A New Statistical Approach to Geographic Variation Analysis

Author(s): K. Ruben Gabriel and Robert R. Sokal

Source: *Systematic Zoology*, Sep., 1969, Vol. 18, No. 3 (Sep., 1969), pp. 259-278

Published by: Taylor & Francis, Ltd. for the Society of Systematic Biologists

Stable URL: https://www.jstor.org/stable/2412323

A NEW STATISTICAL APPROACH TO GEOGRAPHIC VARIATION ANALYSIS

K. Ruben Gabriel and Robert R. Sokal

Abstract

The authors discuss the problems of describing geographic variation data and develop statistical methods for categorizing sets of populations sampled from different localities. The general approach of the simultaneous test procedures, available with a variety of statistical tests and for continuous as well as for categorical data, is employed with these techniques. Geographical regions are defined as sets of connected localities, with connectedness being defined geometrically. Maximal acceptable connected sets of localities (defined as regions) or coarsest acceptable connected partitions of the entire set of localities are found by these procedures. These are illustrated with several examples.

I. Aims and Purposes

The primary aim of geographic variation analysis in biological systematics is the *description* and *summarization* of patterns of variation and covariation of characteristics of organisms that are distributed over an area. Such analyses are generally applied to species populations to study the variation of diverse characters. The most frequently studied characters are morphological, but recently there have also been such studies of biochemical, physiological, behavioral, cytological, immunological, as well as genetic characters.

The basis for studies in geographic variation rests on the existence of populations of comparable organisms at a number of localities in the area under study. Comparisons of these populations are made in terms of one or more observable characters and the analyses relate these comparisons to differences in location.

Data for studies in geographic variation consist of samples taken from the populations at a given number of localities, with a set of characteristics observed for each organism sampled. Summary statistics may be computed for each sample, as, for example, means and standard deviations for single measurable characteristics, correlations for pairs of such variables, etc. A summary of the data would then consist of a list of localities, each accompanied by its set of summary statistics for observed characteristics. In most cases, the information in such a list would be difficult to grasp and much would be gained by plotting the statistics on a map (or several maps) according to their location. Such graphic representation often reveals a good deal about the geographic variation pattern involved.

A step beyond mere description of the pattern of variation of the characteristics of organisms is *categorization*. Usually, one prefers to group together localities that are geographically adjacent and whose populations are similar in their characteristics. This may be desired merely for purposes of simplification and summarization, or for the formal or semi-formal recognition of a population or series of populations in terms of the Linnean system.

The study of patterns of geographic variation will often lead to *causal analysis*. One may attempt to interpret the variational and correlational patterns of a species as adaptations to variation in known environmental factors, such as climatic, topographic, or edaphic variables. Other possible causes of variation may be differences in associated species populations, such as host plants, parasites, predators, etc. Marked and abrupt changes in characters between close localities may be related to abrupt changes in the above factors, to strong barriers to dispersal of the organisms or to secondary zones of intergradation between allopatrically differentiated populations. Unusual patterns of variation may also lead to the discovery of previously unsuspected causal variables. Certain variational patterns may be related to historical factors, such as the movements of populations induced by physiographic factors, or to distributional changes of allied species.

Ultimately, one would like an analysis of geographic variation to provide the tools for *allocation of unknown specimens* to a given population or geographic locality (with a stated probability of success). Thus, given an organism or a sample for which one or more characters have been measured, one would like to be able to know the probability that it came from a particular geographical area.

II. OBSERVATIONS AND ERRORS

Types of data.—The characters with which geographical variation studies may be concerned are of several kinds, as are biological variables in general. It is useful to distinguish *measurement* variables from *categorical* ones (attributes). The former include ratios as well as direct measurements, and for present purposes no distinction will be made between continuous variables proper and meristic variables, such as number of teeth or scales, which range over an appreciable number of possible values. (Meristic variables with a small range, such as number of digits in vertebrate forelimbs, are not explicitly dealt with here.) Categorical variables, on the other hand, are not arranged in a linear quantitative order, but merely allow classification into a number of distinct categories. An example is that of three different color patterns on individuals of a polymorphic species. Geographic variation analysis would investigate geographic differences in the proportions of these patterns. Other examples of categorical observations are mating types, host or food preferences, or morphotypes.

Geographic variation studies may be univariate or multivariate. A *univariate* study considers only one character at a time, irrespective of whether other characters are also being measured and reported. In a *multivariate* study, on the other hand, one investigates not only the geographic variation of each variable separately, but also the covariation of all the variables together. Recent studies have tended to emphasize multivariate techniques (Jolicoeur, 1959; Sokal and Thomas, 1965; Thomas, 1968a, 1968b). Furthermore, attempts at explaining the pattern of variation in terms of underlying causes frequently require a multivariate approach.

Types of error.—The data for studying geographic variation of biological organisms are affected by a number of sources of random variability of organisms and errors of measurement. Inferences based on such data necessarily require statistical methods and will be probabilistic in character, i.e., will involve significance and confidence statements rather than absolute determinations.

The present discussion deals with observations on organisms sampled at a number of localities. The errors involved are therefore of three kinds (1) sampling error due to natural variability of organisms at a given locality, (2) measurement error for each sampled organism, and (3) errors in representation of the entire area of study by means of a particular set of localities. Each of these sources of error will be discussed in turn.

The sampling error due to natural variability of the organisms of any one population at any given locality is the phenetic variance of characteristics of biological organisms. This basic biological "error" may be estimated by the conventional statistical methods from the within-sample variability. The reliability of sample estimates may then be judged by means of their standard errors and by confidence bounds which are computed by well known statistical techniques from the variance estimate and the sample size. Systematic errors may creep in, largely as a function of the times at which the observations were taken. Samples of organisms collected in one locality at different times of the year may yield statistics different not only because of the basic, biological variability, but also because of seasonal phenetic variation. This may result from changes in the genetic composition of the population or from changes in the environment which affect the appearance of the organisms (see Heryford, Kishpaugh and Sokal, 1969, for a recent example). Similarly, there may be annual fluctuations in these statistics for a variety of reasons. This source of error assumes importance when samples are collected in different years. Finally, there may be secular change (over relatively long periods of time) which presumably can be ignored in most studies. The respective magnitude of these sources of variation will depend on the biology of the organisms studied. For example, the relative length of the life-span will determine how a population meets the challenge of seasonally or annually changing environmental factors.

Errors of measurement due to imprecise laboratory techniques are unlikely to be important relative to the basic biological error.

Errors of locality selection; so far only errors *in situ* have been considered, i.e., biological error and measurement error at any given locality. Of no less importance are errors due to the sampling of localities within the area studied. How large a number of localities is needed? And how should they be located? On a lattice? Completely at random? Randomly within regions? The answers will depend on the aims of the study, and on the extent of geographic variability expected.

If one requires precise interpolation to unsampled points of the area one may need a systematic lattice of many locations. If one is content with a general indication of geographical trends or regions, a scattered small number of localities may suffice. If one wants to interpolate blood group frequencies between sampled points one would need a very much denser locality network in Israel, or in Manhattan, than, say, in central China or in the province of Quebec.

These are most weighty questions that have to be carefully considered in any geographical variation study. However, they are essentially preliminary to the statistical analysis of the data and will therefore not be taken up here. The present discussion starts off from a given set of localities and is concerned with inference from data collected at these localities. The situation is somewhat analogous to the distinction between design and analysis of experiments. The analysis is conditional on the given design, and cannot give better results than what is built into the design, and yet it is useful to discuss the analysis per se for *given* designs. This does not detract from the importance of good design of experiments (and of good sampling of localities in a geographical area).

III. STATISTICAL APPROACHES

Questions and methods.—Methods of statistical analysis must be chosen according to stated goals. Whether the question is formalized in a mathematical statistical model, or is merely guiding the exploration of data, there must be some question or questions for which answers are sought.

The present paper does not attempt a general discussion of problems of geographic variation and the related statistical methods. For a review of such methods the reader is referred to Sokal (1965). The discussion here is limited to the question of partitioning an area into regions according to the characters of the organisms inhabiting different localities of the area. This question is closely related to that of fitting a surface (surfaces) which describes the variation of a character (characters) over the area. It is well therefore to review the two questions and their appropriate statistical methods.

Trends versus partitions.—If the characters under study vary continuously over the area considered, a description of the geographic variation by means of trend surfaces is appropriate. A trend surface may be fitted to the observations of each character by least squares and contour lines plotted on a map for easy visualization of the trends. These contour lines estimate the loci of all points with a given value of the character, and are therefore referred to as *isarithmic* lines, or *isophenes*. A first study of geographic variation by such means is due to Marcus and Vandermeer (1966) and an application to zoogeographic data has been made by Fisher (1968). A general discussion of this approach can be found in Krumbein and Graybill (1965), and Harbaugh and Merriam (1968).

The local errors of estimating parameters of characteristics at any locality can have serious effects on the confidence that may be placed in fitted trend surfaces. On the other hand, the lack of continuous observations over the area is of little concern provided the characteristics themselves really do vary continuously in a simple fashion, i.e., as described by a low order polynomial in the coordinates.

A different situation arises if one has to forego the continuity assumption and attempts to take into account abrupt changes in characteristics from one locality to another, nearby locality. For that case one may simplify matters by assuming the area to consist of separate regions, with organisms being homogeneous in their characteristics within each region, but differing from one region to another. The statistical problem is then that of identifying and circumscribing these regions. Maps may show the regions and indicate for each region the typical values of the characteristics of the organisms. The boundaries of the regions are not isophenes in the sense of continuous variation, but must be understood as the line across which there is sharp change. (Consider the analogy of indicating cliffs on a topographical map.)

Statistically, the problems here are quite different from those encountered in continuous variation. The errors at each locality may still be of importance, but will generally affect the results less than in the contour fitting procedures described above. On the other hand, interpolation between adjacent localities of different regions becomes entirely impossible. All one knows is that the region boundary is between the two points, but there is no way of telling where exactly it may be. Statistical analysis may concern itself with grouping localities into apparently homogeneous regions, but it cannot lead to inferences about points beyond those sampled. Whenever a map displays such regions, the localities, i.e., the sampled points, should be drawn in to allow proper understanding of the meaning of the displayed regions. Between any two adjacent localities of different regions the boundary could be drawn anywhere at all, at least from a purely statistical point of view. However, there may be valid biological or physiographic reasons for locating the boundary in a certain area.

As far as statistical analysis of such data is concerned, a region is not a continuous area but a set of localities. However, to make sense geographically one would not consider any possible set of localities but try to restrict onself to sets whose localities migh be thought of as belonging to one continuous region. Thus, adjacent localities which are distant from all other localities might be considered to belong to the same region, whereas a set of localities scattered individually among other localities would not be so considered. An attempt to formalize an appropriate criterion as to when a set of localities may be said to belong to one region will be made below in section IV.

In reality, it may well be that neither the continuous variation model nor that of abrupt changes between homogeneous regions are sufficient to describe patterns of geographic variation. A trend surface may be regarded as an approximation of one kind, and a partition into regions as an approximation of another kind. A complex model with both types of variation might possibly be analyzed by considering the residuals from the first type of analysis as data for the second. This is not taken up here. Since trend surface analysis has been discussed elsewhere (Marcus and Vandermeer, 1966; Krumbein and Graybill, 1965;

Harbaugh and Merriam, 1968), the present discussion is restricted to the study of partition into regions.

Statistical inference on partitions.—The statistical technique chosen depends on the situation. In the most definite case one has to test a single statistical hypothesis, that is, whether the regions of a given a priori partition are internally homogeneous, and possibly also whether they are heterogeneous from one region to the other. Ordinary tests of significance are then appropriate, such as an analysis of variance which compares (1) the variation among regions, (2) among localities within regions, and (3) among organisms within localities. Tests of homogeneity within regions compare (2) against (3); tests among regions check heterogeneity (1) against (3) or against some pooling of (2) and (3).

When no a priori partition is indicated, one must choose between the large number of possible partitions. The present approach is to check them all and list those that are acceptable in the sense of fitting the data without significant deviations. This involves several, and possibly a large number, of significance tests on the same data, a procedure that is fraught with dangers of misleading probability statements as well as of contradictory decisions when different partitions are tested. Statistical procedures using the same data simultaneously to make tests of several related hypotheses are known as methods of *multiple comparisons* (see, Miller, 1966, Sokal and Rohlf, 1969, or other textbooks on statistics). There have been a number of applications of such methods to geographic variation analyses (Sokal, 1965; Sokal and Rinkel, 1963; Sokal and Thomas, 1965; Rinkel, 1965; Thomas, 1968a).

A particular class of such methods, known as *simultaneous test procedures* (STP) will here be applied to the geographic variation partition problem. This particular type of procedure, implicit in Tukey's early work (1951), has been developed for a variety of types of data [analysis of variance: Tukey (1951), Scheffé (1953), Gabriel (1964); multivariate analysis of variance: Gabriel (1968, 1969b); nonparametric analysis of variance: Steel (1966), Sen (1966); nonparametric anova and manova: Gabriel and Sen (1968); categorical data: Gabriel (1966)], and has been generalized under the name of STP by Gabriel (1969a). Its logical and probabilistic neatness and its wide applicability have led to its choice for the present purpose.

Simultaneous test procedures.—The arithmetic of an STP is to compute a certain test statistic for each hypothesis and compare it with the critical value for the STP, rejecting the hypothesis if the statistic exceeds that critical value, accepting it otherwise. The essential feature of STP's is that the *same critical value is used for the statistics of all the different hypotheses.* In Tukey's (1951) original procedure one would test the homogeneity of each region by comparing the (standardized) range of all the sample means of that region with a critical point from the (standardized) range distribution, always taking the distribution for as many means as there were localities in the *entire* area. In Gabriel's (1964) adaption of Scheffé's (1953) procedure one uses the sum of squares between localities of the region instead of the range of means. In other STP's one uses other statistics in essentially the same way.

One of the tests of an STP will be of overall homogeneity, i.e., for all localities in a geographic variation analysis. The critical value is chosen to make the overall test for the exact level desired for the entire STP. This ensures not only that the probability of falsely rejecting the overall hypothesis be α, but also that the probability of falsely rejecting any true hypothesis, overall or subsidiary, be no more than α. In other words, the probability that all true hypotheses will be accepted by an α-level STP is at least $1-\alpha$.

The probability of acceptance of any particular true hypothesis may well be above $1-\alpha$, and indeed will be so in some cases. Low probabilities of false rejection would in themselves be desirable, but they are unfortunately associated with reduced probabilities of correct rejections, that is, of detecting heterogeneity. This loss of power is the price one pays in testing many hypotheses simultaneously. To preserve a low probability of any false rejection, one loses some chances for finding differences when they really exist.

The hypotheses to be tested all relate to the homogeneity of populations from different localities. They may range from a general hypothesis of overall equality, through hypotheses of equality within some region (subset of localities), to hypotheses of pairwise equality of two localities. They will also include, most importantly, hypotheses on partitions of the entire area into regions, the hypotheses being that the localities are homogeneous within each region. Clearly, there exist among all these hypotheses many relations of implication: overall homogeneity implies homogeneity within each region and for all partitions; if one partition is subpartitioned into another, homogeneity of the regions of the former certainly implies the same for the latter; if one region contains another, the former cannot be homogeneous unless the latter is; etc., etc. It is essential that the decisions of multiple comparisons do not violate any of these implications. Thus, if the homogeneity hypothesis for a region is accepted, it should not be rejected for any subregion of that region. This property, which is referred to as *coherence* ("transitivity" in Gabriel, 1964), holds for all STP's.

Relations of implication among hypotheses also work in another direction. If a particular hypothesis is untrue, this may mean that the other hypotheses it implies cannot all be true. However, it does not indicate which of these others are untrue. Thus, if a partition of the entire area into several regions is not homogeneous at least one of its regions must also be heterogeneous. Similarly, if a region is heterogeneous, at least one pair of localities in the region must differ. It would be desirable to have methods of multiple comparisons which always provide resolution of a rejected hypothesis into rejection of some more detailed hypotheses, i.e., which indicate at least one pair of differing localities in every significant region. However, most methods, including most STP's, do not have this property of *consonance*, but allows *dissonant* decisions such as rejecting homogeneity of a partition as a whole whilst accepting homogeneity for each of the regions by itself. This is a drawback of these methods, as of significance tests in general, in which "acceptance" has to be understood as insufficient evidence for rejection, rather than as evidence supporting the hypothesis.

It will be clear at this stage that STP's, in particular if they are not consonant, allow a good deal of nonuniqueness in their decisions. Take the simplest instance of a region with three localities, A, B, C which has been found significantly heterogeneous. Clearly at least two pairs of localities must also be heterogeneous, but a nonconsonant STP may not indicate any such pair and a consonant STP may merely indicate a single one. If A is significantly different from C, but (A,B) and (B,C) are nonsignificant, the decision is indeterminate in that one has not decided whether B differs from A or from C (or from both).

The indeterminacy is equally clear in testing partitions where several alternatives may all be nonsignificant and thus "acceptable," and the choice between them remains undecided. Again, this is an essential feature of any method based on significance testing, which merely asks, "what is acceptable in view of the data." The estimation approach, on the other hand, asks, "what fits the data best," and ignores all other answers that are not best, though they may be almost as good. There is a good deal to be said for presenting the variety of acceptable regions and partitions and thus stressing the limitation of reliable statistical inference from a body of data. The alternative of choosing one particular partition may be more satisfactory to the research worker, but it introduces a spurious element of

TABLE 1. EXAMPLE OF AN F-STATISTIC ANOVA STP DATA[1]. CUBIC ROOT OF BODY WEIGHT OF FEMALE RED-WINGED BLACKBIRDS (*Agelaius phoeniceus*), FROM SEVEN LOCALITIES IN NORTH AMERICA (LOCATED AND IDENTIFIED IN FIG. 5, BELOW).
a = number of localities = 7

	Locality						
	1	2	3	4	5	6	7
	3.708	3.708	3.583	3.448	3.503	3.451	3.332
	3.733	3.530	3.634	3.503	3.624	3.517	3.362
	3.570	3.672	3.530	3.530	3.634	3.575	3.391
	3.659	3.622	3.476	3.476	3.583	3.476	3.271
	3.609	3.659	3.609	3.530	3.443	3.672	3.391
	3.744	3.634	3.583	3.517	3.224	3.420	3.448
	3.476	3.609	3.503	3.476	3.471	3.377	3.391
	3.708	3.596	3.583	3.557	3.634	3.503	3.476
	3.915	3.659	3.530	3.609	3.498	3.434	3.391
	3.530	3.684	3.583	3.503		3.476	3.332
	3.744	3.634		3.634		3.462	3.302
	3.708	3.647		3.503			3.420
		3.570		3.557			3.362
		3.544		3.503			3.557
		3.570		3.557			
				3.667			
				3.674			
n_i	12	15	10	17	9	11	14
ΣY_i	44.104	54.338	35.614	60.245	31.614	38.363	47.426
$\bar{Y}$	3.675333	3.622533	3.561400	3.543824	3.512667	3.487545	3.387571
s	0.1163	0.522	0.0495	0.0667	0.1309	0.0804	0.0196

Anova

Source of variation	*df*	SS	MS	F
Among localities	6	0.6887	0.1148	16.818***
Within localities	81	0.5529	0.0068	

Critical SS $= s^2(a-1)\ F_{\alpha[a-1,\ \Sigma n_i - a]}$

For $\alpha = 0.05$ the critical SS $= 0.0068(7-1)2.22 = 0.0906$ *** $= P < 0.001$

[1] Data by Power (1970).

definiteness which is most unlikely to be borne out by any future observations.

An example of an STP.—These concepts will become clearer by considering an example (Table 1) from a geographic variation study in ornithology (Power, 1970). The overall analysis of variance for the seven localities is shown in the table. It is highly significant by the *F*-ratio or, equivalently, when the sum of squares among localities (0.6887) is compared with the critical SS (0.0906).

Next, one proceeds to find the largest homogeneous subsets. One may employ various strategies. One can test the homogeneity of every pair of localities in combination, and, remembering the property of coherence, test only those triplets made up entirely of homogeneous pairs. Similarly, one would test only those subsets of four that contain only homogeneous triplets. On the other hand, one may suspect certain large subsets to be homogeneous and proceed to test these right away. If their SS is greater than the critical SS one rejects the null hypothesis and tests smaller subsets. If, however, the larger set is not significant there is a saving of much computational labor. When done on the computer it is simplest to test all combinations and draw inferences from their significance or the lack thereof.

Each sum of squares is computed from the general formula

$$\sum^{l} \frac{(\sum^{n_i} Y_i)^2}{n_i} - \frac{(\sum^{l}\sum^{n_i} Y_i)^2}{\sum^{l} n_i},$$

where i is the index of locality, l is the number of localities in the subset, n_i is the number of observations, and $\sum^{n_i} Y_i$ is the sum of the observations at locality i. Thus the sum of squares for the subset of localities (1, 2, 3) is computed as

$$\frac{(44.104)^2}{12} + \frac{(54.338)^2}{15} + \frac{(35.614)^2}{10} - \frac{(44.104 + 54.338 + 35.614)^2}{12 + 15 + 10} = 0.0708.$$

By proceeding in the manner described above one arrives at the following largest homogeneous sets of localities (1, 2, 3), (2, 3, 4, 5), (3, 4, 5, 6,), and (6,7). Each one of these sets is "largest" in the sense that addition of any further locality to one of them would make it heterogeneous (i.e., it would have a sum of squares greater than the critical SS). Thus, for instance, the SS for the set (1, 2, 3, 4) equals $0.1441 > 0.0906$.

Further details on the computation of STP's are furnished by Gabriel (1964) and in the textbook by Sokal and Rohlf (1969), where Section 9.7 and Boxes 13.7 and 14.9 deal with this subject. A computer program to carry out the type of STP (SS-STP) illustrated above is furnished in Appendix A3.6 of the same book.

IV. GEOGRAPHIC PARTITIONS, REGIONS AND CONTIGUITY

Geographic variation analysis is concerned only with a particular type of partitions, namely those in which the parts are geographic regions. Thus, for example, if Maine localities are associated in one region with other localities from New England and the Maritime provinces of Canada, whereas Oklahoma localities are associated in another region with other southwestern localities, one may have a geographically reasonable partition. On the other hand, if a region consists of one sample from Bar Harbor, Maine, a second from Little Rock, Arkansas and a third from Casper, Wyoming, one doubts whether the resulting partition would make any biogeographical sense. It is well to decide beforehand what one may or may not accept as of interest in a geographic variation study. This also reduces the vagueness of the questions asked and the indeterminacy of the answers provided. Clearly, the definition of a geographic partition hinges on that of its component parts, that is, of regions.

Regions defined.—The operational definition of a region has to be given in terms of a fixed number of sampled localities. Yet these localities should be said to belong to a region only if it is plausible to think of them as representing a continuous region including an infinity of other, continuously connected points.

In a continuous area a region is visualized as a connected set of points, of whatever shape. The crucial aspect is the connectedness, which means that any two points A and Z of a region **R** are linked by a continuous closed band of points belonging to **R**. The topological properties of such regions will not be discussed in detail. It suffices to note that no two mutually exclusive regions **R** and **P** could intersect with one another. For if any band of points from **R** did intersect with a band from **P**, the point of intersection would be common to **R** and **P**, which is not possible since **R** and **P** are mutually exclusive.

One needs to extend the definition of connectedness to a set of localities from among a finite collection on a map. Again, one will define such a set as connected, and take it to belong to one region, if any two localities A and Z of the set **R** are connected by a chain of localities B, C, . . . , X, Y of

R, such that the pairs (A, B), (B, C), . . . , (X, Y), (Y, Z) are all contiguous. This definition hinges on the concept of contiguity of pairs of localities (defined in the next section). Contiguities must satisfy the following two requirements; (1) if (A, B) are contiguous there exists a closed continuous band of other points (not sampled) from A to B, (2) if (A, B) and (C, D) are each contiguous, the band joining A and B has no point in common with the band joining C and D. The first requirement is obvious, the second is necessary to avoid the contingency that two mutually exclusive connected sets should belong to two regions which have points in common.

It will have been noted that *the definition of a connected set, or region, that is sought here involves only its geographical features, not the characters of the organisms found in it.* The ultimate aim of the present treatment is indeed to find regions which are meaningful both geographically and in terms of homogeneity of characters of the organisms. But the approach of this paper is to seek first a purely geographic definition of a region and then, using that definition, to search for those regions which in some optimal way interpret the geographic variation of the characters.

A definition of contiguity.—The intuitive ideas of contiguity and connectedness suggest that contiguous localities should be close rather than distant, both distant from any other localities rather than intermixed with many others, joined by an unbroken straight line rather than by a line passing through other localities between them, etc. And the two requirements mentioned in the previous section must, of course, be satisfied.

It is possible to find many definitions that satisfy these requirements for any given set of localities. For example, the straight lines of any triangulation grid joining the localities satisfy them. But there are many possible such grids and there seems no way of determining a unique and intuitively acceptable triangulation grid as a basis for a definition of contiguities. Joining all localities within some given distance of one another (single linkage, Sokal and Sneath, 1963) will generally violate requirement (2) for contiguity relations. Joining only mutually closest localities will leave many localities unconnected, resulting in the representation of regions by single localities.

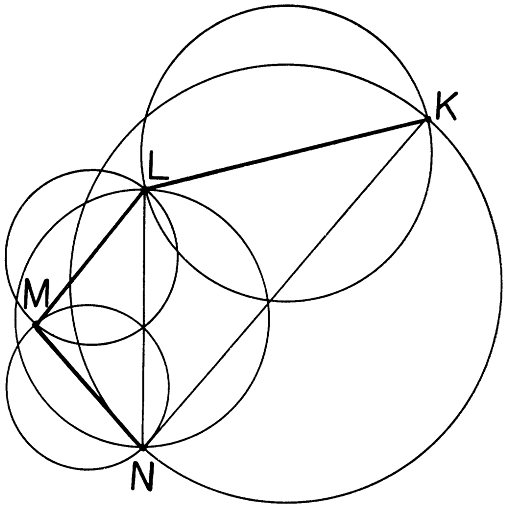

FIG. 1.—Contiguity defined. For explanation see text.

A simple definition, which is unique, is that any two localities A, B are said to be contiguous if, and only if, all other localities are outside the *A-B circle*, that is, the circle on whose circumference A and B are at opposite points. In other words, two localities A and B are contiguous unless there exists some other locality C such that in the triangle ABC the angle subtended at C is of 90° or more. This seems an intuitively acceptable definition, for if such a locality C exists it should really be considered as being "between" A and B, so that a closed continuous band of points from A to B would most likely include C, and thus A and B should not be considered contiguous. In Figure 1 the following pairs of localities are contiguous by this definition: K, L; L, M; and M, N. The pair K, N is not contiguous since locality L falls within the circle whose diameter is KN. The other way of showing that K and N are not contiguous is by

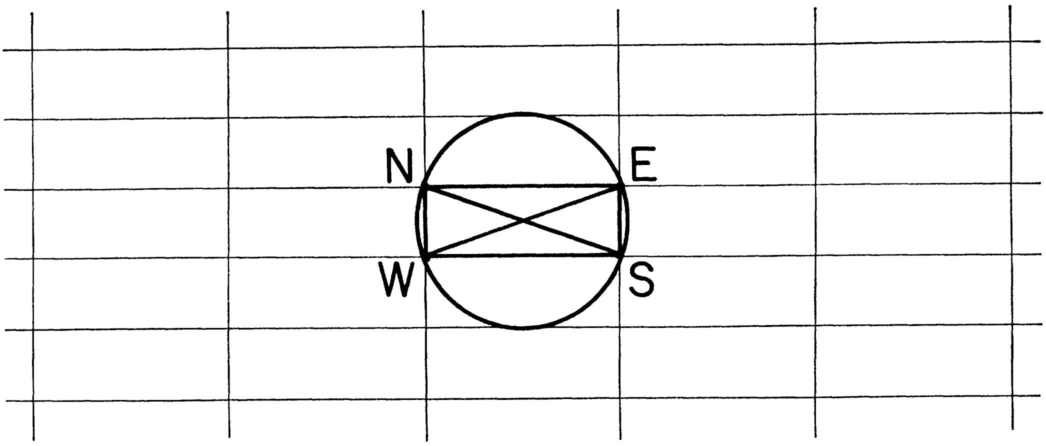

FIG. 2.—Contiguity in a rectangular grid. For explanation see text.

demonstrating angle KLN to be greater than 90°.

The following example may substantiate the proposed criterion of contiguity. Consider a lattice of points, i.e., the intersection points of a rectangular grid, equally spaced horizontally and equally spaced vertically (see Figure 2). If one takes two adjacent points N, E on a horizontal line, and the corresponding two points W, S on the next horizontal line below, one obtains the four vertices of a rectangle. If one now looks for reasonable definitions of contiguities on that lattice, one would certainly agree that adjacent N and E, W and S, N and W, and also E and S should be considered contiguous. But what about N and S, or E and W? In terms of distances and the configuration of the lattice, one could equally well justify calling either pair contiguous. But they cannot both be contiguous, for the straight line bands of points joining them do intersect; hence a reasonable definition is to consider neither of them contiguous. And this is exactly what the proposed criterion of contiguity does. For the N-S circle has W and E on its circumference, and the W-E circle, being the very same circle, has N and S on its circumference. Hence, the proposed criterion assigns contiguity relations exactly as one would on intuitive grounds. Some graphic experimenting will probably convince most readers that this is indeed an intuitively acceptable criterion.

Pending a better definition, or, indeed, any useful alternative definition, this definition of contiguity is proposed as a good working criterion. If another criterion were used instead, particular results would be affected, but the general considerations of this paper would remain unchanged.

Note that the proposed definition takes account only of the location of the localities on the plane map. It ignores not only characteristics of the organisms but also the features of the map itself. Existence of rivers, lakes, oceans, mountains, deserts, etc., have been entirely ignored. Two distant localities may be considered contiguous because an intervening natural obstacle prevents any localities being studied between them, whereas close localities may not be considered contiguous because of the proliferation of suitable sampling localities between them. This points to a drawback of the proposed definition which might be reduced by taking into consideration certain natural barriers and disallowing any contiguities which cross them. It would probably be difficult to do this in an entirely objective manner, but it may well be worth while in an attempt to allow only meaningful regions. A decision will have to be made in any study as to whether and how

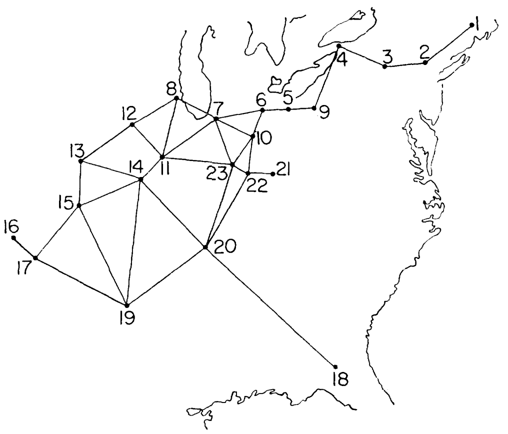

FIG. 3.—Twenty-three localities in eastern North America sampled by Sokal and Rinkel (1963) in a geographic variation study of the aphid *Pemphigus populi-transversus*. Contiguous localities are indicated by lines connecting the locality numbers.

to restrict contiguities beyond the restriction imposed by the definition proposed above.

Contiguity partitions.—Note first that the entire collection of localities in a study always forms a connected set under the proposed definition, so that the entire area under study may always be considered as a region.

A less trivial question is whether a given subset of localities forms a connected set and may be taken to represent a region within the entire area under study. And further, it is of interest to consider partitions of the collection of localities into mutually exclusive subsets, each one of which is a connected set. This will be referred to as a contiguity partition and corresponds to the decomposition of the entire area into a number of disjoint regions.

Fig. 3 illustrates these concepts. The numbered points are 23 localities in eastern North America (sampled by Sokal and Rinkel, 1963, in a geographic variation study of the aphid *Pemphigus populi-transversus*). The lines joining points indicate all contiguity pairs by the proposed definition. No triangle in the Figure has an angle of 90° or more, but drawing in any additional line between any two localities would create a triangle with such an angle. Note, first, the trivial fact that all 23 localities are connected. Note, further, that some subsets are connected, e.g., localities 1–10, localities 11–17, localities 18–23, whereas other subsets are not connected, e.g., localities 1–6, localities 7–9, localities 20–21. Finally, note that 1–10, 11–17, 18–23 form a possible contiguity partition into three mutually exclusive connected sets, each representing a separate region. Also note that 1–6, 7–9, 10–19, 20–21, 22–23 do not form a contiguity partition since not all the parts are connected sets.

There is not one unique contiguity partition for any collection of localities, but quite a large number of such partitions may be possible. The criterion for choosing among them will be mainly in terms of the characters of the organisms sampled at the localities, with a view to having homogeneity within regions and heterogeneity among regions. However, one may also wish to impose some further, purely geographical or biological restrictions on what sets will be considered to represent regions and what partitions decomposing the area into regions may be reasonable.

Stronger criteria for regions and partitions—The connectedness criterion accepts such strung out sets of localities as 1, 2, 3, 4, 9, 5, 6 as well as more highly interconnected sets such as 7, 8, 11, 12, 13, 14, 15 (both in Fig. 3). One might prefer to consider only those sets as representing regions that are more highly interconnected in some sense. This requires a measure, or index, of *degree of connectivity*. Once such a criterion is developed one can employ it to eliminate from the large number of statistically acceptable regions and partitions those that do not meet the established standards of connectivity.

Most measures of connectivity for a single set relate the number c of contiguity pairs (connected localities) of the set, to the total number l of localities in the set in some form of a ratio. A number of such measures are available from graph theory. For example, Garrison and Marble (1961) defined

$\alpha = (c - l + 1) / (2l - 5)$ for $l = 3, 4, \ldots,$. This coefficient ranges from zero for a completely strung out set, to increasing values as the number of contiguities per given number of localities increases. Thus, the two sets mentioned above have α's of 0 and 0.44, respectively. The range is always from zero up, but with the criterion of contiguity proposed here the maximum is 1 for $l = 3$, $2/3$ for $l = 4$, $3/5$ for $l = 5$, $5/7$ for $l = 6$, and $2/3$ for $l = 7$. No explicit formula is at present available for this maximum, though it seems always to be close to $2/3$ (except for $l = 3$).

In the present context it may be more relevant to seek a measure of separateness of a set from all other localities. Thus, in Fig. 3, both sets of localities 4, 9, 5 and 7, 23, 22 are strung out, but whereas the former is little connected with other localities, the latter is highly connected. Again, the triangular set 15, 17, 19 is more separated out from other points than the triangular set 7, 10, 23.

A suitable measure of separateness of a set **R** is defined as follows. First, define a *contiguity triangle* as a set of three localities of which each pair is contiguous. Then count the number t of contiguity triangles which have exactly two localities in **R**. The proposed index of separateness is then

$$\lambda = 1 - \frac{t}{2c},$$

which clearly varies between zero and one in an appropriate manner. Thus, the index λ assumes the values $1 - (0/4) = 1$, $1 - (4/4) = 0$, $1 - (1/6) = 5/6$ and $1 - (3/6) = 1/2$ for the four triplets mentioned above, and 1 and 0.9 for the two sets of seven mentioned earlier.

The two measures discussed above are but two of a substantial array that could be defined and might be applicable. The interested reader will find a comparative account in Kansky (1963). To determine the most suitable measures of connectivity and separateness for geographic variation analysis will require extensive empirical investigation. Possibly other approaches for defining regions (see for example Neely, 1969) may prove more suitable. In any event it should be stressed that the statistical techniques proposed in this paper would be compatible with almost all of these criteria of connectivity and separateness. One would be able to choose some such index as the above and consider only sets for which the index exceeded a certain cut-off point.

A note on analogies with graph theory.—A close correspondence exists between the concepts of the present section and those used in graph theory. (For a recent exposition see Busacker and Saaty, 1965). If localities are considered as *vertices* in graph theory, and contiguities as *edges*, then the regions or connected sets of this discussion are *connected graphs*. Since the localities are in a plane, and the edges are not allowed to intersect, the present sets are *connected planar graphs*. However, in view of the particular geometric definition of contiguity adopted here this is a special type of planar graph which has apparently not received attention in graph theory. Thus, for example, in defining an index α of *connectivity* graph theoreticians divide the *cyclomatic index* $(c - l + 1)$ by the maximum it can attain which is $(l-1)(l-2)/2$ for a general graph, and $2l - 5$ for a planar graph. But it was noted above that this maximum is unattainable with the contiguity edges used in the present study, so that $\alpha = (c - l + 1)/(2l - 5)$ usually has a maximum of about $2/3$.

Further study of the analogies of graph theory with the present concepts might lead to fruitful applications of graph theoretic techniques in the analysis of geographic regions.

V. STP'S FOR REGIONS AND FOR PARTITION

STP's for connected sets.—Some geographic variation studies are concerned only with the identification of regions homogeneous with respect to some character of the organisms under study. The regions may be identified by means of connected sets of localities—or possibly by some stronger criterion for sets as discussed in section IV above—and the homogeneity of each such set may be decided on by an STP of the character under study.

TABLE 2. ABO BLOOD GROUP FREQUENCIES[1] FROM FOURTEEN SAMPLES IN JAPAN (LOCATED AND IDENTIFIED IN FIGURE 4).

Localities	Blood Types				Row Totals
	O	A	B	AB	
1	421	582	298	127	1428
2	437	582	332	131	1482
3	9316	10814	6688	2887	29705
4	1767	2106	1301	550	5724
5	1403	1785	924	431	4543
6	2967	3978	1959	904	9808
7	4854	6379	3243	1561	16037
8	817	1031	510	291	2649
9	1109	296	906	31	2342
10	869	959	696	261	2785
11	727	796	563	283	2369
12	3303	4119	2267	1041	10730
13	543	775	429	188	1935
14	1738	2088	1437	649	5912
Column totals	30271	36290	21553	9335	97449

[1] Data extracted from Mourant, Kopeć and Domaniewska-Sobczak (1958:152–157).

The STP will be used to test hypotheses of homogeneity only for those sets meeting the criterion of connectedness—or some other criterion. The significance of the test statistic for each set will decide whether the set is rejected, or acceptable as homogeneous. In view of the coherence property of STP's—section III above—any subset of an acceptable set must also be acceptable and, conversely, no set can be acceptable if it contains a rejected set. An acceptable set might therefore be enlarged by addition of further localities and become rejected, but a rejected set can never be enlarged so as to become acceptable. This leads to the identification of sets which are, in themselves, acceptable, but which result in rejected sets on addition of *any* further locality. Such sets will be called *maximal acceptable*. A little reflection will show that every acceptable set must be contained in, or at most equal to, some maximal acceptable set, whereas all sets that are not so contained must be rejected. The results of the STP may therefore be summarized by a list of all maximal acceptable sets, which will in general be very much briefer than the entire list of all acceptable sets.

In the geographic variation study of regions as defined in this paper, the STP results are summarized by a list of all maximal acceptable connected sets. These sets represent the largest homogeneous connected regions, such that different homogeneous connected regions may well overlap, but none may be part of another. Where the overlaps of such largest regions are not too numerous, the list of maximal acceptable connected sets may by itself provide for adequate categorization of the variation pattern of the organism. It may then be feasible to shade each largest region on a map, with cross-shading showing overlaps.

Example of finding maximal acceptable connected sets (using categorical data).—The method of determining regions will be illustrated by finding maximal acceptable connected sets in an example chosen to demonstrate the versatility of the STP approach. The data employed are categorical (frequencies of the ABO blood group types—Table 2) but one could as easily have used continuous data such as the body weights of Table 1. The blood group samples are of people from 14 Japanese localities identified and shown on a map in Figure 4. Contiguities are indicated on the map as lines linking contiguous localities.

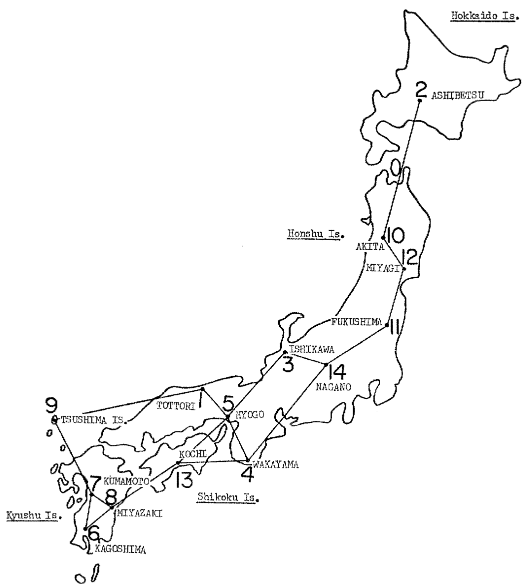

FIG. 4.—Fourteen localities in Japan for which ABO blood group frequencies were extracted from Mourant, Kopeć and Domaniewska-Sobczak (1958: 152–157). The locality names refer to prefectures, except for 2 (which is a town) and 9 (which is an island). Contiguous localities are indicated by lines connecting the locality numbers.

The STP for categorical data (Gabriel, 1966), is based on the G-statistic, which is also known as the information or likelihood ratio statistic. This statistic is discussed in detail by Sokal and Rohlf (1969, Ch. 16), who illustrate its employment for an $R \times C$ test of independence such as these blood group data (Box 16.8) and show an STP for categorical data (Box 16.5). They also provide an appropriate computer program in their Appendix A 3.15.

The actual data are given in Table 2. The computation of the G-statistic involves transformation of frequencies f to $f \ln f$. For desk calculator work this is best done by means of special tables, such as Table **G** in Rohlf and Sokal (1969). Of course, when done by computer, the transformation is carried out as part of the program.

To carry out this STP one first tests the hypothesis of overall homogeneity by computing

$$G = 2\Big[\sum^{a}\sum^{b} f_{ij} \ln f_{ij} - \sum^{a}\Big(\sum^{b} f_{ij}\Big) \ln \Big(\sum^{a} f_{ij}\Big) - \sum^{a}\Big(\sum^{b} f_{ij}\Big) \ln \Big(\sum^{b} f_{ij}\Big) + \Big(\sum^{a}\sum^{b} f_{ij}\Big) \ln \Big(\sum^{a}\sum^{b} f_{ij}\Big)\Big],$$

where f_{ij} is the frequency of the jth blood group at the ith locality, a is the number of localities ($=14$) and b the number of blood types ($=4$). One may express the above formula in words as follows: two [sum of transforms of the frequencies in the body of the contingency table minus sum of transforms in the column totals minus sum of transforms of the row totals minus transforms of the grand total]. For the entire data of Table 2 one obtains

$$G = 2[780{,}524.758 - 905{,}456.953 - 993{,}751.820 + 1{,}119{,}404.891] = 1441.752.$$

This is compared with the critical point of the χ^2- distribution with $(a-1)\ (b-1) = (14-1)(4-1) = 39$ degrees of freedom, being 54.471 for an STP of level $\alpha = 0.05$. Since G greatly exceeds this critical value, the hypothesis of overall homogeneity is rejected significantly.

One now proceeds to test particular subsets, confining oneself to connected sets, since others are not of interest in a geographic variation study according to the approach outlined above. Thus, set (1,2, 3,4,) will not be tested, but (4,6,7,8,9,13,14) will be. For the latter set one computes G as before, except that one now includes only the seven rows representing the localities in the set. Of course, the column sums, the sum of the row sums, and the grand total will change accordingly. The statistic is found to be

$$G = 2[343{,}538.340 - 399{,}990.320 - 418{,}075.305 + 475{,}206.074] = 1357.578.$$

This is still well above the critical 54.571, so the set (4,6,7,8,9,13,14) is not acceptable as representing a homogeneous region.

By successively removing localities from the set without breaking its connectedness, one will finally arrive at an acceptably homogeneous connected set. In this example removal of localities 9 and then 6 still leaves an unacceptable set (4,7,8,13,14) with $G = 87.820$ (> 54.571). Further removal of locality 7 yielded set (4,8,13,14) which has $G = 45.776$ (< 54.571) and is therefore acceptable as representing a region. One must now proceed to attempt to enlarge this set, seeing if any of the larger connected sets (4,6,8,13,14), (4,5,8, 13,14), (3,4,8,13,14), (4,8,11,13,14) may be acceptable. The respective G values 94.719, 63.144, 57.374, and 104.101 all exceed the critical value of 54.571, hence the set (4,8, 13,14) is a maximal acceptable connected set.

It is well to go through this procedure in some systematic manner. After a few steps one will require increasingly fewer actual calculations because the sets at issue often either are subsets of sets already accepted or contain rejected sets. In those cases the acceptance or rejection, respectively, follows at once from the coherence property and G need not be computed.

The following is the list of maximal acceptable connected subsets for the present example. This list was computed in less than 7/100th of one minute on a G.E. 625 computer at The University of Kansas.

Maximal set of size 1: (9)
Maximal set of size 3: (11,12,14)
Maximal sets of size 4: (1,4,5,14), (2,10,11,12), (3,4,5,14), (4,5,13,14), (4,5,11,14), (4,8,13,14)
Maximal sets of size 5: (1,3,4,5,13), (1,3,5,8,13), (3,4,5,8,13), (3,4,11,13,14)
Maximal set of size 7: (1,4,5,6,7,8,13)

There is a certain amount of indeterminacy in that, for instance, locality 11 may be in a northern region starting from locality 2, or it may belong to a central region extending as far as localities 4 and 5. Even so, certain salient features do emerge: the northern localities of Ashibetsu on Hokkaido (2) and Akita (10) belong to a region including Northern Honshu but not extending south of Fukushima. Tsushima Island forms a separate region of its own. The island of Kyushu belongs to a southern region which may extend through Shikoku Island into central Honshu but certainly not beyond the Kyoto-Osaka area. However, it may also be that the localities of central and southwestern Honshu form one or more regions by themselves, separate from the island of Kyushu. The data do not allow more definite unequivocal conclusions, but for some purposes this may already be of value. Using a stronger criterion of connectivity (as suggested in section IV) would have reduced the number of maximal acceptable sets and quite likely would have made the interpretation of these data simpler and less ambiguous.

STP's for contiguity partitions.—Most geographic variation studies will require STP's not only for testing of single regions, but also, or primarily, for the testing of contiguity partitions. The hypotheses of concern will be those of homogeneity of partitions, meaning homogeneity within each region of the partition. The regions themselves, as before, are identified by means of connected sets of localities, possibly with some further restrictions in terms of connectivity or separateness.

The criteria of contiguity, connectedness and possibly of separateness serve to determine which partitions are of interest, that is, which hypotheses are to be tested. The STP provides the significance tests for these hypotheses and hence the decisions as to which partitions are acceptable and which must be rejected for showing significant heterogeneity within parts, i.e., within regions. The coherence property of the STP's plays a role here, too. Consider two contiguity partitions one of which is a subpartition of the other, that is, each part of the latter is made up of one or more parts of the former. One may then speak of the former as being a finer partition, and the latter a coarser partition. The more the parts are subpartitioned, the finer the partition, the more parts are amalgamated, the coarser the partition. Clearly, if a partition is homogeneous, i.e., each of its parts is homogeneous, so is every one of its subpartitions. And conversely, a partition must

FIG. 5.—Seven localities in North America represented by blackbird samples (D. M. Power, 1970; see Table 1). Contiguous localities are indicated by lines connecting the locality numbers.

be heterogeneous if any its subpartitions is. This imposes logical implication relations on the homogeneity hypotheses for partitions and these are preserved by the STP's. Thus, if a partition is acceptable, so is every one of its subpartitions; conversely, a partition must be rejected if any of its subpartitions is rejected.

Starting from any contiguity partition which is rejected, one might subpartition until one arrived at a partition fine enough to be acceptable. (The finest partition, with a single locality in each part, is clearly always homogeneous). Conversely, starting from an acceptable partition one might amalgamate parts until one arrived at a partition coarse enough for homogeneity to be rejected. (The coarsest partition, in which all localities belong to the same single part, may be assumed to be rejected; for in cases in which it is accepted—acceptance of the null hypothesis for the overall study—there is no need for further detailed study of partitions.) At some stage of these processes one comes across the coarsest contiguity partitions that are acceptable, but which are such that any amalgamation of their parts results in a rejected partition. Clearly, any acceptable partition is either coarsest or a subpartition of a coarsest acceptable partition. Therefore the results of an STP on geographic partitions may well be summarized by a list of all *coarsest acceptable contiguity partitions.* There will be several, possibly many, coarsest acceptable contiguity partitions. The nonuniqueness of the result, as expressed in the plurality of solutions, is an essential feature of the method, and indeed of any method based on tests of significance. It rejects some possibilities, but does not choose among the others.

Example of finding coarsest acceptable contiguity partitions (using continuous data).—The data are the cube roots of body weights of female red-winged blackbirds given in Table 1. The localities are plotted in Fig. 5. In addition to the original observations, Table 1 also shows sample sizes and means as well as an analysis of variance which provides an error variance estimate of 0.0068 with 81 degrees of freedom and an $\alpha = 0.05$ level critical sum of squares value of 0.0906. It was already noted above that the overall hypothesis of homogeneity of all localities is rejected decisively.

One may now proceed to test various contiguity partitions of the entire set of localities. In line with our previous approach one will be interested only in contiguity partitions. In view of the simple contiguity relations between the localities of this study (Fig. 5) one readily identifies the six possible contiguity partitions into *two* regions as (1|2,3,4,5,6,7), (1,2|3,4,5,6,7), . . . , (1,2,3,4,5,6|7). For any one of them one computes the sum of squares and tests it against the critical SS = 0.0906, mentioned above. Thus, for the second of the above partitions one computes the sum of squares between localities 1 and 2, 0.0185, and the sum of squares among localities 3,4,5,6 and 7, 0.2495. The sum, 0.2680, is the required sum of squares. It exceeds 0.0906 and so this partition is rejected. Regions containing only a single locality do not contribute to the sum of squares. In a similar way all the other five partitions above would also be rejected in this example.

One proceeds to three region partitions (1|2|3,4,5,6,7), (1|2,3|4,5,6,7), (1|2,3,4|5,6,7), . . . , (1|2,3,4,5,6|7), and then (1,2|3|4,5,6,7), (1,2|3,4|5,6,7), . . . , up to (1,2,3,4,5|6|7),—computing the sums of squares among localities for each region and summing them to obtain the partition sum of squares. Again, a partition is accepted only if its sum of squares is less than the critical sum of squares. Thus, in the present example, partitions (1,2|3,4,5,6|7) and (1,2|3,4,5|6,7) are the only acceptable ones. Clearly, these are also coarsest acceptable contiguity partitions. However, one further has to check other partitions into four parts which are not subpartitions of these. Thus, one has to check the partitions (1|2,3,4,5|6|7), (1|2,3,4|5,6|7), (1|2,3,4|5|6,7), (1|2,3|4,5,6|7), etc., whose sums of squares are, respectively, 0.0821 (acceptable), 0.0553 (acceptable), 0.1138 (rejected), 0.0441 (acceptable), etc. The final list of coarsest acceptable contiguity partitions is

	No. of parts
(1,2|3,4,5,6|7)	3
(1,2|3,4,5|6,7)	3
(1|2,3,4,5|6|7)	4
(1|2,3,4|5,6|7)	4
(1|2,3|4,5|6,7).	4

Thus, there clearly are two extreme regions, each containing the most outlying locality in one direction and at most one other locality. The other localities belong to one or two intermediate regions.

VI. COMPUTATIONS

For any sizeable number of samples the number of tests required for an STP to determine the class of maximal acceptable connected sets is too large for practical use without computer facilities. Programs in FORTRAN IV have therefore been written to carry out these computations and print out the list of these sets. Such programs are available from the second author for use with various STP's according to the type of data at hand: single variable normally distributed, single variable nonparametric, multivariate normally distributed and categorical. These programs differ from those available in Appendix A3 of Sokal and Rohlf (1969) in enabling the user to test connected sets only.

An example with 22 univariate samples took 8 minutes on the IBM 7040 to run with the nonparametric STP and 10 minutes with the sum of squares STP. In the former case there were 8 maximal acceptable connected sets, in the latter 30. The latter sets were generally smaller, as analysis of variance is often more powerful than nonparametric tests, and hence that analysis took less time on the computer. On the G.E. 625 these running times were considerably reduced.

Input to these programs consists of (1) sample data (2) a critical value and (3) all contiguity relations between pairs. Output is then printed out as a list of maximal acceptable connected sets.

For contiguity partitions no systematic computation has yet been programmed, but programs are available which allow the questioning of significance for any number of partitions. Here the input is, in addition to (1) data and (2) critical value, (3) a list of partitions, detailing the points belonging to each subset part. Output is in form of a significance decision for each partition. Again, these questioning programs are available for STP's for a variety of types of data.

VII. DISCUSSION

To the extent that categorization is required in the interpretation of geographic variation data, the methods proposed here would appear to have utility. For many characteristics it might be argued that representation of geographic variation patterns might best be done on a continuous trend surface basis without categorization. The argument in favor of the trend surface approach is that if variation is truly continuous, any categorization into mutually exclusive groups is a falsification of the true relations among the observations. In a sense trend surface analysis is the logical culmination of the recent trend toward direct description of population patterns and away from categorization (Ehrlich and Holm 1962, Sokal and Sneath 1963). Among the difficulties that accompany the employment of trend surface analysis is the lack of adequate methods for assessing the statistical confidence that can be ascribed to fitted contour lines in trends. This is one of the problems which led to the development of the techniques in the present paper. Furthermore, theoretical objections to categorization notwithstanding, there continues to be a need for the categorization of homogeneity subsets for a variety of purposes ranging from straight taxonomy to population genetics. In such cases the methods presented above will be of value.

Even when the regions are screened statistically and by contiguity, one often finds multiple solutions that involve overlapping subsets resulting from STP procedures (and other multiple comparisons tests). While these seem unsatisfactory from the point of view of neat partitions of the area under study, they may arise both because of the inherent error in the data, and, possibly, because the true nature of the variation over the area is continuous. The nonuniqueness of the categorization is regrettable from the point of view of orderliness, but does point out the indeterminancy of conclusions from limited sample data and the problems of continuous geographic variation. It should caution the practicing systematist to shy away from the hasty, subjective judgements on boundaries of ecotypes, races and subspecies which abound in the literature.

Regions need not only be defined by geographic continuity, but factors such as topography, rainfall, elevation and others need also be considered. Statistical regions may thus be defined not only as geographically contiguous sets, but also as containing ecologically or topographically similar localities. The statistical methods presented here would be equally applicable to any other definition of regions and partition.

ACKNOWLEDGMENTS

Contribution No. 1299 from the Department of Entomology of The University of Kansas. This research was supported by Grant No. GB-3520 from the National Science Foundation to Robert R. Sokal. The authors appreciate the computational assistance of Ronald L. Bartcher, David R. Fisher, and John R. L. Kishpaugh. Some of this work was carried out while K. R. Gabriel was a visiting professor at Princeton University.

REFERENCES

Busacker, R. G., and T. L. Saaty. 1965. Finite graphs and networks. McGraw-Hill, New York. 294 p.

Ehrlich, P. R., and R. W. Holm. 1962. Patterns and populations. Science, 137:652–657.

Fisher, D. R. 1968. A study of faunal resemblance using numerical taxonomy and factor analysis. Systematic Zoology, 17:48–63.

Gabriel, K. R. 1964. A procedure for testing the homogeneity of all sets of means in analysis of variance. Biometrics, 20:459–477.

Gabriel, K. R. 1966. Simultaneous test procedures for multiple comparisons on categorical data. J. Amer. Statist. Assoc., 61: 1081–1096.

Gabriel, K. R. 1968. Simultaneous test procedures in multivariate analysis of variance. Biometrika, 55:315–331.

Gabriel, K. R. 1969a. Simultaneous test procedures—some theory of multiple comparisons. Ann. Math. Statist., 40:224–250.

Gabriel, K. R. 1969b. A comparison of some methods of simultaneous inference in MANOVA. Second Symposium on Multivariate Analysis, Dayton, Ohio: *in press.*

Gabriel, K. R., and P. K. Sen. 1968. Simultaneous test procedures for one–way ANOVA and MANOVA based on rank scores. Sankhya A, 30:303–312.

Garrison, W. L., and D. F. Marble. 1961. The structure of transportation networks. Unpublished report for the U.S. Army Transportation Res. Command by the Transportation Center at Northwestern University.

Harbaugh, J. W., and D. F. Merriam. 1968. Computer applications in stratographic analysis. John Wiley & Sons, New York. 282 p.

Heryford, N. N., J. R. L. Kishpaugh, and R. R. Sokal. 1969. Seasonal and annual variation in *Pemphigus populi-transversus.* (In preparation).

Jolicoeur, P. 1959. Multivariate geographical variation in the wolf *Canis lupus* L. Evolution, 13:283–299.

Kansky, K. J. 1963. Structure of transportation networks. University of Chicago, Depart. Geography, Research Paper No. 84. 155 p.

Krumbein, W. C., and F. A. Graybill. 1965. An introduction to statistical models in geology. McGraw-Hill, New York. 475 p.

Marcus, L. F., and J. H. Vandermeer. 1966. Regional trends in geographic variation. Syst. Zool., 15:1–13.

Miller, P. G. 1966. Simultaneous statistical inference. McGraw-Hill, New York. 256 p.

Mourant, A. E., A. C. Kopeć, and K. Domaniewska-Sobczak. 1958. The ABO blood groups. Charles C. Thomas, Springfield, Ill. 276 p.

Neely, P. M. 1969. Neighborhood limited classification: a graph theoretic approach to cluster analysis (In preparation).

Power, D. M. 1970. Geographic variation of red-winged blackbirds in central North America. Univ. Kansas Publs. Mus. Nat. Hist.: *in press.*

Rinkel, R. C. 1965. Microgeographic variation and co-variation in *Pemphigus populi-transversus.* Univ. Kansas Sci. Bull., 46:167–200.

Rohlf, F. J., and R. R. Sokal. 1969. Statistical tables. W. H. Freeman and Co., San Francisco and London. 253 pp.

Scheffé, H. 1963. A method for judging all contrasts in the analysis of variance. Biometrika, 40:87–104.

Sen, P. K. 1966. On nonparametric simultaneous confidence regions and tests for the one criterion analysis of variance problem. Ann. Inst. Statist. Math., 18:319–336.

Sokal, R. R. 1965. Statistical methods in systematics. Biol. Rev. (Cambridge), 40:337–391.

Sokal, R. R., and R. C. Rinkel. 1963. Geographic variation of alate *Pemphigus populi-transversus* in eastern North America. Univ. Kansas Sci. Bull., 44:467–507.

Sokal, R. R., and F. J. Rohlf. 1969. Biometry. W. H. Freeman and Co., San Francisco and London. 776 pp.

Sokal, R. R., and P. H. A. Sneath. 1963. Principles of numerical taxonomy. W. H. Freeman and Co., San Francisco and London. 359 p.

Sokal, R. R., and P. A. Thomas. 1965. Geographic variation of *Pemphigus populi-transversus* in eastern North America: stem mothers and new data on alates. Univ. Kansas Sci. Bull., 46:201–251.

Steel, R. G. D. 1961. Some rank sum multiple comparisons tests. Biometrics, 17:539–552.

Thomas, P. A. 1968a. Geographic variation of the rabbit tick, *Haemaphysalis leporispalustris*, in North America. Univ. Kansas Sci. Bull., 47: 787–828.

Thomas, P. A. 1968b. Variation and covariation in characters of the rabbit tick, *Haemaphysalis leporispalustris*. Univ. Kansas Sci. Bull., 47: 829–862.

Tukey, J. W. 1951. Quick and dirty methods in statistics, part II, simple analyses for standard designs. Proc. Fifth Annual Convention, Am. Soc. for Quality Control: 189–197.

Department of Statistics, The Hebrew University, Jerusalem, Israel, and Department of Biological Sciences, State University of New York, Stony Brook, New York 11790.