first two, and so on. Those components explaining the major portions of the variance should depict the basic patterns of spatial patches and trends, when mapped.

An important difference between this approach and PCA must be pointed out. Unlike **R**, the product-moment correlation matrix that is decomposed in PCA, **M** is not positive definite. That is, **M** can have negative eigenvalues, which **R** cannot.

These negative eigenvalues are as important as positive eigenvalues but are of a qualitatively different type. They represent spatial interaction (covariance) that is more important than spatial pattern (variance). A thorough discussion of this topic is beyond the scope of this paper and will be presented elsewhere. To avoid this situation, data yielding negative eigenvalues are not used in this paper. All examples have large eigenvalues that are positive only. Sums of eigenvalues used for comparisons are all sums of the absolute values of eigenvalues.

Loosely following the methodology of principal components analysis and factor analysis, it is possible to extract the eigenstructure of **M** and derive the variable loadings on these resulting axes. Each eigenvector is that linear combination of the original variables, orthogonal to all the earlier eigenvectors, that explains the maximum amount of variance not already explained. The component loadings are the correlations of the original variables with these orthogonal axes. The axes retained as significant can then be rotated to simple structure (i.e., rotated to maximize the number of squared loadings near 1.0 or 0.0), and the locality scores can be derived by projecting the original data points onto these rotated component axes. The structure of the axes should reflect the coincidence of spatially important (i.e., highly weighted) variables. The locality scores should show the contributions of the individual samples to this structure, that is, which localities are most important in determining this pattern.

In interpreting these results, it is important to remember that eigenvectors represent contrasts between variables and explain a maximum amount of variance. Eigenvectors are rotated to simple structure, so that squared loadings approach either 1.0 or 0.0, to facilitate easier interpretation and description. Large positive (and large negative) loadings are emphasized to maximize the variance explained by each axis, and it is the magnitude of these loadings that is of importance. The signs of the loadings on any one factor are arbitrary and all could be multiplied by -1 with no change in meaning.

To demonstrate the utility of this approach for detecting spatial pattern, I analyze 3 types of data. First, the sensitivity of the technique to detect a single patch (small-scale homogeneity, large-scale heterogeneity) or a single linear trend (heterogeneity at all scales) in simulated data is tested. Replicate tests are run to estimate their reliability. Second, two more sets of artificial data are examined to demonstrate qualitatively the accuracy with which the technique recovers patterns. Within each of these data sets, again, there are patches or trends. Third, I analyze two real data sets. The first is the distribution of 21 HLA blood group allele frequencies among 58 localities in Europe. The geographical pattern of these data have been studied to infer migrational history of European peoples (Menozzi, Piazza, and Cavalli-Sforza 1978; Sokal and Menozzi 1982; Wartenberg 1985a). In this paper, I reassess these patterns. The second is abundances of 26 species of Foraminifera in Atlantic and Indian Ocean sediment core tops. These data were used to construct a regional grouping of species distributions that was then related to climatic parameters (Imbrie and Kipp 1971). I reassess Imbrie and Kipp's regionalization as well as another geographic analysis of these data that I have done (Wartenberg 1985a).

For all analyses in this paper, the interlocality weights used in the calculations are proportional to the inverse of the square of the geographical separation distance between localities. Other functional forms of the separation distance could be used for weighting (e.g., inverse distance, inverse log distance, inverse distance to the fourth power, etc.), but the weights chosen have been shown to be generally most reliable for geographic analysis (Crain and Bhattacharyya 1967). Variables other than geographic separation can be used for weighting, if investigators wish to assess pattern as a function of these other variables (e.g., using group membership in

discriminant analysis or another set of characteristics of the same objects—G. F. Estabrook, personal communication).

RESULTS

1. Sensitivity

Two types of patterns are simulated to evaluate the sensitivity of this proposed technique for detecting simple spatial structure. The first is created by filling a square grid of a specified size with random, normal (0, 1) deviates for each of 10 variables. To the first variable, for the left half of the grid, a specified increment is added to simulate a patch. Thus,

$$Y_{ijk} = \epsilon_{ijk} + INC \text{ for } k = 1 \text{ and } i < \frac{I}{2}$$

$$= \epsilon_{ijk} \text{ otherwise},$$

where

i is the row index
j is the column index
k is the variable index
I is the number of rows
J is the number of columns
K is the number of variables (10 in this case)
ϵ_{ijk} are random, $N(0, 1)$, independent deviates
INC is the increment added to the ϵs
Y_{ijk} is the observed grid value.

One variable is spatially patterned and nine are not. Calculated next are the values of Moran's I for the patterned variable and the ratio of the first eigenvalue to the sum of the absolute values of all the eigenvalues for both the spatial correlation matrix for all variables and the Pearson product-moment correlation matrix for all variables. The single spatial autocorrelation coefficient for the first variable is an index of the univariate spatial structure for the variable with the added increment. Spatial autocorrelation for the other variables should not be significantly different from expectation and should not vary as the increment added to the first variable changes. The ratios represent the relative magnitude of the first eigenvalue in spatially weighted and unweighted correlation matrices, respectively, to the total variance. They reflect the effectiveness of each method in detecting structure of the one spatially patterned variable in an otherwise unpatterned multivariate data set. For an increment of 0.0, no pattern should be detected. This experiment was replicated 25 times for each set of parameter values. The means and standard errors of the indexes are tabulated in Table 1 and Table 2 for grid sizes 36 and 100, respectively, for increments ranging from 0. to 10.0.

Table 1 shows that for a 6-by-6 grid, an increment of 1.0 produces an increase in Moran's I and a slight change in the eigenstructure of the spatial correlation matrix. Larger increments show marked changes in both of these. The Pearson product-moment correlation matrix shows no overall change.

Table 2, for the 10-by-10 grid, shows a similar pattern, although the change appears to begin earlier for Moran's I, at an increment of 0.5. The change in eigenvalue ratio for M is suggestive at an increment of 0.5 and marked at an increment of 1.0. Thus, for these sample sizes, this technique is marginally able to detect displacements of one standard deviation of the overall surface values. For

TABLE 1
Results of Simulation 1: The Spatial Structure of a Patch Model on a 6-by-6 Grid

	Spatial Ratio		Pearson Ratio		Moran's *I*	
Increment	Mean(%)	SE	Mean(%)	SE	Mean	SE
0.00	24.10	0.53	19.51	0.30	− 0.04	0.01
0.10	23.30	0.52	19.23	0.41	− 0.02	0.01
0.20	23.74	0.67	19.37	0.29	− 0.04	0.01
0.50	24.11	0.66	19.81	0.32	− 0.01	0.01
1.00	25.47	0.67	19.43	0.28	0.08	0.01
2.00	32.49	1.09	19.57	0.28	0.22	0.01
5.00	41.57	0.81	19.00	0.29	0.42	0.01
10.00	42.49	0.76	18.92	0.29	0.45	0.00

NOTES: The increment is the value added to each random, normal deviate for the first variable. The two ratios are the ratio of the first eigenvalue to the absolute value of the sum of all the eigenvalues times 100 percent, for each of the specified correlation matrices. Moran's *I* is the value of that coefficient for the first variable. The mean and standard error of 25 replicates are given. See text for details.

TABLE 2
Results of Simulation 2: The Spatial Structure of a Patch Model on a 10-by-10 Grid

	Spatial Ratio		Pearson Ratio		Moran's *I*	
Increment	Mean (%)	SE	Mean (%)	SE	Mean	SE
0.00	21.90	0.72	15.08	0.16	− 0.01	0.01
0.10	22.46	0.77	15.51	0.22	− 0.01	0.01
0.20	23.33	0.68	15.22	0.23	− 0.01	0.01
0.50	24.03	0.79	15.26	0.16	0.03	0.01
1.00	31.63	1.00	15.48	0.20	0.12	0.01
2.00	45.18	0.68	15.26	0.16	0.29	0.01
5.00	56.83	0.57	15.53	0.16	0.50	0.00
10.00	59.81	0.48	15.56	0.15	0.56	0.00

NOTES: See notes to Table 1.

larger height changes, the method depicts clear change. As sample size increases, there is also a suggestion of increasing sensitivity.

For the second sensitivity test, a linearly increasing trend term is added to the first variable of random normal deviates. The total displacement of the trends ranges from 0.0 to 5.0 across the entire grid area along one axis. Thus,

$$Y_{ijk} = INC * \frac{i}{I} + \epsilon_{ijk} \text{ for } k = 1$$

$$= \epsilon_{ijk} \text{ for } k > 1$$

variables defined as above.

This simulation was also replicated 25 times. The means of the indexes and the standard errors are given in Tables 3 and 4. For both grids, the value of Moran's *I* for the first variable increased steadily throughout the range of trends tried. For the 6-by-6 grid, the spatial correlation matrix index begins to show change for a maximal displacement of 2. For the 10-by-10 grid, it begins to change at 1.0. Again, for displacements just over one standard deviation of the overall surface values, this method begins to detect structure, and sensitivity appears to increase with sample size. The ratios for the Pearson product-moment correlation matrix show no overall change.

The sensitivity of this method of analysis seems dependent on grid size and maximal displacement. For small grid sizes, for example, 36 localities, it may be

TABLE 3
Results of Simulation 3: The Spatial Structure of a Trend Model on a 6-by-6 Grid

	Spatial Ratio		Pearson Ratio		Moran's *I*	
Increment	Mean (%)	SE	Mean (%)	SE	Mean	SE
0.00	23.34	0.73	19.57	0.30	− 0.02	0.01
1.00	24.08	0.96	19.18	0.33	0.00	0.01
2.00	26.30	1.28	18.56	0.20	0.08	0.01
3.00	30.32	1.14	19.35	0.32	0.19	0.01
4.00	33.77	1.14	19.40	0.34	0.25	0.01
5.00	37.60	0.92	19.69	0.38	0.31	0.01

NOTES: See notes of Table 1, except increment in this case refers to the maximal surface displacement along one edge of the grid.

TABLE 4
Results of Simulation 4: The Spatial Structure of a Trend Model on a 10-by-10 Grid

	Spatial Ratio		Pearson Ratio		Moran's *I*	
Increment	Mean (%)	SE	Mean (%)	SE	Mean	SE
0.00	22.13	0.64	15.13	0.16	− 0.01	0.00
1.00	36.96	1.03	15.06	0.14	0.17	0.01
2.00	46.60	0.98	15.66	0.22	0.32	0.01
3.00	54.11	0.64	15.18	0.14	0.45	0.01
4.00	55.62	0.53	15.00	0.18	0.49	0.00
5.00	58.86	0.61	15.35	0.18	0.53	0.00

NOTES: See notes to Table 1, except increment in this case refers to the maximal surface displacement along one edge of the grid.

necessary to have changes of 1–3 standard deviations to be detected. For 100 localities, changes of 1 standard deviation were clear-cut. For more complex patterns, sensitivity will vary but regular patterns superimposed on random surfaces can be detected.

2. *Accuracy*

To assess the multivariate resolution of this method, two more simulated data sets were created. These data sets have spatial patterns for more than one variable. The sampling design is shown in Figure 1. The 36 localities are located on a square grid and are divided into four groups: A, B, C, D. Each locality is assigned a random, independent, normal (0, 1) deviate for each of eight variables in each study. For the first simulation in this set, an increment of 3.0 is added to all localities in sections A and B for the first variable, and a like increment is added to all localities in sections B and D for the second variable. Six additional spatially random variables, equivalent to an increment of 0.0 in all sections, are included (Figs. 2A–2H). I then calculate the M matrix and extract the eigenstructure. Two components account for over 98 percent of the variance. The eigenvectors are rotated obliquely to simple structure using the Harris and Kaiser (1964) criterion, and the standardized data are then projected onto these axes. The results are shown in Figures 2I and 2J. Figure 2I shows the contrast between AC and BD, as in variable 2. Figure 2J shows the contrast between AB and CD as in variable 1. The input data structure is clearly revealed by these analytic results. A similar analysis by PCA with oblique rotation of the first 2 components (Figs. 2K, 2L) reveals no discernible geographic patterns and the first two components explained only 42 percent of the variance. (Since the locality scores and projections of data observations onto eigenvectors, the definition of positive and negative is arbitrary and can be reversed. It is the magnitude that is of interest.)

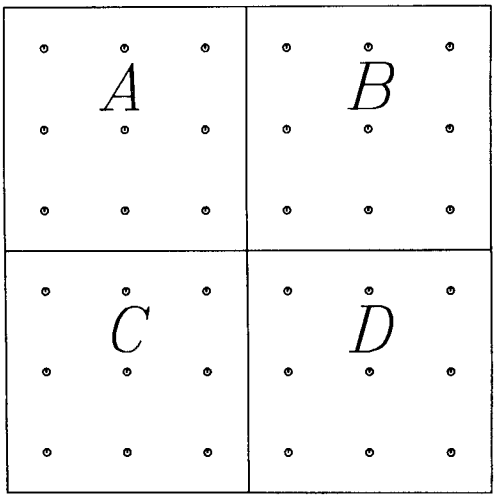

FIG. 1. The Sampling Design for the Simulation Experiments. There are 36 localities divided into 4 regions, A, B, C, and D.

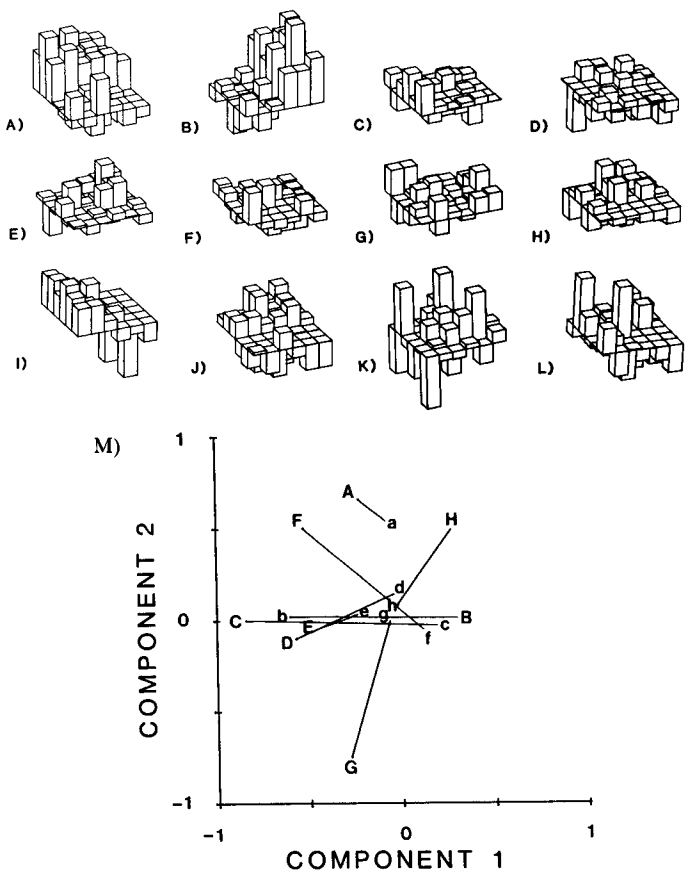

FIG. 2. Geographic Maps for the First Simulation. Frames A–H are input variables. Various values are added to underlying random, $N(0, 1)$ deviates as described in the text. Frames I, J are maps of rotated MSC scores, frames K, L are maps of rotated PCA scores. Frame M is a plot of the component loadings for PCA and for multivariate correlation on the first two component axes, after oblique rotation. The uppercase letters are the PCA loadings and the lowercase letters are the loadings from M. The letters correspond sequentially to the variables (i.e., *A* and *a* are the first variable, *B* and *b* are the second variable, etc.).

Another way to look at these data is to plot on the same set of axes the loadings on the first two components for each variable for PCA and for MSC. One can then assess how the relative position of each variable in this space changes, based on the spatial weighting. Variables with strong spatial structure should remain far away from the origin, although their orientation may change. Variables with weak spatial structure should end up closer to the origin.

A plot of this type, in which loadings from PCA and MSC have been rotated obliquely, is shown in Figure 2M. The uppercase letters represent the PCA loadings and the lowercase letters the loadings from MSC. The solid lines depict the change in position of the variables from the PCA solution to that for MSC, that is, that due to spatial weighting. In this case, only the first (A) and second (B) variables are far away from the origin for MSC while most variables are far away from the origin for PCA. As A and B have spatial structure, by design, while none of the other variables do, this representation is consistent with what we know about the variables and emphasizes the spatial pattern.

The next simulation introduces terms with trends rather than patches. Again, the grid in Figure 1 is filled with independent, random, normal (0, 1) deviates for each variable. The first variable is incremented from left to right, by values ranging from 0.0 to 3.0 (Fig. 3A). The second variable similarly is incremented from front to back increment (Fig. 3B), while the next six variables are left spatially random (Figs. 3C–3H). The data are analyzed as above and yield two components that account for 91 percent of the variance. The first is a front-to-back contrast (Fig. 3I), the second is a left-to-right contrast (Fig. 3J). The PCA results for the same data (Fig. 3K, 3L) do not show distinct geographic patterns. The PCA solution explains only 37 percent of the variance.

The plot of PCA loadings and the MSC loadings from (Fig. 3M) is similar to that for the first simulation. The first two variables (A and B) maintain their importance in both types of analysis, although their orientation switches, while the other variables lose some of their importance (i.e., end up closer to the origin) in the spatially weighted case. The component scores are more informative than the loadings, but the loadings generally are consistent with our knowledge of the data.

Additional simulations were run for more complex patterns and the results were consistent with those reported here. In summary, in all simulations MSC depicted the geographic pattern that was put in. PCA was much less effective at describing these patterns. Plots of the component loadings helped describe the way in which MSC was sensitive to geographic pattern.

3. *HLA Human Blood Group Data*

The next test of the proposed methodology is to analyze a real rather than a simulated data set. The data I have chosen are gene frequencies of 21 alleles of the HLA-A and HLA-B human blood systems measured in 58 European and Near Eastern populations (localities). The geographic patterns of these data have been studied by Menozzi et al. (1978), Sokal and Menozzi (1982), and Wartenberg (1985a). Blood type characteristics are indicative of a population's origin and heritage. Differences in blood types between populations dissipate through interbreeding. The expressed goal of Menozzi et al. (1978) was to map synthetic variables, statistical composites of genetic (blood type) variables, from which to infer the evolutionary history of the populations studied. They constructed these synthetic variables using PCA. Since genetic distance between populations should be proportional to the time of separation and inversely proportional to the intermigration between them, the history of geographic movement should be revealed from the study of these maps (Cavalli-Sforza and Bodmer 1971; Menozzi et al. 1978). Using PCA, Menozzi et al. were able to summarize over half of the

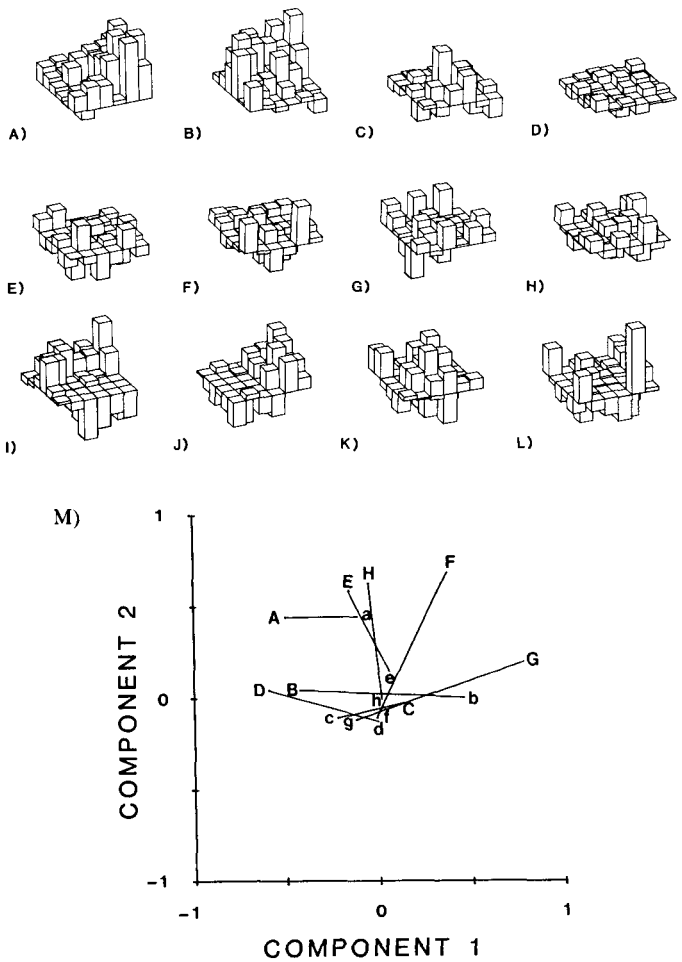

FIG. 3. Geographic Maps for the Third Simulation. Frames A–H are input variables. Various values are added to underlying random, $N(0, 1)$ deviates as described in the text. Frames I, J are maps of rotated MSC scores, frames K, L are maps of rotated PCA scores. Frame M is a plot of the component loadings as in Figure 2.

variation of 38 human blood group allele frequencies at 67 localities into 3 orthogonal principal component axes, or synthetic variables. They constructed maps of these synthetic variables and argued that these maps represented a geographic depiction of the covariance of allele frequencies. From these maps, they inferred the migrational history of early European populations and found patterns roughly coincident with the hypothesized spread of early farming from the Near East.

A complementary approach for indirect geographic analysis was used by Sokal and Menozzi (1982) on a subset of the same data (21 allele frequencies of the HLA-A and HLA-B blood systems at 58 European localities). These authors, applying univariate spatial autocorrelation analysis, described the spatial correlograms of the geographic pattern for each variable. Then, they looked for pattern among correlograms of the allele frequencies from which to infer the covariation of the allele frequencies. Cluster analysis of the similarity of the correlograms and the similarity of the original data observations calculated for all pairs of variables yielded 3 basic patterns. These patterns were similar to those

found by Menozzi et al. (1978). Sokal and Menozzi (1982) emphasized the relationships of the parameters of the geographic patterns (similarities of the correlograms of the variables) to parameters of the multivariate pattern (correlations between the variables). Their conclusions about the migrational history of early European populations were consistent with those of Menozzi et al. (1978).

Wartenberg (1985a) studied the same data subset as Sokal and Menozzi (1982) and applied the method of canonical trend surface analysis (CTS). By constructing variance-covariance matrices of the genetic variables (blood type alleles) and the geographic variables (coordinates, their squares and cross products) and taking the joint eigenstructure, he constructed maps of the overall geographic patterns. These, too, were consistent with the earlier analyses. He also showed, however, that if additional data without geographic pattern were included in the analysis, only CTS would be able to recover the underlying geographic information.

Details of the data set used in this study are given in Sokal and Menozzi (1982). A map of the localities is shown in Figure 4. The spatial correlation matrix is calculated using inverse distance squared weighting, and the eigenstructure extracted (Table 5). From consideration of a scree plot (Cattell 1978), I retain two components as most important. They account for 80.6 percent of the variance. The next few components account for patterns of lesser importance (corresponding to a second scree) and the final components correspond to the error variance (the first scree). I obliquely rotate the first two components to simple structure using the Harris-Kaiser criterion (Harris and Kaiser 1964) and project the standardized data onto these axes. The resulting locality scores are shown in Figures 5A and 5B.

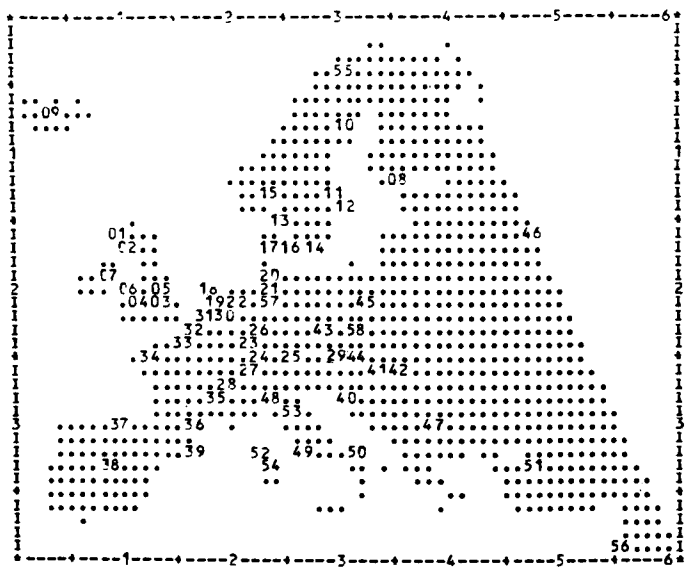

FIG. 4. A Map of Europe and the Near East Showing the 58 Localities Where the HLA Blood Group Gene Frequencies Were Sampled

All maps of the HLA data in this paper were produced by means of the SYMAP computer contouring program (Dougenik and Sheehan 1979) using a Lambert Azimuthal, equal area projection, centered at 0 degrees latitude and 7.5 degrees east longitude. There is a one-to-one correspondence between areal sizes on such a map and true areal sizes on the spherical Earth.

A clear north-south pattern exists across the entire map of the first component, with various aberrations toward the center. Highest values are noted in Scandinavia,

TABLE 5

Results of MSC on HLA Data Set I: Absolute Value of the Eigenvalues of the Spatial Correlation Matrix M

Component Number	Eigenvalue	Percentage of Total
1	7.65	61.75
2	2.33	18.81
3	0.64	5.15
4	0.56	4.51
5	0.44	3.57
6	0.25	2.06
7	0.15	1.18
8	0.12	0.95
9	0.07	0.60
10	0.04	0.32
11	0.03	0.21
12	0.02	0.21
13	0.02	0.17
14	0.01	0.12
15	0.01	0.11
16	0.01	0.11
17	0.01	0.06
18	0.01	0.06
19	0.00	0.03
20	0.00	0.02
21	0.00	0.02

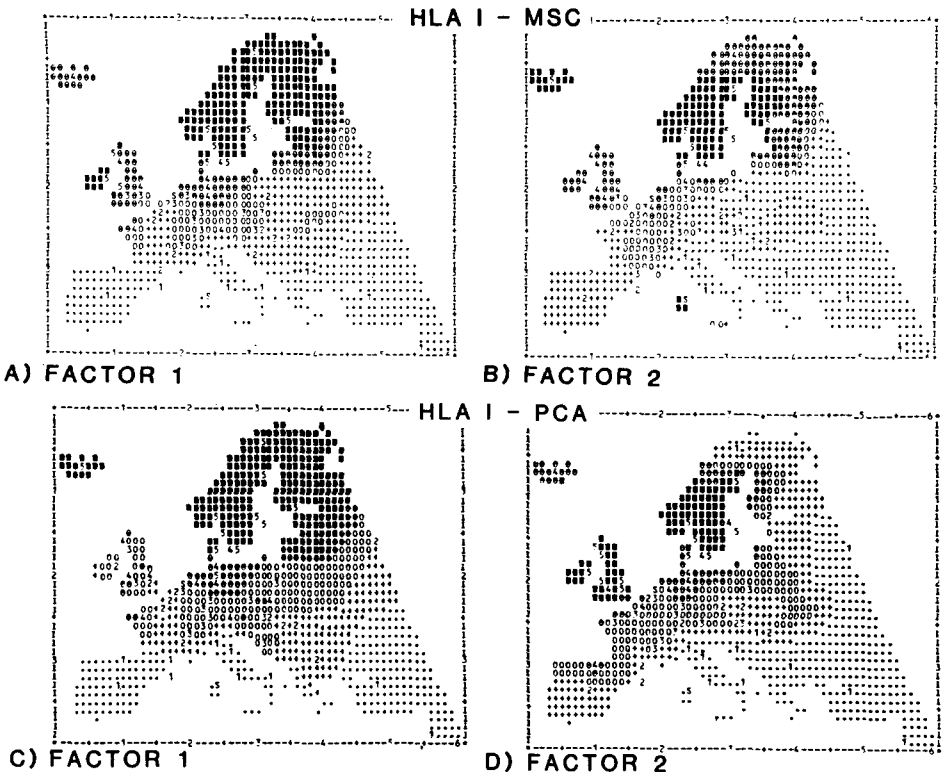

FIG. 5. Contour Maps of the PCA and MSC Rotated Component Scores from HLA Blood Group Data Set I Analysis. Frame A is the first MSC component, frame B the second MSC component, for the data alone. Frame C is the first PCA component, frame D the second PCA component.

Ireland, and Scotland with values of decreasing importance in England, Iceland, and central Europe. Lowest values occur in the Near East, Italy, Greece, and Spain.

The second map shows a superficially similar but noticeably different pattern. High values occur in southern Scandinavia and Iceland, with intermediate values throughout the United Kingdom, in northern and eastern Scandinavia, and central Europe. Low values are mainly in Spain, Italy, Greece, and the Near East. These patterns are consistent with the earlier analyses of these same data, which reported north-south or northwest-southeast trends. The earlier studies postulated that this pattern reflected the migrations of people from the Near East, first west and then north, and that the migrating people carried the technology of farming with them. I have found more localized patterns with MSC. A north-south pattern is separated from a northwest-southeast pattern, and enhanced structure is found in central Europe, where sampling is most dense. An analysis with weights that change more gradually with distance (e.g., inverse distance) would detect even more broad scale patterns.

The maps of the PCA scores, also rotated obliquely to simple structure, show similar patterns (Figs. 5C, 5D). As in the other analyses of these data, since the pattern in the data is primarily geographic both MSC and PCA give similar results.

The map of the component loadings shows a complex pattern of realignment. The diagram, however, with two labeled endpoints and one connecting line for each of 21 variables, is too complicated to include here. To be able to view them, variables must be examined one at a time. Most of the variables move, reflecting particular spatial covariance and only a few (d, f, h, t) move appreciably closer to the origin.

It is interesting to compare these results (i.e., obliquely rotated components) quantitatively with those obtained with other methods (i.e., orthogonal components). PCA yielded three basic patterns (Menozzi et al. 1978) as did univariate spatial autocorrelation (Sokal and Menozzi 1982) and canonical trend surface analysis (Wartenberg 1985a). Comparison of the resultant factor scores by Spearman's rank-order correlation coefficients (Table 6) shows that the first spatial factor of MSC corresponds to the first principal component axis and the first canonical trend surface. The second component from MSC corresponds to a contrast between the first and third from PCA and a contrast of the first and second surfaces versus the third from CTS. Since the axes derived for the present method are rotated obliquely, the simpler description obtained in this study may be a more realistic

TABLE 6

Spearman's Rank Correlations for HLA Data Set I of Scores on the First Two MSC Components with Component Scores on the First three PCA Components and the First Three CTS Components Different Techniques as well as Scores on the First Two MSC Components for HLA Data Set II

	HLA Data Set I	
Method	Component 1	Component 2
MSC-HLA I		
Component 2	0.669	
MSC-HLA II		
Component 1	0.774	0.834
Component 2	0.480	0.794
Orthogonal PCA		
Component 1	0.973	0.707
Component 2	0.005	− 0.296
Component 3	− 0.237	− 0.717
CTS		
Surface 1	0.914	0.611
Surface 2	0.194	0.423
Surface 3	− 0.268	− 0.401

representation of the data. Forced orthogonality may force the appearance of additional factors.

To test the sensitivity of the results of MSC to nongeographic information, I analyze a second set of data constructed to confound the allelic-geographic covariance of the HLA data with nongeographic but structured variation of hypothetical variables. Ideally, these additional variables will not affect the results. In a study of the sensitivity of PCA and CTS to the addition of spatially unstructured data to a spatially structured data set (Wartenberg 1985a), PCA responded to the overall variance pattern even though it did not have spatial pattern while CTS remained largely unchanged, still depicting only the geographic pattern.

This second data set, HLA data set II, is the original HLA data set with 10 correlated but spatially unpatterned variables added. Each set of 5 variables is based on a series of 58 random, independent, uniform numbers between 0.0 and 10.0. Each locality was assigned one of these numbers at random, and then 5 random, independent, uniform numbers between 0.0 and 1.0 were added to it separately to generate variates for that locality for each of the 5 variables. This procedure produced correlation structure among the 5 variables of each of the two sets, but resulted in no significant geographic pattern (i.e., autocorrelation) among any of the 10 new variables.

The first two eigenvalues of **M** account for 75.6 percent of the total variance. The component score patterns that result from the analysis of these data (Figs. 6A, 6B)

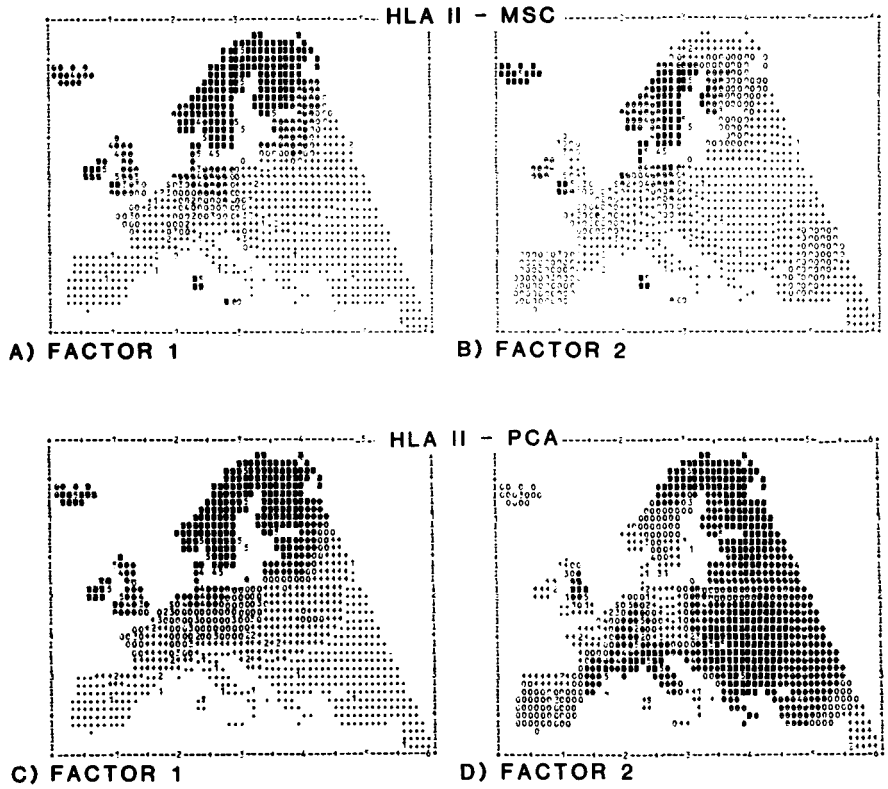

FIG. 6. Contour Maps of the PCA and MSC Rotated Component Scores from HLA Blood Group Data Set II Analysis, with Random Noise. Frame A is the first MSC component, frame B the second MSC component, for the data alone. Frame C is the first PCA component, frame D the second PCA component.

are very similar to those discussed above. Most differences in the maps occurred in areas of low data density, perhaps an artifact of the contouring program. The correlations between these component scores and the corresponding scores from data set I are all above 0.77 (Table 6). The technique is insensitive to spatially unpatterned noise.

The PCA results for the HLA data set II (Figs. 6C, 6D) are not as resistant to geographically random information. The first component is still similar to that obtained with HLA data set I, but the second component is quite different.

4. *Foraminifera Data*

The final data set I analyze is a set of species abundances of 26 species of Foraminifera sampled from the sediment core tops at 61 locations (Fig. 7) throughout the Atlantic and Indian Oceans collected by Imbrie and Kipp (1971). The goal of their original study was to derive statistically independent assemblages of species that could be used in multiple regression analysis for paleoecological reconstruction of climate. They discussed the geographic distribution of the species and argued that components derived by PCA would be geographically coherent. They mapped the component loadings (from a Q-mode analysis) which showed patterns corresponding to the basic climatic regimes (i.e., polar, subpolar, subtropical, and tropical) and circulation patterns (i.e., gyre margins, transitional zones) of the oceans (see also Kipp 1976; Wartenberg 1985b).

FIG. 7. A Map of the Atlantic and Indian Oceans Showing the 61 Localities Where Core Tops Were Taken by Imbrie and Kipp (1971) and the Foraminifera Identified

These data have also been examined by CTS (Wartenberg 1985a). The CTS results showed a regional structure in the first and second components that was similar to that summarized by the first four PCA axes. The next three CTS axes showed more detailed geographic structure that is coincident with circulation and biological production patterns.

Again, the patterns depicted by MSC are somewhat different from those derived with other methods, although the broader features are recovered similarly in all techniques. Two factors recover 69 percent of the variance. The subsequent 8

factors, the second scree, also appear to be indicative of pattern, but to a much lesser degree. The rest of the higher-order components, the first scree, seem unimportant. For simplicity, I will concentrate on the first two factors.

The first factor (Fig. 8A) shows a latitudinal zonation and is most highly concentrated in the trade wind region of the North Atlantic Ocean. The pattern falls off to the north and south, but there is a brief rise in the South Atlantic and Indian Oceans, also in the trade wind region. The contrast between regions corresponds well with climatic zones, as was depicted by the other methods of analysis, but the pattern in the trade wind regions is strongest. More fine scale detail with greater geographic relief is afforded by the surfaces of MSC than the smooth surfaces produced by CTS.

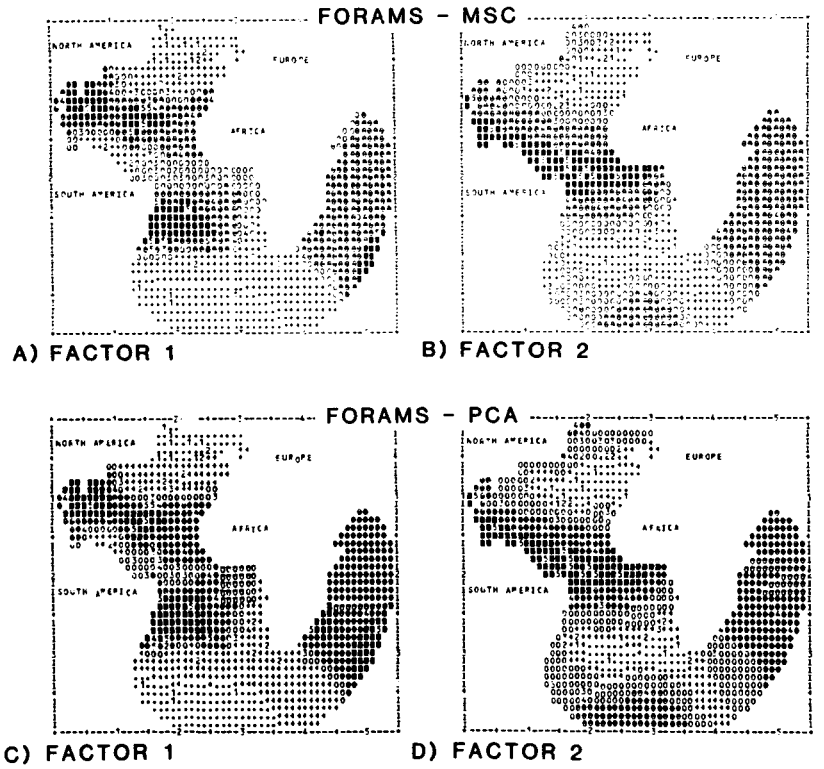

FIG. 8. Contour Maps of the PCA and MSC Rotated Component Scores for the Foraminifera Species Abundance Data from the Atlantic and Indian Oceans. Frame A is the first MSC component, frame B the second MSC component. Frame C is the first PCA component, and frame D is the second PCA component.

The second factor (Fig. 8B) has its highest values in the equatorial region of the oceans. The values fall off towards the poles, with intermediate values (i.e., those close to 0) in the trade wind regions. The variation in these intermediate areas was described by the first component. There is also a suggestion of pattern corresponding to the strong coastal margin currents along the eastern United States, western Africa, and southeastern Africa.

Maps of the obliquely rotated PCA components show a fairly similar pattern (Figs. 8C, D). As with the HLA data, since I have picked a data set with strong geographic pattern, most analytic techniques will depict the same geographic

pattern. In general, the pattern is that of the climatic zonation noted in the other analyses.

The joint picture from these two components (either the MSC or PCA) seems to be a composite of those represented by other (orthogonal) methods. Much of the information isolated into separate components by orthogonal PCA or separate surfaces by CTS are merged into a more highly patterned representation. An oblique representation is more complicated to interpret, but more economical in depiction. Rank correlations of the component scores from different methods are given in Table 7.

TABLE 7
Spearman's Rank Correlations between Scores on Components from MSC, PCA, and CTS for the Foraminifera Data

	Foram Data	
Method	Component 1	Component 2
MSC		
Component 2	0.209	
Orthogonal PCA		
Component 1	0.798	0.657
Component 2	− 0.346	0.778
Component 3	− 0.272	0.149
CTS		
Surface 1	0.550	− 0.697
Surface 2	0.023	− 0.534
Surface 3	0.108	0.290

The factor loading plot is a complex picture representing the correlational structure of all the variables and, again, too complicated to include here. Features of note are that two species (*Pulleniatina obliquiloculata* and *Sphaeroidinella dehiscens*) have taken on increased importance in the spatial analysis while four other species (*Globigerinita glutinata*, *G. crassaformis*, *G. scitula*, and *Globigerina digitata*) have ended up closer to the origin.

The MSC patterns depicted are interesting in that they emphasize the regions that show the most coherent geographic patterns and highlight the most geographically reliable species. While other methods may resolve differences in abundances most clearly, MSC highlights geographically predictable areas. The first MSC component has highest locality scores in the equatorial regions and gyres. These areas are thought to be stable in terms of species composition but are unproductive biologically. They are contrasted with the polar regions and the upwelling zones, which are less stable in composition but, when productive, have a relatively simple but different biological makeup. It is this predictable composition that is reflected in the sediment material. Other regions are less predictable and do not provide as much geographically useful information. This contrast between predictability and unpredictability was not emphasized by the other methods, although the difference between polar and tropical faunas was. The second MSC component shows a superficially similar but distinct pattern. The first two MSC components differ from each other in that they separate the equatorial regions from the gyres, which are separate systems of biological production. Thus, the MSC components both emphasize the overall biological structure of the Atlantic and Indian Oceans as depicted by the other methods, but they also describe additional features of stability (predictability).

DISCUSSION AND CONCLUSIONS

The purpose of this paper is to propose a new technique to study the spatial structure of multivariate data observations. Spatial autocorrelation is the dependence

of values of a variable on values of the same variable at geographically adjoining locations. Indexes of spatial autocorrelation have been studied extensively as indicators of univariate spatial pattern. By analogy, spatial correlation is the dependence of values of a variable on values of a second variable at geographically adjoining locations. A matrix of such correlations, when analyzed in a manner analogous to principal components analysis, yields a set of spatial factors, linear combinations of localities each of which jointly contribute to certain aspects of the overall spatial pattern. The result is an ordination of sites based on their multivariate similarity and conditioned on their geographic proximity.

A similar development can be made for Geary's c (Geary 1954). Lebart (1969) has developed a related approach for Geary's c in which locality values first are "differenced" with each other for the same variable (the method in the present paper does not restrict comparisons to the same variables), and then the multivariate covariance is assessed. It is a covariance analysis in difference space rather than in the original data space, as described above. The relationship between these two approaches is not addressed here.

To show how the proposed technique works, simple examples were generated. More complex tests using data of more intricate multivariate and spatial structure are needed to further understand the usefulness and limitations of this method. In general, the technique recovers the input patterns, and it also resolves plausible solutions for the two real data set. These data sets both had strong spatial patterns that were detected by MSC as well as other methods. However, when nongeographic information was added to the data, only the results from those methods that specifically assess spatial pattern (i.e., MSC and CTS) were not altered dramatically. Thus, MSC is insensitive to nonspatial information. Unlike nongeographic methods of analysis, only the geographically variable portions of the data are resolved. This decomposition into geographical and nongeographical should be useful in the study of the processes that control geographic patterns. Further, MSC is able to give a more local and higher-order result than CTS, when necessary.

In this paper, I have made little mention of the problems of component correlations (i.e., the effect of oblique rather than orthogonal rotations) and nonlinear responses. The distribution patterns of the data in both the simulated and real data are geographically overlapping and thus correlated. To accommodate this, I employed an oblique component rotation with MSC which allows components to be nonorthogonal. These results were compared to results from oblique PCA and from CTS which is constrained to produce orthogonal surfaces. The agreement for the most important component was good in both cases. It could not be expected to be perfect for additional components given this orthogonality constraint. The second oblique PCA component, however, was similar to the second MSC component. The orthogonality constraint is a methodological limitation of CTS, although it has been suggested that it would be useful to try various rotations of the canonical axes too (Cliff and Krus 1976; Wartenberg 1985a). For the technique proposed in the present paper, orthogonal results would have been less informative, as we know that the underlying variables (at least in the simulations) were not independent. In addition, the MSC can only describe data relationships in terms of linear composites. Correlated factors are often derived to depict nonlinear features of data distributions (e.g., quadratic surfaces). This is a second reason to rotate obliquely, but does not apply to canonical trend surface techniques, which have nonlinear terms in them. Complex nonlinear features, however, still may not be resolved adequately.

Finally, I note that with this method both patches and trends were recovered. In exploratory analyses, the investigator often does not know what type of structure to look for. CTS is restricted to large-scale trends. Nonspatial techniques look at

multivariate structure, which the investigator hopes will correspond to spatial structure. Traditional spatial techniques look at univariate structure, which investigators hope will generalize to include many variables. The method proposed here combines all these aspects to give a unified, overall picture. It is left for further study to determine if this desire for generality has obscured resolution so much as to render this approach uninteresting in real data analytic situations.

APPENDIX

The Relation between Ordinary Least Squares, Generalized Least Squares, Principal Components Analysis, and Multivariate Spatial Correlation

Consider the standard linear model

$$\mathbf{Y} = \beta\mathbf{X} + \epsilon, \tag{A1}$$

where $\epsilon \approx N(0, \sigma^2\mathbf{I})$.

In regression analysis by ordinary least squares methods (OLS), we estimate β as

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}. \tag{A2}$$

In situations where there is covariance among the error terms, the standard linear model is

$$\mathbf{Y} = \beta\mathbf{X} + \epsilon, \tag{A3}$$

where $\epsilon \approx N(0, \sigma^2\mathbf{V})$.

In this case, to estimate β, we use generalized least squares methods (GLS):

$$\hat{\beta} = (\mathbf{X}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{V}^{-1}\mathbf{Y}. \tag{A4}$$

By analogy, for principal components analysis (PCA), we assume a standard model:

$$\mathbf{X} = \mathbf{Z}\mathbf{F}^t + \epsilon, \tag{A5}$$

where $\epsilon \approx N(0, \sigma^2\mathbf{I})$.

We estimate $\mathbf{F}$ by taking the eigenstructure of $\mathbf{R}$,

$$\mathbf{R} = \mathbf{X}^t\mathbf{X}. \tag{A6}$$

For generalized principal components analysis, (GPCA), I propose the following model:

$$\mathbf{X} = \mathbf{Z}\mathbf{F}^t + \epsilon, \tag{A7}$$

where $\epsilon \approx N(0, \sigma^2\mathbf{V})$.

We can estimate $\mathbf{F}$ by taking the eigenstructure of $\mathbf{M}$,

$$\mathbf{M} = \mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}. \tag{A8}$$

Note that in this derivation we do not have an arbitrary weight matrix $\mathbf{W}$ as in Moran's I derivation. Rather, I use a weight matrix called $\mathbf{V}^{-1}$, as it is typically

referred to. **V** is the variance-covariance matrix of the errors, ϵ. For a particular error model, $\mathbf{V}^{-1}$ can be set equal to some function of the identity matrix, **I**, a binary neighborhood or connectivity matrix, **C**, and a scaling factor, ρ. For example, in the conditional autoregressive model (Cliff and Ord 1981, p. 148),

$$\mathbf{V}^{-1} = (\mathbf{I} - \rho\mathbf{C}). \tag{A9}$$

Then in GLS,

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'(\mathbf{I} - \rho\mathbf{C})\mathbf{X})^{-1}(\mathbf{X}'(\mathbf{I} - \rho\mathbf{C})\mathbf{Y}) \\ &= (\mathbf{X}'(\mathbf{I} - \rho\mathbf{C})\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y} - \rho\mathbf{X}'\mathbf{C}\mathbf{Y}).\end{aligned} \tag{A10}$$

After appropriate normalizations, the rightmost term of the equation is the OLS solution term minus a term for spatial covariance. Similarly in GPCA,

$$\mathbf{M} = \mathbf{X}'(\mathbf{I} - \rho\mathbf{C})\mathbf{X} = \mathbf{X}'\mathbf{X} - \rho\mathbf{X}'\mathbf{C}\mathbf{X}. \tag{A11}$$

The first term on the right of the equation is the PCA solution and the second is the MSC solution. Further development of this approach will be presented elsewhere.

LITERATURE CITED

Cattell, R. B. (1978). *The Scientific Use of Factor Analysis*. New York: Plenum Press.

Cavalli-Sforza, L. L., and W. F. Bodmer (1971). *The Genetics of Human Populations*. San Francisco: W. H. Freeman.

Cliff, A. D., and J. K. Ord (1981). *Spatial Processes: Models and Applications*. London: Pion.

Cliff, N., and D. J. Krus (1976). "Interpretation of Canonical Analysis: Rotated vs. Unrotated Solutions." *Psychometrika*, 41, 35–42.

Crain, I. K., and K. Bhattacharyya (1967). "Treatment of Non-Equispaced Two-Dimensional Data with a Digital Computer." *Geoexploration*, 5, 173–94.

Dougenik, J. A., and D. E. Sheehan (1979). *SYMAP User's Reference Manual. Version 5.20*. Cambridge, Mass.: Laboratory for Computer Graphics and Spatial Analysis, Harvard University Graduate School of Design.

Geary, R. C. (1954). "The Contiguity Ratio and Statistical Mapping." *The Incorporated Statistician*, 5, 115–45.

Griffith, D. A. (1978). "A Spatially Adjusted ANOVA Model." *Geographical Analysis*, 10, 296–301.

Harris, C. W., and H. F. Kaiser (1964). "Oblique Factor Analytic Solutions by Orthogonal Transformations," *Psychometrika*, 29, 347–62.

Hubert, L. J., R. G. Golledge, and C. M. Costanzo (1981). "Generalized Procedures for Evaluating Spatial Autocorrelation." *Geographical Analysis*, 13, 224–33.

Imbrie, J., and N. G. Kipp (1971). "A New Micropaleontological Method for Quantitative Paleoclimatology: Application to a Late Pleistocene Caribbean Core." In *The Late Cenozoic Glacial Ages*, edited by K. K. Turekian, pp. 71–181. New Haven: Yale University Press.

Kipp, N. G. (1976). "New Transfer Function for Estimating Past Sea-Surface Conditions from Sea-Bed Distribution of Planktonic Foraminiferal Assemblages in the North Atlantic." In *Investigation of Late Quaternary Paleoceanography and Paleoclimatology*, edited by R. M. Cline and J. D. Hays, pp. 3–41. Geol. Soc. Amer. Memoir 145.

Klauber, M. R. (1975). "Space-Time Clustering Tests for More than Two Samples." *Biometrics*, 31, 719–26.

Lebart, L. (1969). "Analyse Statistique de la Contiguite." *Publication Institut Statistique de L'Universitate Paris*, 18, 81–112.

Mantel, N. (1967). "The Detection of Disease Clustering and a Generalized Regression Approach," *Cancer Research*, 27, 209–20.

Menozzi, P., A. Piazza, and L. Cavalli-Sforza (1978). "Synthetic Maps of Human Gene Frequencies in Europeans." *Science*, 201, 786–92.

Moran, P. A. P. (1948). "The Interpretation of Statistical Maps." *Journal of the Royal Statistical Society*, Series B, 10, 245–51.

_____. (1950). "Notes on Continuous Stochastic Phenomena." *Biometrika*, 37, 17–23.

Morrison, D. F. (1976). *Multivariate Statistical Methods*. Second edition. New York: McGraw-Hill.

Sokal, R. R., and P. Menozzi (1982). "Spatial Autocorrelations of HLA Frequencies in Europe Support Demic Diffusion of Early Farmers." *American Naturalist*, 119, 1–17.

Wartenberg, D. E. (1985a). "Canonical Trend Surface Analysis: A Method for Describing Geographic Pattern." *Systematic Zoology*, in press.

_____. (1985b). "Spatial Autocorrelation as a Criterion for Retaining Factors in Ordinations of Geographic Data." *Mathematical Geology*, 17, 665–82.