Journal of Classification 7:53-75 (1990)

Approximate Analysis of Variance of Spatially Autocorrelated Regional Data

Pierre Legendre
Université de Montréal

Neal L. Oden
State University of New York at Stony Brook

Robert R. Sokal
State University of New York at Stony Brook

Alain Vaudor
Université de Montréal

Junhyong Kim
State University of New York at Stony Brook

This work was supported by NSERC grant no. A7738 to Pierre Legendre and by grant BSR 8614384 from the National Science Foundation to Robert R. Sokal. This is contribution No. 366 of the Groupe d'Ecologie des Eaux douces, Université de Montréal, and contribution No. 727 in Ecology and Evolution from the State University of New York at Stony Brook.

Authors' addresses: Pierre Legendre and Alain Vaudor, Département de Sciences biologiques, Université de Montréal, C.P. 6128, Succursale A, Montréal, Québec, Canada H3C 3J7; Neal L. Oden, Department of Community and Preventive Medicine, Health Sciences Center, State University of New York at Stony Brook, Stony Brook, New York 11794-8036, U.S.A.; Robert R. Sokal and Junhyong Kim, Department of Ecology and Evolution, State University of New York at Stony Brook, Stony Brook, New York 11794-5245, U.S.A.

Abstract: The classical method for analysis of variance of data divided in geographic regions is impaired if the data are spatially autocorrelated within regions, because the condition of independence of the observations is not met. Positive autocorrelation reduces within-group variability, thus artificially increasing the relative amount of among-group variance. Negative autocorrelation may produce the opposite effect. This difficulty can be viewed as a loss of an unknown number of degrees of freedom. Such problems can be found in population genetics, in ecology and in other branches of biology, as well as in economics, epidemiology, geography, geology, marketing, political science, and sociology. A computer-intensive method has been developed to overcome this problem in certain cases. It is based on the computation of pooled within-group sums of squares for sampled permutations of internally connected areas on a map. The paper presents the theory, the algorithms, and results obtained using this method. A computer program, written in PASCAL, is available.

Résumé: Cet article présente une solution au problème de l'analyse de variance, pour certains cas où la variable à analyser est spatialement autocorrélée alors que le critère de classification représente des sous-régions connexes du territoire à l'étude. On sait que les méthodes classiques d'analyse de variance ne sont pas applicables dans ce type de situation puisque la condition d'indépendance des échantillons n'est pas respectée; l'autocorrélation positive réduit la variabilité intragroupe, si bien que la quantité relative de variabilité intergroupe s'en trouve artificiellement augmentée. Cette situation correspond en réalité à une vaste catégorie de problèmes en génétique des populations, en écologie et dans d'autres branches de la biologie, ainsi qu'en épidémiologie, en géographie, en géologie, en science économique, en science politique et en sociologie. Ce nouveau test appartient à la famille des tests par permutation. Nous calculons la somme des dispersions intragroupes et testons contre une distribution de référence obtenue en permutant les régions géographiques un grand nombre de fois sur la carte. La véritable difficulté de ce test est d'ordre algorithmique, puisqu'il n'est pas facile de permuter des régions sur une carte, de façon à ce que chaque groupe demeure connexe, et que la carte permutée occupe le même espace total que la carte d'origine. Cet article présente la théorie, les algorithmes, ainsi que des résultats obtenus par cette méthode. Un programme écrit en PASCAL est disponible.

Keywords: Analysis of variance; Choropleth map; Ecology; Genetics; Geography; Permutation test; Spatial autocorrelation.

1. Introduction

Let us consider a common problem. We are studying a variable at various locations in space and we want to test whether predetermined geographic areas are significantly different in terms of the means of this variable. We assume, under the null hypothesis of no difference, that the means do not depend on the areas. This situation presents a problem common in such fields

as ecology, economics, epidemiology, genetics, geography, geology, marketing, political science, and sociology, when dealing with data associated with maps.

The problem as stated above is clearly one of single classification analysis of variance (hereafter referred to as a one-way ANOVA). It is well known, however, that spatial autocorrelation within areas, as is likely to be found in geographically-based data, disturbs the tests of significance used in analysis of variance and its nonparametric equivalents. The reference distributions used in these tests are valid only if the observations arise by adding independent and identically distributed error terms to means that may or may not depend on the areas. Autocorrelated data do not fit this model (Griffith 1978, 1987; Cliff and Ord 1981; Millard et al. 1985). The problem of autocorrelated data can be viewed from another perspective: if one knows the shape of the autocorrelation function and the values the variable takes at some of the points, then one can predict with a given error the values at other points in space. Observing the actual values of the variable at these other geographical locations can only refine our knowledge, since the set of values the variable may take is restricted by our previous knowledge of the values at other space locations; as a consequence, each new observation does not bring with it one full degree of freedom. On the other hand, it is difficult to establish what fraction of a degree of freedom each new observation represents. So, the problem of determining the appropriate null reference distribution remains unsolved. When the effect of spatial autocorrelation is not taken into account, the test becomes too liberal in the case of positive autocorrelation. That is, differences among groups that are not truly different are too often declared to be significant. The probability of a Type I error is larger than its assumed α value. Negative autocorrelation will produce the opposite effect.

Although various approaches may be used to relate geographically-based generating processes to the spatial distribution of variables (see Discussion), these approaches do not exhaust the problem, and an ANOVA can still be seen as the test of choice for examining potential differences in means among geographic areas. Carrying out an ANOVA requires, of course, that one be able to assess correctly the statistical significance of the result. This paper proposes such a procedure, based on a computer-intensive generation of a reference distribution from the actual data. Since it is a permutational method, it differs from the parametric approach of Griffith (1978), for instance. It also differs from the permutation approach of Edgington (1987) in that we preserve the autocorrelation structure of both the variable and the classification criterion.

However, as we stress again below, we advocate this method unreservedly only when the localities are roughly on the nodes of a regular lattice. Use of the method in any other case must be accompanied by

simulations validating the method for that application. We give an example of this below.

2. A Proposed Solution

Suppose that we have observed data values at n localities, and suppose that the localities are spread evenly over the study area. Suppose also that the localities have been exhaustively divided into k nonoverlapping geographic areas $(A_1, \ldots, A_k)$ according to some criterion that is independent of the data values. If the geographic areas are assigned the means of their data values, the resulting map is called a choropleth map in geographical research (Muehrcke 1978; Thrower 1972). We may represent the data as values x_{ij} in a table with k columns, the j-th column with n_j values, and $\sum_{j=1}^{k} n_j = n$, as in one-way analysis of variance. To give an example, suppose the x_{ij} represent gene frequencies in n cities, and $(A_1, \ldots, A_k)$ represent k countries, where $k < n$, into which these cities fall. We wish to decide if we can believe the null hypothesis

H_0: The means of the variates in the areas are the same.

The above situation would be a classical analysis-of-variance problem, were it not for the fact that the x-values of the variable are spatially autocorrelated, but in an unknown way. As pointed out earlier, spatial autocorrelation breaks the classical ANOVA assumption of independence of the error terms, and necessitates an alternative technique.

If we understood the way in which the data are spatially autocorrelated under H_0, we would be able to simulate sets of data values having this spatial autocorrelation, and use them to test the null hypothesis. One possible solution to our problem would be to estimate the spatial autocorrelation function, and use this in the simulation.

Let us suppose, however, that the spatial patterns we deal with are quite complex, and that we have low confidence in our ability to estimate the spatial relations between the localities correctly. Instead, it seems likely to us that it is easier to generate geographic areas that share the properties of the observed areas; by repeating the process a number of times, we can obtain a distribution of the statistic (below) against which the actual map can be tested. Accordingly, we take the data values as fixed, and make the following mimicking assumption.

Assumption: The observed set of shapes can be viewed as a single realization from the ensemble of sets of shapes that can be generated by one of our

computer algorithms. The shapes of the geographic areas $(A_1, \ldots, A_k)$ are therefore acceptably mimicked by a computer algorithm.

To obey the mimicking assumption, the general characteristics of the shapes in the ensemble must be determined. Two computer algorithms are presented in the Appendix. These produce sets of simulated areas (called pseudoareas) that are more or less compact and in which localities are geographically contiguous within each area. These specifications should correspond to most situations found in the field. Implementing geographic contiguity presents a difficult problem in computer algorithms. We overcome this problem by representing the geography as graphs or networks. Graphs that connect nearby points can be used to represent geographical contiguity as edge links between the sampling points. All points that are linked graphically will also be geographically contiguous. We require furthermore that each pseudoarea produced contains the same number of sampling localities as the original area. Therefore, the requirement that the localities be evenly spaced will preserve not only sample size, but also the geographical size of the area. The algorithms presented in the Appendix attempt to satisfy these requirements. Both algorithms work by exhaustively partitioning the graph network into connected subgraphs that contain the same number of localities as the original areas. The connectedness assures that the pseudoareas contain geographically contiguous points. Whether the pseudoareas mimic relatively compact geographical areas is discussed below.

In addition to the requirement that the localities be evenly spaced, we also stipulate that the density of graphical connections should be approximately even over the study area. That is, we wish the incidence variance (i.e., the variance of the number of edges per node of the graph) to be low. In a simulation reported below, when this stipulation was not met, certain pseudoareas were differentially attracted to various parts of the study area. Various graphical connection schemes will fulfill the incidence requirement. In particular, nearest-neighbor graphs have a low incidence variance; in simulations that we performed on randomly located points, the incidence variance of nearest-neighbor graphs was about one fourth that of the corresponding Delaunay triangulations. The incidence variance of graphs based on chess moves (rook's connections, king's connections, etc.) is nearly zero, since only marginal points receive a smaller number of edges.

The test statistic we use is the classical pooled sum of squares within areas:

$$SSW = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$

where k is the number of groups, n_j is the number of localities in group j, and $\bar{x}_j = (1/n_j) \sum_{i=1}^{n_j} x_{ij}$ is the mean value of the variable in group j.

SSW is calculated for the observed geographic areas, and again for each realization of a pseudomap (which is a map of pseudoareas). We reject the null hypothesis if the observed statistic is small relative to the distribution given by the pseudoareas. Because the number of localities per area is preserved in each realization of a pseudomap and the total sum of squares is a constant over all permutations, this decision is equivalent to rejecting the null hypothesis when the among-areas sum of squares (*SSB*) is too large, where

$$SSB = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2$$

and

$$\bar{\bar{x}} = (1/n) \sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}$$

is the overall mean value of the variable over the n localities.

The validity of this test depends upon the validity of the mimicking assumption. In general, we will never know if this assumption is true. Indeed, the best we can usually hope for is that it is approximately true. Therefore the test we present here cannot be viewed as exact, but only as approximate. In some cases, even the approximation will be bad.

One should always investigate the validity of the mimicking assumption when using this test. One way to do this is to compare statistics on the shapes of the observed areas to the same statistics on the population of pseudoareas. We have found the *set diameter* most useful. It is defined as the largest geographic interpoint distance between any pair of localities in the area or pseudoarea, measured along the earth's curvature.

In one of the applications presented below (Section 4), we also report, for each observed area, its position in the corresponding distribution of pseudoarea set diameters. A decision as to the validity of the mimicking assumption can be reached as follows. If we compare the position of the observed area's set diameter to the distribution of this same statistic in the population of pseudoareas, we can compute the probability of obtaining, among the pseudoareas, a result as small as or smaller than the actual value. We may expect this probability to follow approximately a uniform (flat) distribution on the [0, 1] interval if there is no bias in the pseudoarea generation process, that is, if the mimicking assumption is met. Considering the probabilities computed for the various areas in the problem under study, we may test the

correctness of the uniform distribution hypothesis using a Kolmogorov-Smirnov test of goodness-of-fit when the number of groups allows ($k \geq 4$).

In addition to a test of difference of means, it is also possible to investigate each geographic area separately. That is, we can see if its individual sum of squares within group j

$$SSW_j = \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$

is unusual when compared to the SSW_j values of the pseudoareas corresponding to that area. If the probability of the observed value of this quantity is low, the geographic area in question is more internally homogeneous than one can expect under the pseudoarea model. Since the various SSW_j values are correlated, one should use the Bonferroni method (Cooper 1968; Miller 1977) or related techniques, when investigating the homogeneity of sets of individual areas in this way, in order to allow for simultaneous inference.

An alternative way to motivate this analysis is to view the areas (countries, in our running example) as fixed and to assert that the data arise by discretely sampling a single realization of a continuous, stationary, isotropic, weakly autocorrelated stochastic process at the study locality. If we had a different realization, we would of course observe different values at the study localities; therefore, the value observed at any single study location has a sampling distribution. If, however, the assumption holds that we are sampling the stationary process described above, then, for any given cluster of localities, the distribution of *SSW* values calculated on those localities remains the same under any rigid motion (including mirror-imaging) of the localities that still leaves them in the study area. A pseudoarea is a locality cluster that is approximately like the observed cluster (same area, same number of localities, roughly the same shape), so that if we are indeed sampling a stationary process as described above, the *SSW* of a pseudoarea has approximately the same distribution as the observed *SSW*. Drawing many pseudoareas will build up a rough empirical distribution for the *SSW* statistic. The "weakly autocorrelated" part of the assumption allows successively drawn pseudoareas to be nearly independent, thus giving a fairly dependable idea of the distribution. In this case the mimicking assumption is recast as an assumption that the pseudoareas are enough like the observed area to be counted as approximations of it. That is, the observed area is assumed to be an observation from the distribution giving rise to the pseudoareas. The various tests of the mimicking assumption then serve to test this new version of the assumption.

The researcher must decide for each application whether either of these motivations is believable before deciding whether to use this approximate

analysis of variance method.

A conversational computer program written in PASCAL is available from P. Legendre and A. Vaudor to perform the computations for the above method, which we term the *contiguity-constrained permutational ANOVA* (acronym COCOPAN). The program contains both the ring and the random tree algorithms described in the Appendix.

3. Properties of Pseudoareas

3.1 Shape Statistics

What shape do pseudoareas receive from the algorithms described in the Appendix? A regular grid of points was constructed, bearing 18 rows and 20 columns, and the points were connected using king's connections (horizontal, vertical, and diagonal links). Groups of 10, 20, . . . ,80 points each (8 groups, a total of 360 points) were defined on this grid. After permuting the areas 300 times, the set diameter statistic was computed for the pseudoareas (Table 1). These evenly spaced locality points produced slightly elongated pseudoareas, with average set diameter 1.4 to 1.7 times the length of the minimum possible set diameter, using the ring algorithm, and 1.7 to 2.0 times for the random tree algorithm (shapes slightly less compact). So, the algorithms are appropriate for mimicking compact geographic areas, while they may not be for very elongated or dendritic-shaped areas.

3.2 Where do Pseudoareas Fall?

This section studies where the pseudoareas fall on the map, under different patterns of connections. This question is another aspect of the problem of randomness of the pseudoareas constructed by our algorithms; it concerns their location on the map instead of their shape.

For this study, we constructed a regular grid of 100 localities (10×10) on which a network of connections was defined (below). The localities were initially divided into three groups of, respectively, 10, 30 and 60 localities, and we permuted the map 1000 times, using the ring algorithm. If localities were randomly and independently assigned to pseudoareas, we would expect each locality to be included the same number of times in the pseudoarea with 10 localities, and similarly for the 30 and 60 locality pseudoareas.

In a first permutation experiment, the points were connected with irregular density, as follows. The four bottom rows of the 100-point grid were connected using the king's connection pattern (horizontal, vertical, and diagonal connections). This connection pattern assures near-neighbor linkages in all possible directions, which assumes that the area can be viewed as covered

TABLE 1

Set Diameter Statistics over Simulated Pseudoareas

n_j in group	Min[1]	Ring algorithm Mean[2]	Ring algorithm Ratio[3]	Random tree algorithm Mean[2]	Random tree algorithm Ratio[3]
10	3.161	4.352	1.38	5.400	1.71
20	4.470	6.564	1.47	8.996	2.01
30	5.826	8.707	1.49	11.425	1.96
40	6.702	10.880	1.62	13.463	2.01
50	7.609	12.327	1.62	14.979	1.97
60	8.536	13.755	1.61	16.445	1.93
70	9.214	15.353	1.67	17.296	1.88
80	9.836	16.955	1.72	18.107	1.84

[1] Minimum set diameter for the same number of points (n_j) as in the group. Unit = length of a horizontal or vertical edge between adjacent points of the grid.

[2] Mean over 300 permutations, regular grid (18 x 20).

[3] Ratio = Mean/Min.

by a collection of connected rectangular tiles. Patterns based on other chess moves, employed below, are less densely connected. The upper section of the grid was divided vertically in the middle, and the right-hand part was connected using the rook's connection pattern (horizontal and vertical connections), while the left-hand part was connected by horizontal links only. After 1000 permutations, the frequency of occurrence of each locality in each of the three groups (10, 30 and 60 localities) was computed, and an analysis of variance showed that there were significant differences ($p < 10^{-4}$) in the allocation of the three groups to the various regions of the map. Parametric *a posteriori* contrasts tests (SNK and Tukey: Nie et al. 1975, section 22.3.3) showed that the mean frequency of occurrence was different for all pairs of regions of the map, and this for each group (10, 30 or 60 localities respectively). Mapping the frequencies of occurrence for each of the groups of localities showed that the 60-point group tended to occupy the horizontally-

connected area of the map (upper left) far more often than expected, while the 10-point group occupied that same area far less frequently than expected. The 30-point group occupied the king's-connected area (lower part of the map) far more often than expected.

To see whether this effect was really caused by the connection pattern, rather than by some other artifact of the program, a second permutation experiment was conducted in which the points in the grid were all connected by the king's connection pattern (horizontal, vertical, and diagonal connections). After 1000 permutations, the frequency of occurrence of each locality in each of the three groups was computed as above; an ANOVA showed that there were no significant differences in the allocation of the three groups to the various regions of the map.

These results show on the one hand that the COCOPAN method may not be appropriate in cases where the localities are connected in a pattern with high incidence variance; in this example the smaller groups of localities tend to avoid the more weakly connected regions of the map. On the other hand, when the connecting pattern was uniform, there was no tendency for some pseudoareas to be found repeatedly in some regions of the pseudomap, or to avoid other regions.

4. An Example: Ecological Application (Lattice Data)

Legendre and Legendre (1984) have studied the postglacial dispersal of freshwater fishes in the Québec peninsula. Part of their data set is used here to test if the number of fish species is related to the geomorphology of the area. The territory under study consists of 64 one-degree-square units (lattice data), located in the western part of Québec, adjacent to James Bay (Figure 1); each unit of territory is about 7000 km^2. It seemed possible that the number of habitat types available for fishes depended to a certain extent on the nature of surface deposits, which may influence, for instance, the primary production of the lakes and rivers, through the types and amounts of dissolved minerals. Three kinds of surface deposits have been recognized in the area: the Tyrrell marine transgression, glacial deposits only, and glacial lake transgression.

There is a north-to-south gradient in species number, common in these northern latitudes. So as not to confound the effect of latitude with that of geomorphology, the effect of the former was eliminated by linear regression. A parametric analysis of variance was first computed on the residual data, using the ONEWAY subprogram of the SPSS package. Significant differences were found among the groups at probability level 0.0009. Parametric *a posteriori* contrasts tests (SNK and Tukey) showed further that although the marine (A in Figure 1) and the glacial lake (C) transgression

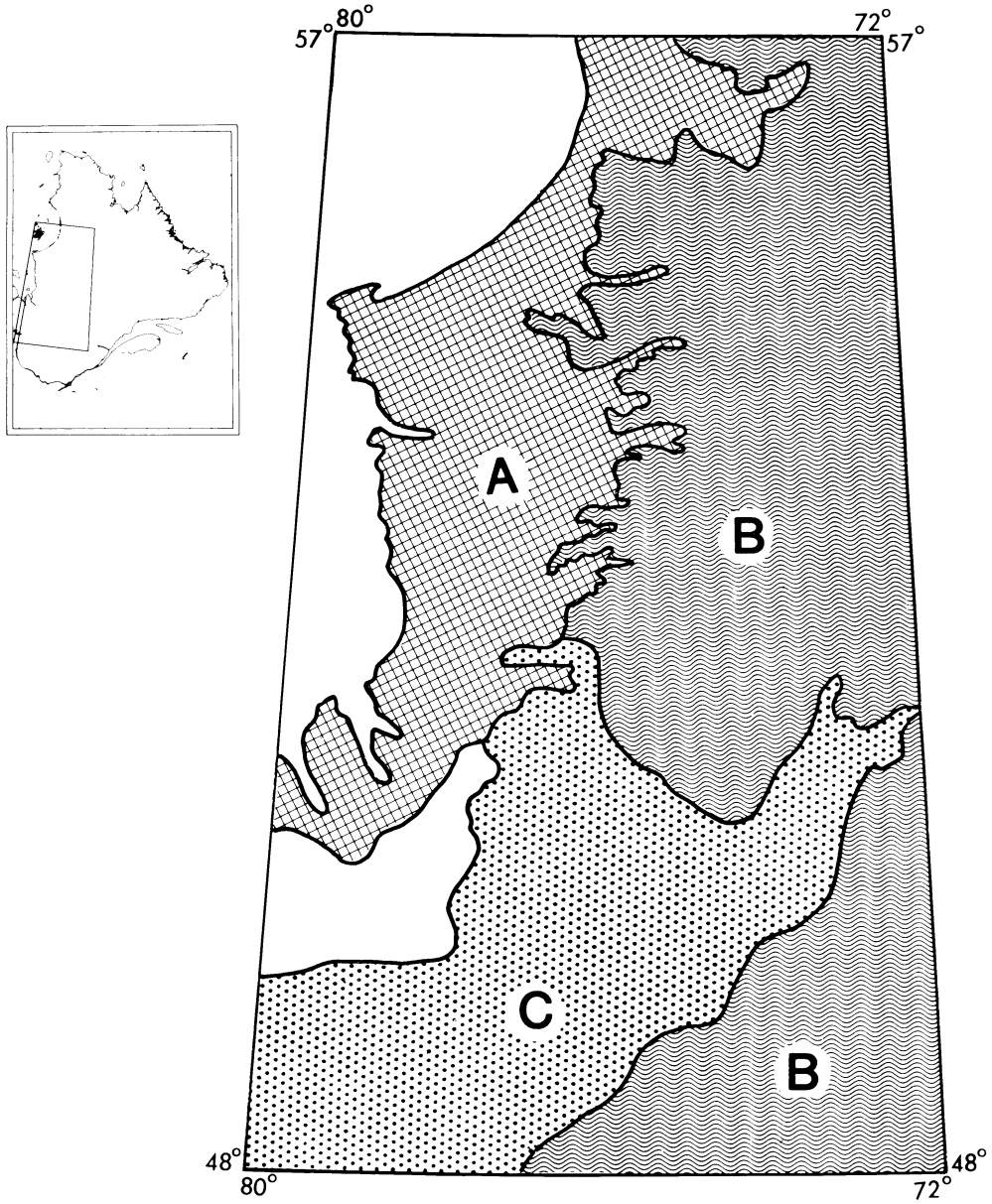

Figure 1. Geomorphological division of the territory, according to the Glacial Map of Canada. **A:** maximum extension of the Tyrrell marine transgression. **B:** Glacial deposits only (Wisconsinian ice sheet). **C:** maximum extension of the Ojibway-Barlow glacial lake.

TABLE 2

Set Diameter Statistics over Pseudoareas, Fish Example

Group name	n_j in group	LT[1]	EQ[1]	GT[1]	Prob LT+EQ[2]
A	19	157	5	88	64.8%
B	29	94	5	151	39.6%
C	16	202	3	45	82.0%

[1] LT = number of pseudoareas with set diameter smaller than the actual value; EQ = equal; GT = greater than. Number of permutations = 250.

[2] Probability of a result as small or smaller.

zones were not significantly different from one another, both displayed significant differences from the third zone (B) where only glacial deposits are found.

Since the detrended data still displayed significant patchiness, when subjected to spatial autocorrelation analysis (Cliff and Ord 1981), they were re-analyzed by COCOPAN; the territory units were connected using king's connections for maximal near-neighbor connectivity. Sixteen of the 250 permuted maps were found to yield smaller within-group sums of squares than the observed data (none were equal), giving a probability that H_0 is true equal to 0.064. This probability level does not allow one to reject H_0 at the 0.05 probability level and so we find no convincing evidence that fish diversity, as measured by the number of species, varies with the "surface deposits" classification criterion in this territory as mapped in Figure 1.

In Table 2 we report results on the position of the shape statistics of the observed areas in the distribution of such statistics for the pseudoareas. Since none of the three areas is aberrant, we tend to accept the mimicking assumption.

Here, then, is a case where spatial autocorrelation in the data would have led one to take the wrong statistical decision after an ordinary analysis of variance, because ANOVA is too liberal when the data are positively autocorrelated, as discussed in the Introduction. COCOPAN, on the contrary, is

more conservative. This difference may be important in a context where results from statistical tests may lead to management decisions (Millard et al. 1985).

5. What to do if the Lattice Condition is not Met

The computer algorithms presented in the Appendix ensure that the number of localities in each of the geographic areas is preserved in every realization of a pseudomap. Variation in the density of localities across the study site prevents the original size of the areas from being preserved. Areas with dense localities will give rise to larger sized pseudoareas (as measured for instance in square kilometers), when projected in more loosely sampled parts of the map. If the data are also spatially autocorrelated, this may influence the *SSW* statistic. In this situation, COCOPAN needs to be applied with caution. However, auxiliary tests may suggest that the method is still reliable in specific applications. An example using this more complex (but commonplace) data structure follows.

6. Gene Frequency Differentiation Among Human Linguistic Groups (Nonlattice Data)

Sokal et al. (1989) have studied whether differences in gene frequencies are associated with language-family areas in human populations in Europe. The data in the study consisted of 69 variables of blood group gene frequencies and cranial measurements sampled from various geographic locations in Europe. Each genetic system was measured at a different set of localities over Europe. The location of the samples was not under the control of the investigators, and the sampling density for the variables was very uneven across the European continent. Is the method reliable when the data do not form a regular lattice? To provide an answer to that question, a simulation study was carried out to assess the degree to which the unevenness of the sampling locations affects the permutation method developed here.

The simulation study used the observed geographic locations and their graph links (next paragraph) and generated data according to three different models, in order to assess the method. The first model, called the null model, assumed that all localities were independent and there were no differences among groups in the means of the variable; an analysis of variance is not expected to find any significant difference among these groups. The second model, called the language model, assumed that there was no spatial autocorrelation within language areas in the simulated variables, but that some of the language areas had different means; an analysis of variance is expected to find these differences. The third model, called the geographic model,

assumed that there was autocorrelation between the locations but no difference in means across language areas; an investigation of differences among means should yield nonsignificant results. The investigators felt that if COCOPAN gave results expected under the various models, the degree of uneveness in the sampling pattern of the data could be ignored when evaluating the actual genetic data in the study.

The localities of two genetic systems were chosen as representative of the sampling pattern of the total data set. They were: ABO gene frequency data, which had the largest number (870) of localities, and haptoglobin gene frequency data, which had about the modal number of sample localities (175). The locality sampling points based on the ABO gene frequency data were tested only against the geographic model because of the heavy computational load incurred when working with this dataset. The graph links on these two locality sets were established by a Delaunay triangulation (Brassel and Reif 1979; Ripley 1981, section 4.3; Upton and Fingleton 1985, section 1.7) of the original geographic locations. The original partitionings of localities into language-family areas were used in the simulation study.

The simulated data for the null model were generated by sampling from a normal distribution of mean 0 and variance 1. We created 50 such surfaces on the geographic locations of the haptoglobin data. COCOPAN was applied to each of these simulated surfaces in turn. The results were summarized by determining whether a dataset was found to lead to significant differences among group means at the 0.05 probability level. All but two replicates from the null model were found to be nonsignificant, which is well within the expected number of nonsignificant datasets (obtained from the binomial probability), so that the method performed satisfactorily.

The language model data were generated, again for the haptoglobin data points, by sampling from the normal distribution, specified as $N(0,1)$, for all language family areas but two. The two exceptions were sampled from normal distributions specified as $N(1,1)$ and $N(3,1)$; this was done in two different ways. In a first set of 50 simulations, the two areas with the largest number of localities were sampled from the normal distributions with offset means. In a second set of 50 surfaces, intermediate size areas were chosen. All of the replicates in the language model were found to differ significantly in their means, regardless of which of the two treatments was applied. This result was also as expected, since the assigned parametric means differ by construction.

The geographic model data were generated by assuming that there was a "lattice" of 100 000 points underlying the continent of Europe. Each lattice point had a value sampled from the normal distribution of $N(0,1)$. The value for each observed locality was obtained by summing all the values of the underlying lattice that were covered by a window centered upon the

geographic location of the observed locality. Neighboring localities will be autocorrelated due to overlap of the windows; the amount of autocorrelation depends on the size of the window. Four window sizes were used in the simulation test.

Each treatment was replicated 50 times for the haptoglobin data points, except for the largest window size, where computational load limited replications to 20; each treatment for the ABO data points was replicated 20 times. The number of significant replicates per set of different autocorrelation strengths were 3, 4, 4, and 0 for the haptoglobin data points, as the strength (i.e., window size) increased; for the ABO data points, the number of significant replicates was 0, 3, 4 and 3 per set, respectively. Since there was no difference in means of the parameters of the simulation, we expected that no significant difference should be found among group means of the simulated data sets. The maximum value of four significant replications found for the haptoglobin as well as for the ABO data points is well within expected 99% confidence limits, based on the binomial distribution.

The least autocorrelated surfaces, among the four data sets generated under the geographic model for the haptoglobin data points, were also subjected to parametric ANOVA for comparison. All 50 replications led to highly significant results. This outcome indicates that contiguity-constrained permutational ANOVA does overcome the problem of autocorrelation in the data while parametric ANOVA does not. Although this simulation study is not complete (e.g., mixed models of language and geography were not tested), we felt that the results gave enough confidence in the method to apply it to our unevenly distributed sample data.

When the original dataset of 69 variables, sampled over Europe, was initially tested using parametric ANOVA, all but 4 variables were found to differ significantly in means over language groups. As suggested at the beginning of the Discussion, the four non-significant variables were immediately eliminated from the study, since adjustment for spatial autocorrelation by our method would probably not have shown them to be significant. When the 65 remaining variables were tested by COCOPAN, the null hypothesis was rejected in only 23 cases (Sokal et al. 1989), which shows again that parametric ANOVA can often give very biased results in the presence of spatial autocorrelation.

The analysis of the European gene-frequency data demonstrated that speakers belonging to different language families differ genetically. These differences are not due to spatial autocorrelation induced by the limited mobility of the speakers (isolation by distance), since the COCOPAN method eliminated the effects of autocorrelation. In the cited study (Sokal et al. 1989), various models bringing about the observed patterns are enumerated and discussed.

It should be noted that our simulation study does not validate the use of the permutational method in all cases of uneven sampling. The tests were done specifically on the sampling patterns of ABO and haptoglobin and therefore only lend confidence to using the method on these and similarly sampled datasets. We recommend that unevenly sampled data be examined individually in a manner similar to the simulation study above.

7. Discussion

We have explained in earlier sections that autocorrelation in a variable can disturb the tests of significance used in the analysis of variance, and we have clearly illustrated this phenomenon in the two examples presented in Sections 4 and 6, where results of parametric ANOVA were compared to those obtained with the contiguity-constrained permutational ANOVA described in this paper. Even in the presence of spatial autocorrelation, one can start by performing standard ANOVA (parametric or non-parametric). If the spatial structure consists of positive autocorrelation for the small distance classes, and if the null hypothesis of ANOVA is accepted, there is probably no need to proceed with any further testing. With positive autocorrelation, ANOVA results that are significant can be rendered nonsignificant with our method, ANOVA being too liberal in that case. So, only when the null hypothesis of ANOVA is rejected should the analysis be repeated using some method, such as the one described in this paper, that takes the spatial structure into account. In the very rare cases of negative autocorrelation, ANOVA will be too conservative.

One might be tempted to use the Mantel test (Mantel 1967) as another way of answering the same question as that addressed by COCOPAN. The test would be performed between (1) an $(n \times n)$ matrix of n localities, containing the differences in values of the variable at the various localities, and (2) a model $(n \times n)$ matrix containing, say, zeroes for all within-group comparisons and ones for all between-group comparisons. The Mantel statistic would measure the degree of agreement between data and model, and this value could be tested for significance against a reference distribution obtained after repeated random permutations of the order of the localities in one of the two matrices. We need to point out that this test would not be equivalent to COCOPAN. In the latter test, we leave the sampling points in their fixed positions on the map, each retaining its value of the variable under study, thus leaving undisturbed their autocorrelation structure. In the Mantel test on the contrary, random permutations of rows and columns destroys the inter-point structure in one of the two matrices (computationally the results are the same whichever matrix is permuted). Either the values of the variable are randomized over the localities, thus destroying their autocorrelation

structure, or it is the group membership which is established at random, so that the connectedness of pseudoareas is destroyed in the permutation process. Used thus, the Mantel test is a nonparametric ANOVA that is appropriate only when there is no spatial autocorrelation within groups. In contrast, COCOPAN is appropriate if one believes that the autocorrelation structure of the variable is a salient property of the problem. Moving the regions around, in the algorithm, is actually equivalent to retaining the initial geographic position of the regions while randomly moving around (i.e., randomizing) selected blocks of localities that retain their autocorrelation structure.

Suppose our data indicate a spatial gradient, with long, narrow areas perpendicular to the direction of steepest ascent. How should such data be analyzed? The surprisingly complex answer to this question is considered below in terms of two separate issues.

The first issue concerns the question of whether the observed gradient should be removed before the analysis of the data, as we did in the example of Section 4. The problem here is that a true systematic change in the underlying expected value of the data (termed here for convenience a "true" gradient) can often be convincingly imitated by highly autocorrelated data with no change in underlying expected value (termed here for convenience a "false" gradient). Indeed, it is this phenomenological similarity between true and false gradients that necessitates the use of special methods with autocorrelated data. If there is prior reason to believe in a true gradient (as would be the case, for example, if we knew, on the basis of other evidence, that the observed data values were driven by an underlying nutritional gradient, or by climate, etc.), then this gradient must be removed before further data analysis with COCOPAN. The reason is that the null hypothesis of COCOPAN holds that there is no underlying change in expected value (the "weak stationarity" assumption of geostatistics). In other words, all variations in the data are caused by autocorrelation; all gradients are false. If we have prior belief in a true gradient, the data are not admissible for study by COCOPAN. We add parenthetically that the removal of a true gradient in this situation is difficult, since, if one admits the possibility of autocorrelation, the true gradient might easily be partially confounded with a false gradient. Special methods are required to disentangle the truth in this type of situation.

The second issue concerns the validity of the mimicking assumption. Suppose we adopt the null hypothesis that the observed gradient is false. We wish to use COCOPAN to test this hypothesis, with the alternative being that the gradient is at least partially true. Note that the areas in the example, being long, narrow, and perpendicular to the direction of the steepest ascent, differ as much as possible from each other in data values. The reason is that each area is as homogeneous as can be, because of the peculiar division of the observed gradient into areas. Either COCOPAN or an ordinary ANOVA

would therefore produce a significant result. Can we now reject the null hypothesis, and conclude that there is at least some true gradient?

The answer to this question clearly reveals the importance of the mimicking assumption. From an intuitive point of view, we might claim that the observed differences are only the result of the shape and orientation of the areas. If these had been constituted more randomly, we might not have found any differences among them. This possibility immediately raises the question of what sort of areas might have been possible besides the ones observed. If, for some reason unrelated to area-specific data values, the only possible areas were those with borders perpendicular to the gradient, then our observed differences are not surprising, and do not support the assertion of true differences between the areas. If, however, under the null hypothesis of no difference between areas, the areas could have arisen any which way, then the observed inter-area differences are surely surprising, and merit rejection of the null hypothesis.

From the point of view of COCOPAN, these meditations translate into questions concerning the validity of the mimicking assumption. Suppose the areas could have arisen any which way. That is, suppose that the observed areas are a sample from the distribution of areas generated by our computer algorithms. The mimicking assumption is now correct. In this case, COCOPAN would produce the correct answer, namely, that the areas differ. When confronted with such peculiar observed areas, the investigator should, however, immediately suspect that the computer algorithms are inappropriate for the areas under investigation. This lack of suitability should be further revealed by the test of the mimicking assumption required with each application of COCOPAN. If an alternative algorithm generated only pseudoareas perpendicular to the gradient, the method would find no significant difference among areas. Which pseudoarea generation method provides an acceptable model for the observed areas depends upon the specific application.

Appendix: Algorithms

The algorithmic problem to solve before permuted maps can be obtained is not simple, since each new permutation of the pseudoareas must contain the same number of observations per group, each pseudoarea must remain connected, there must be the same number of pseudoareas on the map, and the pseudomap must cover the same total surface as the original map. Two algorithms have been developed for achieving these goals, and although they use very different approaches, they produce very similar probabilities when compared over an extensive number of runs. This result is interesting in itself. The first algorithm produces more compact groups, but the second

algorithm is faster with large data sets. As noted above, both methods require initial user-specification of a network connecting all localities.

1. The Ring Algorithm

The first algorithm approaches the problem as follows. Seed points for the pseudoareas are chosen at random among the localities on the map. Then each group is grown in steps, by attaching concentric rings of points around the seed locality (hence the name of the algorithm), following the connecting graph. When pseudoareas meet, growth is no longer possible along their common border and each one has to grow in different directions, as available points permit. When ring growth is no longer possible, another procedure takes over that forces the incomplete groups to grow at the expense of their neighbors. This goes on until all pseudogroups have reached the required number of localities, that is, the same number as in the observed geographic areas they mimick. If this attempt turns out to be too tedious, the incomplete pseudomap is abandoned and the procedure is restarted from the beginning. The step-by-step algorithm goes as follows.

1.1. For all localities, build a connecting network reflecting the graph connection structure of the observed localities, as described in the Methods section. Write it in a file available to the COCOPAN program, in the form of a list of pairs of locality identifiers.

1.2. Compute the pooled sum of squares within observed areas, *SSW*, based on the actual division of the study area into regions. Call this value *VAL*.

1.3. Compute h random permutations of the pseudoareas on the map. Each permutation corresponds to a realization of the null hypothesis. For instance, $h = 250$, or 500, or 1000. The algorithm accomplishes this task as follows:

1.3.1. First step:

(a) Choose a random seed for each of the pseudoareas to be grown on the pseudomap.

(b) All groups grow simultaneously, one ring of points at a time for each group in turn, following the connecting graph. At the end of each cycle, a statistic is computed:

$$\frac{\text{number of points of the group still to be placed}}{\text{total number of points in the group}}$$

and the group that has the highest value of this

statistic is served first during the next cycle. This step is critical to reducing computation time.

(c) The groups that are blocked by not having localities to expand to on the connecting graph are placed on "hold."

1.3.2. Second step:

(a) The blocked groups on hold are forced to grow at the expense of others, one ring of points at a time, insofar as unassigned points are available.

(b) Repeat this step for all groups placed on hold. (The number of repetitions allowed for this operation is set by a parameter of the program.)

(c) Check that every pseudoarea is still connected. If not, discard the map and go to 1.3.1.

1.3.3. Third step:

(a) If only one or a few points are missing for the last pseudoarea to be completed, find the shortest path between the empty spots and the group (pseudoarea) in question, and move all the points along this path in the direction of the empty spots so as to create empty slots adjacent to the group, permitting it to expand.

(b) Check that every pseudoarea is still internally connected. If not, discard the map and go to 1.3.1.

1.3.4. Decision about the map:
Keep this pseudomap if it is complete, or throw it away and start again from the beginning (step 1.3.1).

1.3.5. Statistic:
If it is kept, compute the *SSW* statistic for this pseudomap, which is a realization of the null hypothesis.

[End of the map permutation loop.]

1.4. From the h values of the statistic obtained under H_0, calculate the probability of finding a value as small as, or smaller than *VAL*, if the null hypothesis is true.

2. The Random Tree Algorithm

For large data sets (with more than, say, 500 points), the ring algorithm above has the disadvantage of being slow; its complexity is $0(n^2)$. This disadvantage becomes a real inconvenience when one wishes to produce a large number of maps. For these reasons, an effort has been made to produce

a faster algorithm. We have used this faster algorithm on problems involving more than 1300 points, which is by no means its practical limit.

The problem of partitioning a graph into connected components with a prespecified number of vertices can be viewed as the problem of finding the appropriate edges that must be removed to achieve the partition. In most graphs, the number of edges that must be removed to produce two disconnected components (although each one remains internally connected) is large. However, we can use the fact that if the graph is a tree, the removal of any edge will result in two separate connected components, by definition of a tree graph. Any graph can be reduced to a subgraph that is a tree. For any desired partition, the entire graph may be reduced such that every subgraph is a tree, attached to other subgraphs by only one edge. Our problem can then be approached as that of finding a random reduction of the original graph into a tree, such that the removal of certain edges will result in the desired partition. The step-by-step algorithm is the following.

2.1. and 2.2. Same as 1.1 and 1.2 above.

2.3. Compute h random permutations of the pseudoareas on the map. Each permutation corresponds to a realization of the null hypothesis. For instance, $h = 250$, or 500, or 1000. The algorithm accomplishes this task as follows:

2.3.1. Pick a locality at random, to be used as the starting point for building the randomly reduced tree (called the "random tree" from here on).

2.3.2. Start a random walk over the topology of the original connecting network, with bifurcations, such that the result is a binary tree. The tree will be made of a random selection of certain edges of the original connecting network. This selection may be done in several ways, but we used the simplest method we found, using the following recursive algorithm.

(a) From the starting point, randomly choose one of the localities connected to it and include this locality as the start of the left branch of the tree.

(b) Repeat (a), adding localities until no more can be found.

(c) Trace back the localities until a locality is found that still is connected to at least one not already included in the tree. Add that locality as a right branch of the tree and continue as in (b).

(d) Check that all localities are included in the tree. If not, find the lost ones. To preserve the binary tree

structure, all the vertices must be of degree three or less (i.e., each must have three or fewer localities connected to it). If a lost locality cannot be attached to the random tree without increasing the degree to more than three at the node locality where it is to be attached (the *mother*), a *pseudolocality* is created. This pseudolocality is inserted between the mother and any other locality attached to the mother. The lost locality is then attached to this pseudolocality. Pseudolocalities are never counted in any other procedure and are only used to preserve the binary tree structure.

(e) Continue (b) from the lost locality, and repeat (d) as needed, until all localities have been added to the tree.

[End of the recursive procedure.]

2.3.3. Once all the localities have been added to the binary tree, write tag numbers along the tree from the tip of the leftmost branch. The numbering is done by counting the number of vertices subtending (''descending from'') each vertex, and including that vertex. The pseudolocalities are not included in the count.

2.3.4. Do binary tree search from the tip of the leftmost branch. The tag numbers from step 2.3.3, found at each vertex, are checked against the list of desired group sizes. If a match is found, all of the vertices from that branch are labeled with the name of the corresponding group. The branch containing these vertices is pruned from the tree.

2.3.5. Renumber the remaining binary tree after pruning. Continue with step 2.3.4 until all groups have been fitted, or until no matches are possible. If groups remain unfitted, abort the tree and go back to 2.3.1 to create a new random tree.

2.3.6. Output the assignments of localities to the newly permuted pseudoareas.

2.3.7. Compute the *SSW* statistic for this pseudomap, which is a realization of the null hypothesis.

[End of the map permutation loop.]

2.4. The estimated probabilities are computed as in 1.4, after the desired number of permutations have been completed.

References

BRASSEL, K. E., and REIF, D. (1979), "A Procedure to Generate Thiessen Polygons," *Geographical Analysis, 11*, 289-303.

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

COOPER, D. W. (1968), "The Significance Level in Multiple Tests Made Simultaneously," *Heredity, 23*, 614-617.

EDGINGTON, E. S. (1987), *Randomization Tests*, 2nd Edition, New York: Marcel Dekker.

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

GRIFFITH, D. A. (1987), *Spatial Autocorrelation: A Primer*, Washington: Association of American Geographers.

LEGENDRE, P., and LEGENDRE, V. (1984), "Postglacial Dispersal of Freshwater Fishes in the Québec Peninsula," *Canadian Journal of Fisheries and Aquatic Sciences, 41*, 1781-1802.

MANTEL, N. (1967), "The Detection of Disease Clustering and a Generalized Regression Approach," *Cancer Research, 27*, 209-220.

MILLARD, S. P., YEARSLEY, J. R., and LETTENMAIER, D. P. (1985), "Space-time Correlation and Its Effects on Methods for Detecting Aquatic Ecological Change," *Canadian Journal of Fisheries and Aquatic Sciences, 42*, 1391-1400.

MILLER, R. G., Jr. (1977), "Developments in Multiple Comparisons," *Journal of the American Statistical Association, 72*, 779-788.

MUEHRCKE, P. C. (1978), *Map Use*, Madison, WI: JP Publications.

NIE, N. H., HULL, C. H., JENKINS, J. G., STEINBRENNER, K., and BENT, D. H. (1975), *SPSS — Statistical Package for the Social Sciences*, 2nd Edition, New York: McGraw-Hill.

RIPLEY, B.D. (1981), *Spatial Statistics*, New York: Wiley.

SOKAL, R. R., ODEN, N. L., LEGENDRE, P., FORTIN, M.-J., KIM, J., and VAUDOR, A. (1989), "Genetic Differences Among Language Families in Europe," *American Journal of Physical Anthropology, 79*, 489-502.

THROWER, N. J. W. (1972), *Maps and Man*, Englewood Cliffs, NJ: Prentice Hall.

UPTON, G. J. G., and FINGLETON, B. (1985), *Spatial Data Analysis by Example. Vol. I: Point Pattern and Quantitative Data*, Chichester: Wiley.