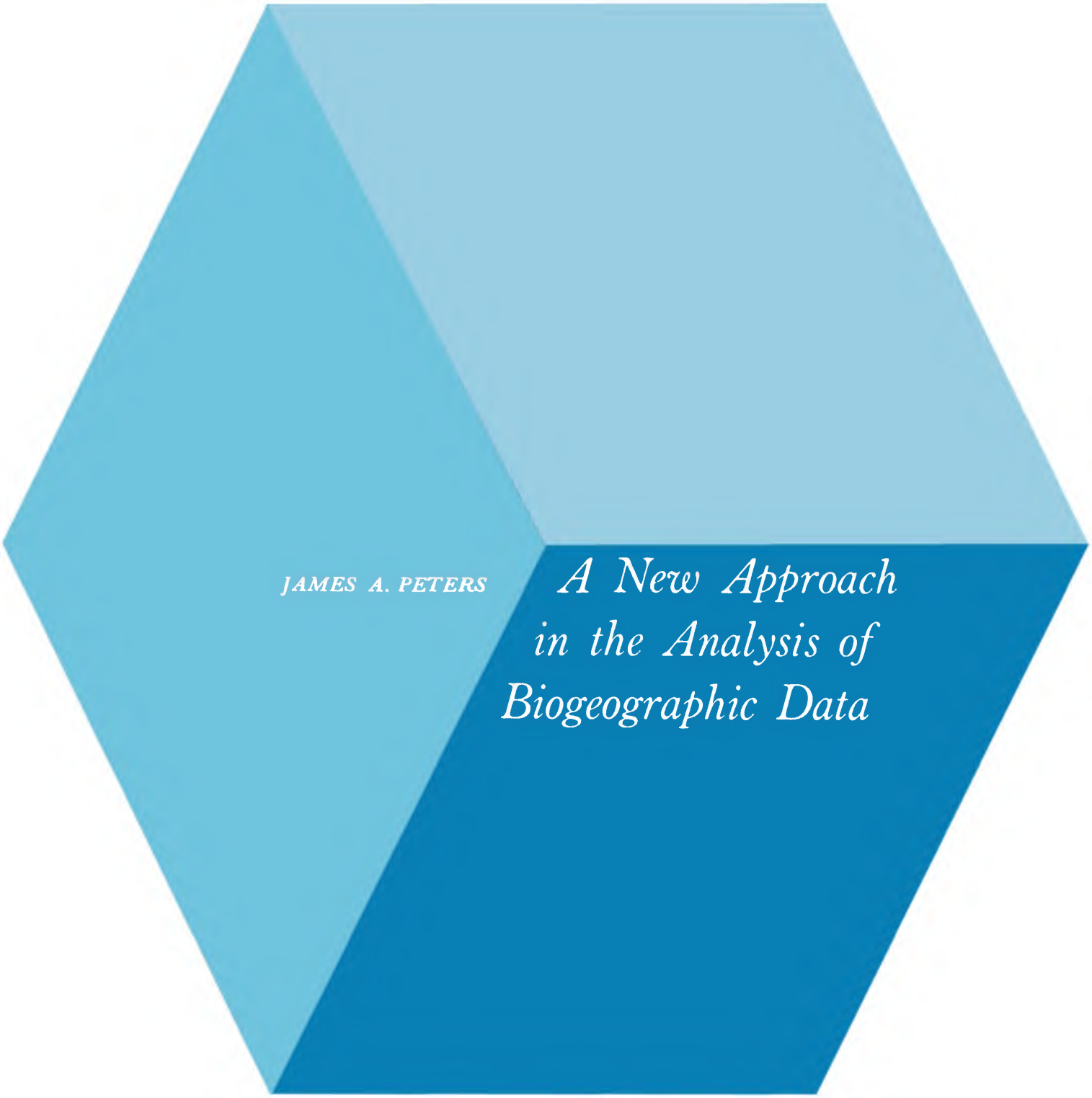


JAMES A. PETERS

A New Approach in the Analysis of Biogeographic Data

SMITHSONIAN CONTRIBUTIONS TO ZOOLOGY NUMBER 107

SERIAL PUBLICATIONS OF THE SMITHSONIAN INSTITUTION

The emphasis upon publications as a means of diffusing knowledge was expressed by the first Secretary of the Smithsonian Institution. In his formal plan for the Institution, Joseph Henry articulated a program that included the following statement: "It is proposed to publish a series of reports, giving an account of the new discoveries in science, and of the changes made from year to year in all branches of knowledge." This keynote of basic research has been adhered to over the years in the issuance of thousands of titles in serial publications under the Smithsonian imprint, commencing with *Smithsonian Contributions to Knowledge* in 1848 and continuing with the following active series:

Smithsonian Annals of Flight
Smithsonian Contributions to Anthropology
Smithsonian Contributions to Astrophysics
Smithsonian Contributions to Botany
Smithsonian Contributions to the Earth Sciences
Smithsonian Contributions to Paleobiology
Smithsonian Contributions to Zoology
Smithsonian Studies in History and Technology

In these series, the Institution publishes original articles and monographs dealing with the research and collections of its several museums and offices and of professional colleagues at other institutions of learning. These papers report newly acquired facts, synoptic interpretations of data, or original theory in specialized fields. These publications are distributed by mailing lists to libraries, laboratories, and other interested institutions and specialists throughout the world. Individual copies may be obtained from the Smithsonian Institution Press as long as stocks are available.

S. Dillon Ripley
Secretary
Smithsonian Institution

SMITHSONIAN CONTRIBUTIONS TO ZOOLOGY

NUMBER 107

James A. Peters

A New Approach in the Analysis of Biogeographic Data

SMITHSONIAN INSTITUTION PRESS
CITY OF WASHINGTON
1971

ABSTRACT

Peters, James A. A New Approach in the Analysis of Biogeographic Data. *Smithsonian Contributions to Zoology,* number 107, 28 pages, 16 figures, 1971.—Similarity coefficients calculated to show relationships between a number of geographic units are ranked for each unit, in order of descending values, and these ranked columns are then compared for all possible pairs of localities. The number of differences or discrepancies in position indicate close biotic relationship if low, or existence of a biotic barrier if high. The ranking of similarity coefficients permits use of all relationships known for a single locality. Earlier work established closeness of relationship on a single calculation—the highest known similarity coefficient. Tests of the method were run on fish collections from transects in the Atlantic and Pacific oceans, and on herpetological collections from Puerto Rico and neighboring islands. The technique is described and illustrated. The inadequacies of the numerical taxonomic techniques in dealing with biogeographic data are pointed out. The computer program used to analyze the data is included.

Official publication date is handstamped in a limited number of initial copies and is recorded in the Institution's annual report, Smithsonian Year.

UNITED STATES GOVERNMENT PRINTING OFFICE
WASHINGTON : 1971

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402 - Price 40 cents (paper cover)

Stock Number 4700–0150

James A. Peters

A New Approach in the Analysis of Biogeographic Data

Introduction

Biogeographers have always recognized the existence of areas of the world that can be characterized by the presence of certain taxa and the absence of others. The numbers of taxa that need to be present or absent, the kinds of organisms that are used as the basis for presence-absence data, and the interpretation of the significance of the groupings are matters of considerable debate and differences between biogeographers. Usually, there has been a rough correlation between the size of the area and the taxonomic level of the organisms used in its definition. The largest biogeographic area is the realm, which is usually characterized by large taxonomic groups. For example, the Australian Realm is distinguished by the diversity and adaptive radiation of the mammalian order Marsupialia, as well as by the occurrence of the subclass Prototheria. The smallest biogeographical area, often called the "biotic province," has usually been defined on the basis of presence or absence of taxa such as species and subspecies, although genera and even families may be either characteristic of or restricted to a particular province.

Part of the difficulty in working with the biogeographic division and subdivision of the world has been the nature of the data. Biogeographic information tends to come in overwhelming amounts, and workers usually restrict themselves to a workable amount of it, either taxonomically or through a geographic restriction. Even so, the details of distribution patterns of individual taxa can represent almost uncontrollable magnitudes of data, and short cuts are attempted even when they recognizably obscure the data. The advent of computers has begun to break this kind of logjam, and recent years have seen the first steps toward quantification of the data to permit full utilization.

Another difficulty for biogeographers has arisen from the revolving nature of biogeographical input versus output. There has been a tendency to accept earlier conformations of units, against which new data are tested. The new data, if they fit somewhat reasonably close to the earlier conformation, are considered as in agreement with it. Thereafter, the new data are used as additional demonstrations of the validity of the earlier conformation. A taxonomist who examines particularly carefully material of a species collected in a biotic area from which it was not previously known, in order to discern nomenclatorially recognizable subspecies, is clearly prone to circular reasoning. If he does give taxonomic recognition to the material, it immediately becomes significant in the definition of the biotic area (Peters, 1955:27). This requires, perhaps, a rather outdated approach to the recognition of taxa, but nonetheless must be recognized as a danger in the analysis of biogeographic data.

These difficulties have pushed authors to attempt to handle biogeographic data in a quantitative manner. The earliest work involved little more than a calculation of a percentage of taxa shared or not shared by two areas. The inability to demonstrate clear-cut relationships using these percentages has been recognized by most workers. Hobart Smith (1949:227) anticipated the trend of much of the work of the following twenty years:

James A. Peters, Department of Vertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington, D.C. 20560

> Standardization of biotic areas would be extremely difficult by precise boundary analysis on the basis of amount of change in fauna per given linear unit (as for instance 1 kilometer). A more reasonable method of standardization would be the arbitrary establishment of limits of faunistic distinctiveness of various subdivisions—including biotic provinces—of major regions.

He did not suggest how this was to be done, although he wrote that

> Formulation of such a rule is not the work for any one man.

He proposed that a committee of the Ecological Society of America do the task. While that society did not see fit to establish a committee, and the work has not been done as a group function, it is interesting to note that both of the methods of standardization suggested by Smith have been used, and that individuals have not hesitated to embark on the formulation of the rules Smith felt too much for "any one man." The work of Webb, Huheey, and Hagmeier, described in detail below, has involved the first alternative, which Smith felt would be "extremely difficult." The computer has facilitated such effort a great deal, and has almost eliminated the difficulties. The advent of numerical taxonomy and the construction of phenograms using its techniques has made it possible for authors such as Hagmeier, Holloway and Jardine, and Kikkawa and Pearse to establish phenotypic levels, based on cluster analysis, at which regions, subregions, provinces, and so on are defined.

Review of the Use of Quantitative Methods in Biogeography

The first steps toward quantitative analysis were based on the calculation of various measures of faunal resemblance between two samples. G. G. Simpson (1960) summarized the earlier work and discussed the variable aspects of the different coefficients that had been used. Peters (1968) provided a computer program that would calculate any chosen similarity coefficient for the various combinations of a series of localities, and Cheetham and Hazel (1969) demonstrated the identity of many of the coefficients given by Simpson with those used by numerical taxonomists in their construction of similarity matrices. All additional work in biogeography uses the calculation of similarity coefficients as the building blocks for more detailed analysis.

A simple technique for using similarity coefficients in the determination of natural areas was suggested by Webb (1950). He prepared lists of mammals and of snakes occurring at a series of sample points arbitrarily set at intervals of 100 miles criss-crossing the state of Texas. Similarity coefficients were calculated for neighboring points only, and the value was put on a map between the two sample points used in the calculation of the coefficient. Lines were then drawn on the map connecting areas of equal similarity values, as is done in constructing a contour map. Webb's method for determining what species were found at a particular sample point was to use overlay maps showing known distributions for the taxa involved, but actual collections at a series of points could be used as easily. Webb's method has a particular value in that it demonstrates quite clearly the existence of broad transitional zones between the areas of high similarity values.

A technique very similar to that of Webb was devised by Huheey (1965). Huheey felt that one of the major difficulties with the method used by Webb was that "he measured similarity instead of divergence, which led him to abandon attempts to use more refined grids and thus improve his accuracy." Huheey measured the difference or divergence between any two samples, which he defined as "the number of taxa occurring in either (but not both) of two areas under consideration, divided by the total number of taxa in both." Huheey argued that the use of a similarity coefficient interfered with narrowing the limits between sample points, because the similarity coefficient simply increased, while the use of a divergence coefficient permitted more and more sharp distinction between points as the limits were narrowed. One might argue that all that the use of divergence coefficients permits is perhaps easier recognition of low values, since they are identical with similarity coefficients which are subtracted from the value 1.00. In any event, Huheey found differences easier to work with than similarities. Instead of basing his work on occurrence of a taxon at a specific point, he used a grid of 11×20 units, each of which enclosed approximately 400 square miles. As did Webb, Huheey used overlays of known distributions and included a taxon as part of the fauna of a square if it covered more than half thereof. He then calculated his "divergence factor" for each square and its four neighbors, and took the average of the four coef-

ficients as the value of the divergence factor for that square. Again as did Webb, he then connected points of equal divergence factors with lines he termed "isometabases." Using this technique, one will find that areas where barriers exist against species movement will appear on the map as higher areas, and areas of considerable similarity will appear as valleys. Huheey found a great deal of agreement between the map of faunal areas of the state of Illinois, drawn after using his technique, and earlier maps derived through the use of nonquantitative techniques.

One of the first biogeographic analyses to use the techniques of numerical taxonomy was done by Hagmeier and Stults (1964). Their methods were similar to those of Webb and Huheey. They started with 745 sample points, for which checklists of the mammalian fauna were prepared, and constructed a map showing number of species throughout North America. A grid made up of blocks 50 miles square, giving a total of 2,490 blocks, was superimposed on the first map, and the number of species whose ranges ended in each block was determined. An index of faunistic change (IFC) was computed using the formula $IFC = 100L/n$, with L the number of range limits in a block, and n the number of species occurring within the block. The result is a percentage of species which have a range limit within the block. These IFC values were entered on a map, and "isarithms" (which are the same as Huheey's "isometabases") were drawn. The result is shown in their Figure 1, with dark areas representing areas of faunal barriers, and light areas representing regions of faunal uniformity. At this point, they had completed the first stage of their analysis, in what I consider to be a successful manner. Several recent authors (Holloway and Jardine, 1968: 155; Kikkawa and Pearse, 1969:823) have pointed out that the selection of different "primary areas" may produce different results in numerical analysis, and Hagmeier himself published a second paper (1966) in which he pointed out the error in their results, arising from the selection of only 24 North American mammal areas. This problem with the results of Hagmeier and Stults lies in the second stage of their analysis, however, not the first stage, which was not changed in the second paper.

The second stage of the analysis by Hagmeier and Stults was an attempt to discover the degree of relationship that existed among a series of mammal provinces. There was a certain amount of agreement between their map of the North American continent showing IFC values and the provinces as previously described by Kendeigh (1961) and shown in their Figure 3, and they used the latter as the basis for their continued analysis. The use of Kendeigh's provinces is the error which Hagmeier (1966) was attempting to remedy. Once the provinces were delimited, however, the authors were irrevocably committed to an analysis of the relationships between those provinces, regardless of their validity as biogeographic entities. They did not use numerical taxonomic techniques in the first stage, but did so in the second stage. The provinces were subjected to cluster analysis, using the weighted pair-group method, and a phenogram was constructed, showing the degree of relationships between pairs of provinces. It should be recognized that once the primary areas are established, cluster analysis will give a stepwise phenogram of smaller and smaller levels of similarity, regardless of whether the original units are valid or not.

Hagmeier recognized the difficulty, and did much to repair it in his 1966 paper. He stated (1966:289) that the error was corrected by "laying a transparent overlay over the species IFC map . . . and drawing lines through all regions of high IFC value, delimiting ultimately a total of 86 (rather than the original 24) primary areas." The 86 areas were used to calculate similarity coefficients, and then subjected to cluster analysis. All those which clustered at values higher than 65 percent were considered to belong to a single province, and the author found that enough of them combined above that value to reduce the working number to 38. These were then used as the primary areas for the final plotting of the mammalian provinces of the continent; they were also used for further cluster analysis to indicate groupings into superprovinces, subregions, and regions.

It is possible to argue that the second paper by Hagmeier falls into the same error as the first, but I do not think this criticism valid. Hagmeier used the basic data from his IFC analysis throughout the second paper, and the provinces were not drawn from any previous author's concept, but were entirely derived from the basic distributional data. The problem caused by the use of only a single similarity coefficient, which is discussed more fully elsewhere in this paper, could not have been overcome by Hagmeier, because techniques did not exist previously

that permitted the utilization of all coefficients known for a locality, or, in the case of the work by Hagmeier, all primary areas.

Holloway and Jardine (1968) used numerical analysis in an examination of the distribution patterns of various taxa in the Indo-Australian area. They distinguished two approaches to zoogeography. The first of these is based on the grouping together of geographical units into larger areas. The dendrograms from the similarity matrices are based on pairings of areas (shown in the authors' Figures 1–3). Preston's coefficient of dissimilarity, z (Holloway and Jardine, 1968:156), was used. The dendrogram is based on the matrix itself, and no shrinking of the matrix is employed. This is the "single-link cluster analysis" method. Once the dendrogram has been constructed, levels of stepwise discrimination are drawn to distinguish zoogeographic regions, subregions, and provinces.

The second approach is based on the use of faunal elements, which are sets of taxa having similar distributional patterns (Holloway and Jardine, 1968:153, figs. 9–11). The dendrograms are then based on cluster analysis of the faunal elements. The coefficient of dissimilarity used for the calculation of the matrix is $1-m/n$, with m the number of "primary areas" (see below) in which both taxa occur, and n the total primary areas. The dendrogram is derived by single-link cluster analysis, as in the first approach. The clusters formed are designated by numbers or letters at the base of the stem of a recognizable grouping, with numbers representing faunal elements, and letters representing subelements.

Both of these approaches share three steps in the six-stage process used by Holloway and Jardine (1968:154). These are:

1. A selection of taxa of a given rank from a particular group of organisms is made.
2. Primary areas are defined.
3. The distribution of the taxa selected in (1) among the primary areas defined in (2) are tabulated.

The second step is clearly a critical one. The authors point out the difficulties that arose for Hagmeier (1966) when he revised the analysis done by Hagmeier and Stults (1964) on the distribution of North American mammals. They point out that in this case identical numerical methods produced different results, depending upon the method of selection of primary areas. They then stated (Holloway and Jardine, 1968:155) that "Our selection of primary areas for the mainland regions is arbitrary." As an indication of what they mean by "primary areas," the following list gives some of their selections: Formosa, Java, Indochina, Sula Islands, Philippines, Australia, and so on. While the authors indicate that islands such as New Guinea and Java are probably not homogeneous with regard to their faunal elements, and thus are not satisfactory for this type of analysis, "the distributional data used were not precise enough to enable us to recognize more than one primary area for each island." Since the same primary areas were used in both approaches to zoogeographic analysis by these authors, it seems clear that the choice of primary areas played a significant role in their results.

Holloway and Jardine concluded that the two approaches are complementary, and need not be deemed arbitrary in nature. They suggested that both intuitive and numerical approaches for grouping primary areas into zoogeographic regions are in themselves mainly of use for descriptive purposes (1968: 186).

In a numerical analysis of the distribution of the land birds of Australia, Kikkawa and Pearse (1969) pointed out the difficulties attendant upon the use of primary areas by earlier authors, and based their work in 121 selected sites on the Australian continent, for which they prepared lists of bird taxa. They indicated that "ideally, the areas of similar fauna should be identified as primary areas on a grid system," but the information available to them was not accurate enough "to delimit small primary areas contiguously over the whole continent" of Australia (1969:823). This led them to select their sites in a manner that permitted them to "minimize ecological factors contributing to the resultant classification of faunal regions." To achieve this, all niches available at a site must be included in the sample. The authors (1969:824) felt "if too many sites are clustered in a small region or too few scattered in a large region, the relative importance of each region may become unreliable though the resultant boundaries might not be altered in the analysis." This seems to be a consequence of the methods used for the analysis of the data, since ordinarily a large number of well-sampled sites or localities within a single homogeneous faunal unit would be expected to show high

levels of relationship between each other, and uniformly low levels of relationship with other faunal units.

Kikkawa and Pearse used the technique of "divisive information analysis" to analyze their data. This was necessary because of the dimensions of the data matrix, which included 121 sites and 464 species. They used a technique designed to permit combining of sites into groups, and then the heterogeneity of each group was calculated, using the "conventional Shannon-Weiner information content." I do not clearly follow the technique used nor the explanation given by the authors, and as a consequence have not been able to use it in my work. But a final remark by the authors to the effect that "the program termination was specified arbitrarily to produce 20 groups for the classification of sites, and 30 and 20 groups for the classifications of species and genera respectively" (1969:825) leads me to feel that there must be a loss of information content through the use of their technique. Since they start their final analysis of classification of sites into faunal areas with only 20 groups of sites, and these 20 groups of sites are then divided into 11 faunal areas (1969:826, fig. 1), it would appear that the arbitrary nature of selection of the 20 groups is reflected in a fairly arbitrary set of faunal areas.

The authors construct dendrograms to show relationships between faunal areas, and they plot boundaries between groups of classified sites "in such a way as to contain all sites of each group within the same confines" (1969:825). All of their maps and dendrograms are based on their calculated "information statistic," but, unfortunately, they do not present any of the data in a way to permit one to follow the technique. Even though the reader of the paper by Kikkawa and Pearse may be frustrated by the omission of information that he needs to understand the interpretations presented, it is still significant to note that this is the first paper in which numerical methods are used with individual sites or localities as the basic unit for analysis, and that the results closely conform to those derived by intuitive methods.

A New Technique for Biogeographic Analysis

In the past, authors have spent many laborious hours calculating similarity coefficients of various sorts for all combinations of any two geographic units under study, whether they be localities, primary areas, biotic provinces, or something else. At first, the calculated coefficients were used as the basis for decision as to the degree of similarity between two units. More recently these values have been used as the matrix data for cluster analysis, using the techniques of numerical taxonomy. Anyone who has gone through either of these processes is aware of the amount of time one can devote to the calculation of the coefficients, and the computer is almost an indispensable prerequisite for cluster analysis, so it is not surprising that biogeographers continue to work with a series of coefficients, each of which expresses the relationship between any two units. There is, however, a large amount of additional information concerning each of these units that is currently not used—the data on the relationship of each unit with all the other units involved in the analysis. The basic concept underlying the new technique described here for estimating the degree of similarity between any two localities is that one should take advantage of all the available data if it is physically possible to do so.

It should also be pointed out that a serious handicap to the utilization of the value found by direct comparison of the data for any two units is that the result can be highly biased as a consequence of dissimilar levels of adequacy of our information concerning the biota of each unit. The absence of a taxon from a particular unit may be a true reflection of the actual situation, but it can also simply mean that field activity within that unit has not yet produced a voucher specimen. Only if all field samples are made by the same individual or group can one be sure that each is made with the same degree of adequacy and thoroughness, under identical conditions, for the same length of time, and equivalency of all other conditions that can affect validity of a sample. There are few, if any, occasions when a biogeographer works with that kind of rigidly controlled data.

If, however, one were to establish a ranked list for each locality, based on the similarity coefficients calculated between that locality and all others, there would be a greater likelihood that position within the ranking would express biotic similarity than would the actual value of any one similarity coefficient. One might assume here that inadequacy of sampling is basically random and would be reflected in all coefficients calculated for the relationship of

the locality with other localities. While the coefficient expressing the relationship with the most closely related locality might be low because of the inadequacy of sampling, one would expect closeness of relationship to be indicated by a high position in a rank of all localities based on coefficients, because less closely related localities should predictably have lower coefficients.

It should be added that if one accepts the assumption of inadequacy of sampling as a basically random phenomenon, one must also recognize that the ranking itself can easily be affected by that randomness. The ranking of the entire series of localities in a diminishing series based upon their similarity coefficients with any one locality will not produce an unchallengable sequence demonstrating diminishing degree of relationship. It merely indicates approximate position, and presence or absence of a comparatively small number of taxa can elevate or reduce the level of a specific locality to some degree. The advantage of the ranking by similarity coefficient values is that one is then using all available information about the relationships of a locality, rather than only a single expression of relationship.

Once one has a diminishing rank of similarity coefficients for every unit in his series, it is possible to go on to the next step in the analysis. The assumption can be made that if two localities are found within a single biotic unit (i.e., the taxa collected at each can be expected to be very similar), the rankings of all the other localities under each of them should also be very similar. If the rankings were identical, one might feel safe in saying that the taxa living at each were also identical, and that the two samples had been drawn from the same fauna, flora, or biota. Were one to find that some localities formed a group within which they were very similar to each other in rank comparisons and very different from all localities in another such group, the conclusion that some kind of biotic barrier exists between these two groups would be indicated.

The situation described above is shown in Figure 1. In this hypothetical area, the investigator has collected samples at each of the localities one through six. He has counted the number of taxa found at each locality, and has recorded the number of taxa shared between any two localities. These data have been used to calculate a similarity coefficient for each pair of localities. This coefficient is shown within a line connecting the two localities. Localities one through three are shown as included within a single biotic unit, and the similarity coefficients are high, as expected. The coefficients for these localities when compared with localities four through six are very low, of course, because of the existence of a biotic barrier interfering with free movement of members of a species. All coefficients calculated for each locality are now arranged in a descending rank. The coefficient value is associated with the locality number to permit its identification. There will be five coefficients for each locality. In any comparison of ranks only four coefficients will be useful. This is true because each column in a pair includes a coefficient based upon its relationship with the other. This coefficient cannot be compared with the other column because there is no equivalent value. This is shown in Figure 2. In this figure the actual data from Figure 1 have been ranked under the numbers

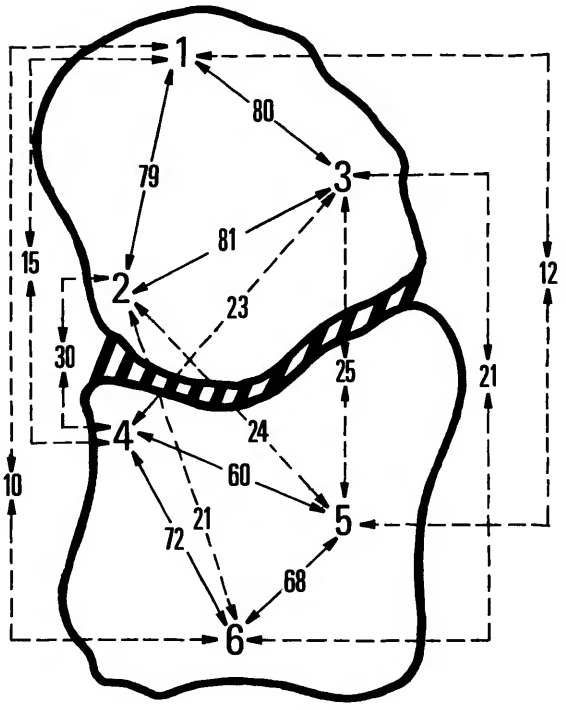


FIGURE 1.—A hypothetical situation with three localities in each of two biotic units. Solid arrows connect localities within a single unit, dashed arrows connect localities in different units. Smaller numbers within arrows are similarity coefficients.

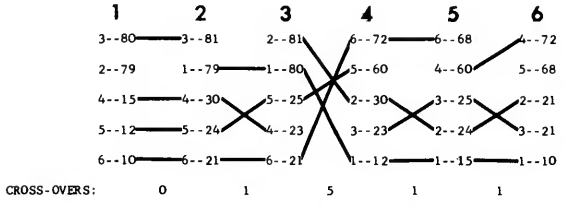


FIGURE 2.—The localities in Figure 1 rearranged to show relationships. Similarity coefficients involving locality one are arranged in descending order beneath it, with second locality number associated with each by dashed line. Solid lines across the figure connect coefficients for a single locality.

assigned each locality, and lines have been drawn from one column to the next to connect the series of coefficients associated with one locality. The intersection of a pair of lines will indicate a shift in position of two localities within the columns connected by the lines. As long as relative positions are the same or similar, the number of intersections should be small, but any comparisons between localities representing different biotic units should show considerable rearrangement of position and, as a consequence, should greatly increase the number of possible line intersections. Sokal and Rohlf (1969: 535) demonstrated that this technique of connecting similar points and counting the intersections can be used as the basis for the calculation of a Kendall Coefficient of rank correlation, although in that particular instance one of the two columns is regarded as a standard against which the other is compared (see elsewhere in this paper for a discussion of this specific problem), and the Kendall Coefficient must be recalculated for each pair of columns. The graphic method of presentation shown in Figure 2 is more satisfactory for presentation of the overall picture, because the area of biotic discontinuity is immediately obvious without the calculation of Kendall Coefficients. There is nothing sacrosanct about the sequence in which the columns are presented, of course, and the arrangement from one to six from left to right can be revised as much as one wishes. The object of revision or switching about of the columns would be to minimize the number of intersections between columns, and we shall see later that this becomes of primary importance when one does not have preconceived notions of the relationships between the localities being investigated.

In the new method proposed here for analysis of biogeographical data, the basic hypothesis is this: if one uses as much as possible of the data available concerning relationships between units, the conclusions drawn from the analysis will be more informative and more reliable. The technique used is to rank all similarity coefficients for any one unit in descending order, and then compare, either visually or mathematically, the derived rank order with all other similarly derived rank orders.

Between Column Discrepancies

If one wishes to give a visual evaluation of relationships among a series of localities, the technique described above, in which the ranked columns are listed and with lines connecting the same locality throughout the sequence of columns, is quite satisfactory. The presentation of the data in that fashion also facilitates the tabulation of the number of crossovers that have taken place. If, however, the number of localities to be used is quite large, and there is no basis for a sequential arrangement of those localities, it becomes difficult to construct the figure and to count the crossovers. For such analyses, I have developed a technique of tabulating the discrepancies that exist between any two columns. If two columns are identical in the sequence of ranked localities, there will be no crossovers, all lines will be horizontal and parallel, and there will be no discrepancies. If any two adjacent localities were to have their relative positions switched, there would be one crossover shown in the figure, and there would be two displacements when the columns were compared. This leads one to think that the number of crossovers should be half the number of displacements, so one could arrive at the number of crossovers without plotting the figure simply by counting the displacements and dividing the count by two. This does not work, because the minimum number of displacements per crossover is two, but that is not the maximum possible number. The displacement figure is in itself an accurate measure of the total difference that exists between two columns, however, and it replaces crossovers as a measure in a run of the computer program JPFRF. Crossovers can only be tabulated by making a figure of the data. Displacement data can indicate the best possible sequence for the columns in a figure, and such a figure will give the minimum crossover count if drawn.

It is, I think, safe to say that this technique of counting discrepancies is physically impossible if the number of localities involved in the analysis is much greater than about 20. In a run of data for 42 localities, I set up the program to print out the ranked pair of columns for each combination, and then attempted to count the number of discrepancies directly from the printout. I found that my accuracy was low, and that each comparison required about thirty minutes. Since 40 localities would involve about 800 such comparisons, it is clearly not feasible to try to do the task by hand. Once the computer program had been written to do the same task, however, all 800 comparisons could be done in less than the time it took me to do one, and could be printed out in a matrix form to permit direct comparison and evaluation.

The total number of discrepancies or differences in position between two columns representing any two localities is recorded by the computer program JPFRF, which first ranks the similarity coefficients calculated for any one locality in descending order, and associates such coefficients with the identification number of the other locality used in its calculation. The technique used associates the identity number with the similarity coefficient as a decimal fraction. Thus, a value in the ranked data for the fourth column might be 75.25, which would mean that the similarity coefficient for columns 4 and 25 is 75. In the ranked data for column 25 there would be a value of 75.04, which has the same meaning. In a comparison of columns 4 and 25, this particular similarity coefficient would not be used in the tabulation of the total discrepancies, but it would be used in any comparison of either of these columns with any other column.

The computer selects two columns to be compared and first reads the uppermost figure in each (the procedure can be followed in Figures 4 and 5). Continuing our example, these would be 100.04 and 100.25, for columns 4 and 25. The computer examines the decimal figure of both and ignores each in further examination of the columns. Next it reads the first ranked value in column 4, notes its position in the column, and separates it into its two component parts, the similarity coefficient and the locality number. Then it searches the other column, which is 25 in our example, for the same locality, and once having found it, notes its position. If the position is the same as for that locality in column 4, it records no discrepancies, and proceeds to the next value in column 4. If the position is different, it counts the number of discrepancies in position. Then it compares coefficients to see if any ties exist, or, if a deviation is permitted, to move above or below coefficients within the tie or the permitted deviation. This movement within ties or the limits of permitted deviation is always such so as to minimize the total discrepancies. The total number is recorded, after the entire two columns have been examined, either as a direct readout of the two localities and the value, or as an entry into a data matrix.

There are several reasons for using this discrepancy value rather than one of the several mathematical coefficients that are available for the comparison of two columns of ranked data. The most commonly used of these are the Kendall Coefficient and the Spearman Coefficient. Either could be used with the data calculated in this program, but there are some problems in doing so. First, both of these coefficients are based on the assumption that the two sets of ranked data have something directly in common. In the example given by Simpson, Roe, and Lewontin (1960:255), the ranking is between the measurements of the first and second molar teeth in a series of fossil mammals. In the example used by Sokal and Rohlf (1969:533), the authors compared the total length of aphid stem mothers and the mean thorax length of their parthenogenetic offspring. The Kendall Coefficient is often used by educational psychologists when comparing the ranks of a group of students as prepared by two different instructors, or the ranks of students on two different tests. In all such cases, one ranking can be held constant, and the position of the same individual or source in a second ranking relative to the constant rank can be determined. There is, however, no such obvious basis for determination of a "constant rank" in the biogeographic data, since both columns are based on a set of similarity coefficients calculated on overlapping but not directly correlated data.

The second minor problem encountered in the use of the data from this program is that the localities are completely reshuffled in every ranking. This makes it impossible to use a Kendall Coefficient, which requires the tabulation of "subsequent ranks greater than the pivotal rank R" (Sokal and Rohlf, 1969:534), unless the locality identifiers are all

changed to actual rank, and then in the second column all the new ranks are identified with the original numbers. Since, however, the data worked with in this biogeographic problem do not bear a direct relationship, there is little logical basis for rearrangement of either column to check its agreement or divergence from the other. As it works out, there is no necessity for any rearrangement of the columns in this way. It can be demonstrated that, if either column is rearranged or renumbered in such a way as to give the ranks numbers from one to N, and the other column renumbered in the same way to agree with the first, there is no difference in the total number of discrepancies from what would be tabulated from the original columns of figures. It has, however, led me to formulate a slightly different coefficient, which will use the discrepancy data without any need to renumber the units in the column. This coefficient will be discussed below.

A third problem in working with these data arises from the occurrence of "ties." This term is used by previous workers when the values of more than one variable are identical, which means they must have the same rank. It is mathematically accurate, and customary, to average all the ranks covered by these identical variables, and then to give each of them the value of the mean of the ranks. In the biogeographical data, however, the tie cannot be given this degree of significance. There are too many variables involved in the preliminary data to permit me to consider two localities identical in their relationships because the similarity coefficient for each of them with a third locality happens to be the same. I have, therefore, given all tied localities complete freedom of movement within the span of tied ranks, and, as discussed above, the computer will move any locality up or down, within the limits of a tie, to minimize the number of discrepancies counted for that locality. Again, since this minimization takes place in every comparison for all localities, it cannot affect the final result to any great degree.

A New Rank Correlation Coefficient

These problems with previously used rank correlation coefficients have led me to devise a slightly different coefficient, which will enable one to compare results from different analyses. It is based on the relationship of the number of discrepancies to the number of localities involved in the study.

If there are no differences between the two columns of ranked coefficients, there will be no discrepancies in position of localities, and the discrepancy count will be zero. If one deliberately arranges the columns to maximize the total discrepancies, the square of the number of localities will be exactly twice the number of discrepancies, if there is an even number of localities, and the number squared will be twice the number of discrepancies plus one, if there is an odd number of localities. In a formula, this would be:

$$2D = N^2 \text{ [If N is even]}$$

$$2D + 1 = N^2 \text{ [If N is odd]}$$

It should be recognized that the maximum number of discrepancies will occur when one column is exactly the reverse of the other. In such a case, the top ranked value in the first column is the terminal value at the bottom of the second column, and the last value in the first column is the top value in the second. This could make a difference in the formulation of a coefficient for the analysis of biogeographical data, because we need to decide whether we wish to maximize the significance of the total number of differences, or if we want to put more emphasis on the fact that what appears to be developing is a negative correlation.

Of the two alternatives, I prefer the first, which will maximize the number of differences. This is because we are not actually trying to say that a "correlation" exists between any two localities. We are merely trying to discover which localities should be included in a recognized biogeographic unit. If any two localities show a number of discrepancies approaching the maximum number possible, we do not regard them as "negatively correlated," and we would simply indicate that they are very different in the taxa which inhabit them.

The following formula will give a range of values from zero to one, with zero indicating no similarity between two localities, and one indicating complete agreement between two localities:

$$1 - \frac{2D}{N^2}$$

The minor difference between even and odd numbers of localities is ignored in this formula, although it

can change a value quite a bit if the number of localities is small. It should be noted that all values calculated for a specific set or group of localities will be equally affected by the same factor, since all will use the same denominator in the equation.

Test of the Hypothesis

The best test of a hypothesis is to see what happens when it is applied to a real situation. For a test run, I selected the data on fishes from the faunal transect made across the Atlantic Ocean from the southeast to the northwest, on cruise 17 by the R/V *Chain,* Woods Hole Oceanographic Institution, in 1961. The mesopelagic fish collections have been reported on by Backus, Mead, Haedrich, and Ebeling (1965), a group hereafter referred to as the "Chain Gang." The Chain Gang report is particularly satisfactory because: (1) it includes a sufficient number of localities well separated geographically, (2) the authors have identified the fish fauna well enough to use presence and absence of various taxa without fear of difficulties from incomplete taxonomic data, (3) the transect was linear, and (4) the data analysis presented by the authors indicated that one major biotic barrier was crossed during the transect. My analysis is restricted to the series of localities numbered from 800 to 813 (Figure 3). There were 116 species of fishes taken at these 14 localities. The Chain Gang analyzed the same series of localities, reducing the number of species to 44 by removing all but two of those known from only one locality (a total of 50 species), and all but two of those known from only two localities (a total of 22 species).

The data were analyzed using the computer program JPFRF, which is discussed in detail later in this paper. The first run maintained the individual identities of all localities, so there are fourteen columns of ranked similarity coefficients to be compared. Since the sequence of localities was linear, there appeared to be no need for rearrangement, and each locality was compared only with its immediate neighbors. Many of the different similarity coefficients described in my earlier paper (Peters, 1968) were used for running the data. Various allowances for errors were made, as well, to see how widely this would affect the result. Figure 4 shows the results of a preliminary run on all 14 localities, using the Coefficient of Community. In this run, only those species collected at a locality were recorded as occurring there. It will be readily seen that the technique did not disclose very successfully any traces of relationships between localities, except perhaps between localities 11, 12, and 13. Practically identical results were produced if the data were tested against any other similarity coefficient.

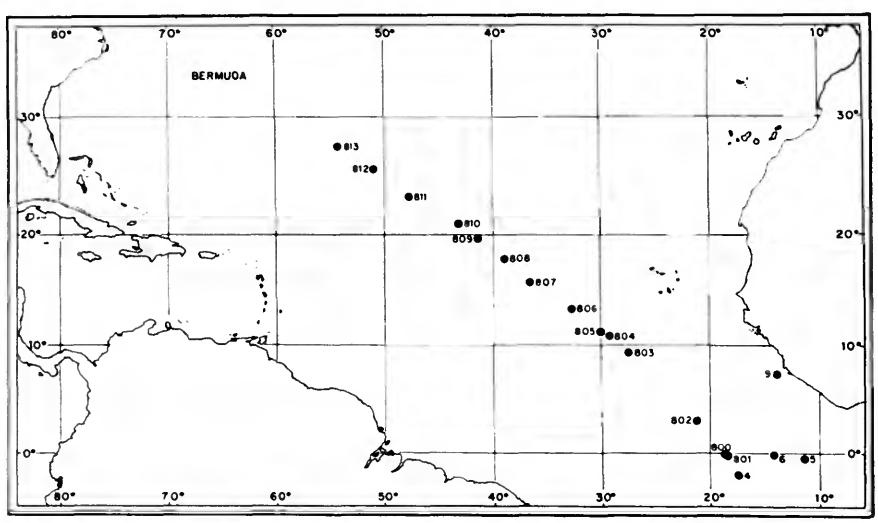


FIGURE 3.—Stations at which fishes were collected by investigators aboard R/V *Chain* (from Backus et al., 1965, fig. 3).

The Chain Gang encountered similar difficulties in the analysis of their data, and they used a simple device in their arrangement of the data that eliminated much of the problem, taking advantage of the linear nature of the collection sites. In their Table 3 (Backus et al., 1965:157), the 44 species are listed, and the number of individuals collected at each site is given. If no specimens were collected at a locality, but that locality lies between two localities where the species was collected, a dash is inserted in the table. The Chain Gang examined only the locality at which a species first appeared and the locality at which the same species disappeared from their collections, referring to this as "first-time and last-time captures" (1965:148). This is equivalent to making the assumption that, even though it may not have been taken at every collection site, a species occurs anywhere between the locality of first-time capture and that of last-time capture. In the Chain Gang table, then, any occurrence of a dash can be regarded as a locality for that particular species. It is not difficult to imagine situations where such an assumption would be dangerous, and it is perhaps not entirely valid for all species in the Chain Gang data. But, since they have analyzed their data in this way, I have done the same in order to make our results comparable. The results of an analysis with the presence of a species assumed in any locality between any two known occurrences are shown in Figure 5. This run used Simpson's Coefficient for the calculation of similarity, although the Chain Gang mentioned (1965:148) that they had attempted to use Simpson's index of faunal resemblance, "but were not successful." Figure 5 shows a distinct change taking place between localities six and eight, with a group of fairly homogeneous localities below locality six, and a second homogeneous group above locality seven.

Based on their method of calculation, the Chain Gang also found a change taking place around locality 807. They say (Backus et al., 1965:150)

> The peak in the interval between collections 807 and 808 ($X^2=3.79$) implies a faunal boundary, for here the odds are about 19 to 1 that the departure of the observed from the expected is not due simply to sampling error. The value for the interval 806–807 is also significant ($P<.10$), suggesting that the boundary is not an abrupt one.

The Chain Gang was able to correlate this faunal barrier with ecological factors. They found that there was a deepening and divergence of the 15° and 20° isotherms, representing a degradation of the thermocline and thickening of the surface isothermal layer, apparently correlated with increasing latitude. They concluded (Backus et al., 1965:152), ". . . that the faunal boundary near Collection 807 corresponds to the boundary between the South Atlantic Central Water and North Atlantic Central Water Masses."

It was a little disturbing that, in order to achieve the same results as the Chain Gang, it was necessary to assume occurrence of the species in localities from whence they had not been collected. It is an artifact of the linear nature of this cruise that makes it possible to do so, but most tests of biogeographical data are not likely to be linear, and such an assumption would have a high probability of error. At the same time, the number of species from certain of the Chain Gang localities is so small that their relationships with other localities tend to obscure the picture. This is particularly true of localities 807, with only thirteen species, and 810, with only seven species. As the map shows, localities 800 and 801 are very close together, and could easily be called a single locality, as can 804 and 805. Localities 809 and 810 can be combined equally appropriately, which leaves only the problem of locality 807. As can be seen from the map, it is quite distant from both 808 and 806, and is perhaps inappropriately combined with either. I have tested it in both combinations, however.

If the original data are used, and actual occurrence only at any locality is recorded, with presence in either of two combined localities recorded as presence in the combination, we can eliminate the effect of comparatively small samples, and also keep out the bias of assumption of occurrence where a species has not been collected. Figure 6 shows that this brings us much closer to recognition of the distinct boundary between the combination of 806 and 807 and the other more uniform areas. When 807 is combined with 808, the number of crossovers and displacements increases considerably, indicating that the 806–807 combination is more valid.

One final step can be taken in the analysis of the Chain Gang data. This involves the use of the combined localities, as well as assuming occurrence of a species in any locality intermediate between two in which the taxon is known to occur. In short, this is the maximum amount of practical manipulation

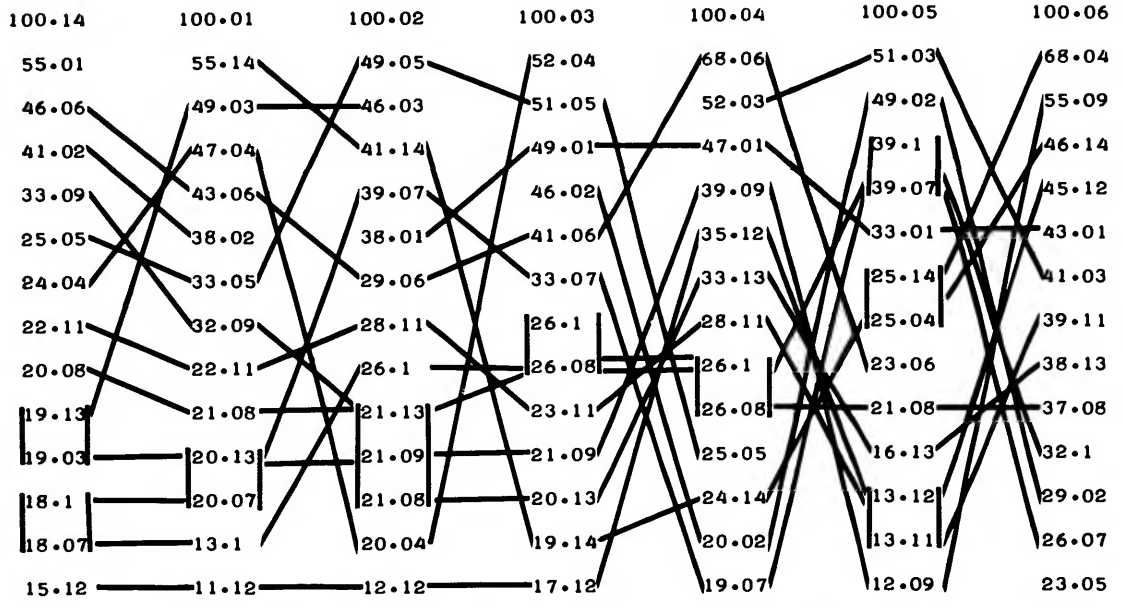


FIGURE 4.—Diagrammatic representation of data from Chain Gang collections, using all 13 stations, with no assumption of occurrence at intermediate stations (see text). The numbers represent the similarity coefficient, to left of decimal point, and the locality number, to right of decimal. A value of 100 indicates the locality dealt with in that column. Since the numbers

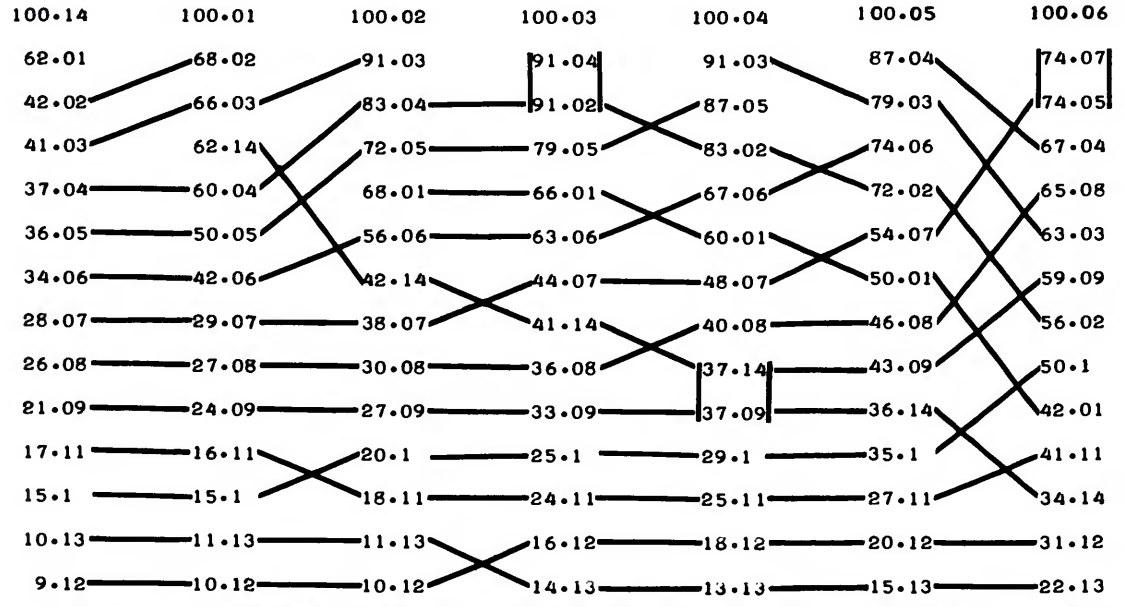
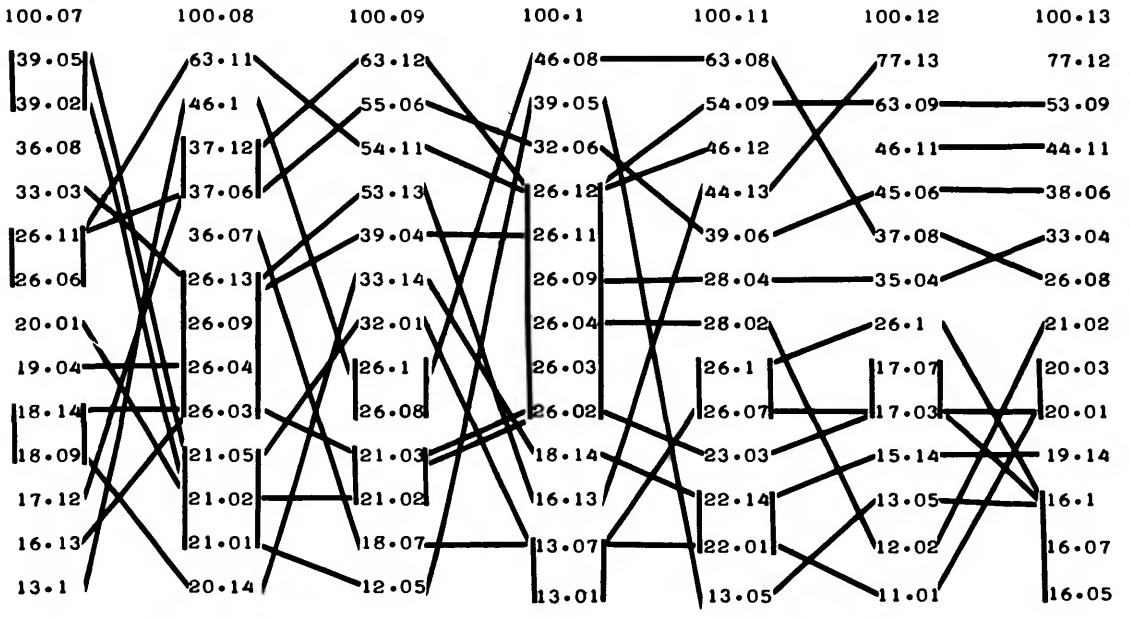
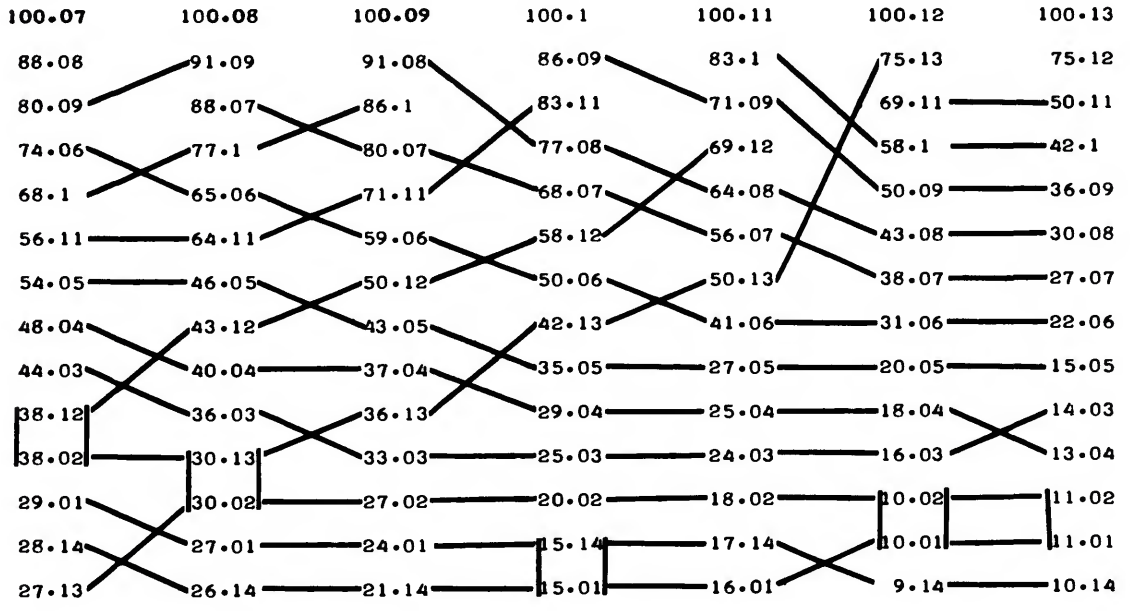


FIGURE 5.—Diagram of the relationships of all 13 Chain Gang stations, with a taxon assumed to be present at any station between two known occurrences. See Figure 4 for an explanation of numbers.



have been ranked and printed by the computer just as shown in the figure, the zero in station "10" does not print, but appears as 100.1. This refers to locality ten, not locality one, which is 100.01.



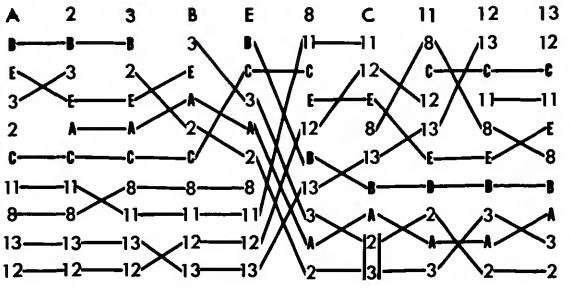


FIGURE 6.—Diagram of relationships after combining some station data. No assumptions are made concerning occurrence. Station A=800+801; B=804+805; C=809+810; and E=806+807.

that can be done with these data. Figure 7 shows the sharpness of the distinction of a faunal boundary at the combination of 806 and 807 (labeled "E" in the figure). Again, if 807 is combined with 808, the number of crossovers and displacements increases considerably.

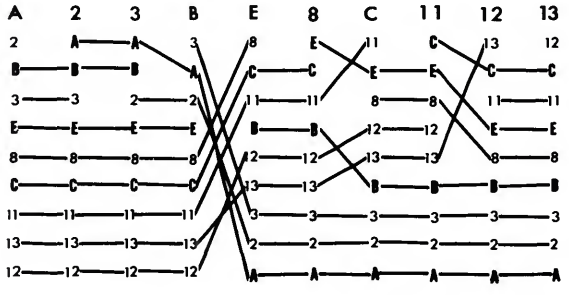


FIGURE 7.—As in Figure 6, except occurrence of a taxon assumed at any station between two known occurrences.

Additional Testing of the Method

As I pointed out above, the analysis of the Chain Gang data was simplified by the fact that the cruise transect was linear, and one did not need to be concerned about the relationships of non-neighbor localities. Such data are exceptional biologically and geographically, and it cannot be regarded as typical of the problems facing biogeographers. It is necessary to test the method against data from non-linear localities as well. The data presented by Heatwole and MacKenzie (1967:434) on the herpetogeography of Puerto Rico and the neighboring islands were easily converted for testing by the JPFRF program, because they had already calculated the faunal similarities, using Preston's coefficient, and published a table showing these values.

The data presented by Heatwole and MacKenzie deal with the "major islands of the Puerto Rican shelf" (1967:434). They included eleven islands (Figure 8). The authors draw certain conclusions based upon the values of the Preston Coefficient, making it possible to compare their results with those drawn from continued analysis of the same data. It is not possible to start with the original presence-absence data, because this is not given by the authors, nor is it possible to use a different similarity coefficient.

The data were first analyzed using the similarity coefficient values as given in the table, except that values of zero were converted to 0.01, and asterisks, which represent uncalculable values in the table, were converted to 0.99. All eleven islands in the matrix were included in the first run, which resulted in some comparatively high discrepancy values. These appeared to result from variation in the position of the island called Monito, and it was then noted that there are only two species of reptiles or amphibians known from that island, one of which is restricted to Mona and Monito, and the other found on all the other islands involved in this analysis. This is too small a sample to use successfully, and the data were run again, leaving out the similarity coefficients given for Monito Island. While this eliminated many of the distorted values, it did not change the relationships expressed by the discrepancy counts.

The results of the analysis are shown in Table 1, with the actual discrepancy counts in the lower left half of the matrix, and the correlation coefficients in the upper right half. The islands are arranged in the same sequence as in the original table in Heatwole and MacKenzie (1967:434). The great similarity between the islands, St. John, St. Thomas, and Tortola, is immediately obvious from the very low discrepancy values. Culebra also belongs in the same closely related group, although the values are slightly higher in each case. Anegada, which lies at some distance away from all the others, has intermediate values in all cases except with Culebra. Another grouping of islands is composed of Puerto Rico, Caja de Muertos, Vieques, and perhaps Mona. The latter, geographically quite isolated from the others, is nonetheless closely related to them. If the islands are

FIGURE 8.—Map showing location of islands on the Puerto Rican shelf (adapted from Heatwole and MacKenzie, 1967, fig. 1).

rearranged to put those showing low discrepancy counts close to one another, the resulting matrix is shown in Table 2. The change from low to high discrepancy values as one passes from one group to the next is obvious, and the intermediate nature of the values for Anegada is equally clear. This is interpreted to indicate that this island is not very similar to any of the others, rather than to indicate that its fauna is an intermediate one.

Perhaps the most interesting observation to emerge from this matrix concerns the island of St. Croix. As pointed out by Heatwole and MacKenzie, this island is characterized by a high degree of endemism in its fauna. There are eight species found on the island, four of which are endemic. As a consequence, in the table of similarity coefficients given by the original authors, St. Croix exhibits very low values, and is sharply set off from the other islands on the Puerto Rican shelf. This fact has long been recognized by herpetologists working on the Carribbean fauna, and it is striking to note that the current analysis, using the ranking technique, results in very low counts of discrepancies between St. Croix and the other islands—St. John, Tortola, and St. Thomas.

TABLE 1.—*Matrix showing relationships among islands on Puerto Rican shelf*

	Puerto Rico	St. Croix	Mona	Anegada	Tortola	St. John	St. Thomas	Culebra	Vieques	Caja de Muertos
Puerto Rico		.54	.84	.76	.62	.58	.60	.68	.86	.92
St. Croix	23		.54	.72	.86	.92	.88	.82	.54	.60
Mona	8	23		.64	.56	.50	.46	.62	.92	.88
Anegada	12	14	18		.80	.76	.74	.90	.74	.76
Tortola	19	7	22	10		.98	.96	.92	.56	.60
St. John	21	4	25	12	1		.96	.84	.52	.64
St. Thomas	20	6	27	13	2	2		.82	.50	.56
Culebra	16	9	19	5	4	8	9		.64	.70
Vieques	7	23	4	13	22	24	25	18		.90
Caja de Muertos	4	20	6	12	20	18	22	15	5	

The values in the lower left half of the matrix are the actual counts of discrepancies between any two localities. The upper right half of the matrix shows the values of the new rank correlation coefficients calculated, using the formula described in the text.

TABLE 2.—*Matrix showing relationships as in Table 1, with islands rearranged into groups as indicated by the discrepancy value*

	St. Croix	St. John	Tortola	St. Thomas	Culebra	Anegada	Vieques	Caja de Muertos	Puerto Rico
St. John	4								
Tortola	7	1							
St. Thomas	6	2	2						
Culebra	9	5	4	9					
Anegada	14	12	10	13	5				
Vieques	20	24	22	25	18	13			
Caja de Muerto	23	18	20	22	15	12	5		
Puerto Rico	23	21	19	20	16	12	7	4	
Mona	23	25	22	27	19	18	4	6	8

The lower triangle includes four islands, the upper includes five (although St. Croix is not a good fit), while Anegada is intermediate in both directions, and is not a good member of either group.

This indicates to me that the ranking technique has another potential value, in that it can give clues as to the origin or derivation of faunas. My interpretation of the figures seen in Table 2 for St. Croix is that the endemism that exists on the island tends to obscure the existence of a considerable similarity between it and its closest neighbors on the Puerto Rican shelf. The discrepancy values permit the recognition of this similarity, and indicate that the fauna of St. Croix is most like the highly similar faunas shared by the three islands mentioned above, which lie in a small group very close to one another, and is either derived from them, or all of them have derived their faunas from a common source.

It is worthwhile to point out that the ranking technique appears to work quite well even though the numbers of species involved are small. Puerto Rico itself has a total of 45 species, but none of the other islands included in this analysis has more than fourteen, and several of them have fewer than ten. One would certainly expect this to have a detrimental effect on the analysis, since addition or deletion of a single species in an island fauna would change that fauna by about 10 percent. It is perhaps an indication of the thoroughness with which these islands have been studied by Heatwole and MacKenzie that such problems do not arise. One suspects that the faunal lists are complete and accurate, so the data is informative even though species counts are small.

It is also important to recognize that in this analysis we are working with a comparatively homogeneous set of samples. The islands analyzed are part of a very compact group, not too distant from each other, and certainly not too different in their respective faunas. Were one to run an analysis of islands from throughout the Caribbean, the group covered in this analysis would appear as a very uniform one, and quite distinct from other such groups in the overall region. It is, therefore, indicative of the sensitivity of the new technique described here that it will pick out and emphasize the comparatively minute differences existing between these islands.

In both the Chain Gang analysis and the Puerto Rican study, described above, the results and interpretations of the original investigators were available for comparison with those derived from the use of the new method described in this paper. While it is rewarding to find that the technique can verify earlier conclusions, it will be limited in value if it cannot be used as an independent, original source of information concerning the biotic relationships as derived directly from field data. In order to check the technique against previously unanalyzed data, a third test was run on the fish collection data from cruise 13

of the research vessel *Anton Bruun,* made off the west coast of South America in 1966. Giles W. Mead, the chief scientist on the cruise, kindly made a preliminary summary of the fish collections available to me for independent analysis. These data (which have recently been published by Craddock and Mead, 1970:Table 6) were in the form of a table, with the fish species listed against the stations where the vessel made Isaacs-Kidd midwater trawl hauls. There was no indication in this table as to any sequence in which the localities were visited, which made it necessary to base the analysis solely on the presence-absence data, and eliminated the possibility of prior bias concerning the results. The collections consisted of 133 species of fishes from 42 stations.

No assumptions were made concerning occurrence. Presence or absence at a station was the only datum used (although the table includes the number of specimens of each species taken at each station). The stations were rearranged into ascending numerical sequence to facilitate their comparison using the computer. Similarity coefficients were calculated for each pair of stations, and these coefficients were then arranged in descending order of rank for each station. A total of slightly more than 800 column comparisons were then made, and the resulting discrepancy values were entered in a matrix in which the stations were again arranged in numerical sequence. As was anticipated, groups of stations showing low discrepancy values between each other and high values when compared with stations outside the group were recognizable. The stations forming the groups were separated within the matrix, however, with very low-numbered stations combining with very high-numbered stations and forming a single group, while slightly more intermediate numbers on both ends of the list formed a second group, and so on. This continued until the stations in the middle range of numbers formed a single group. This is not the pattern one would expect if the numerical series of stations were located along a straight line transect, which would produce groups within the matrix that would be delineated as shown in Figure 9. Each triangle includes all the discrepancy values calculated for pairs of localities within a group. In this illustration, the passage from one biotic area to another is sharp. In a situation where broad transitional zones existed, the borders of the triangles would be blurred at their adjacent corners. When the matrix of discrepancy values resulting from analysis of the fish collections of the *Anton Bruun* cruise is examined, the pattern of relationships between localities is shown as in Figure 10. There is a group of localities forming

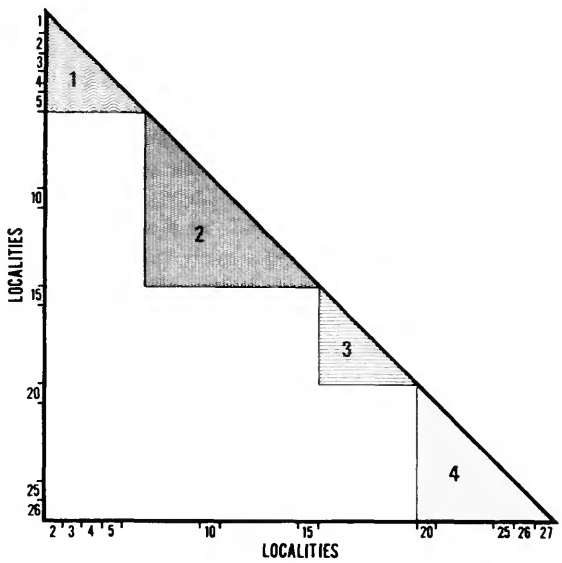


FIGURE 9.—Matrix of localities from a one directional transect. Each triangle includes the localities within a single biotic area.

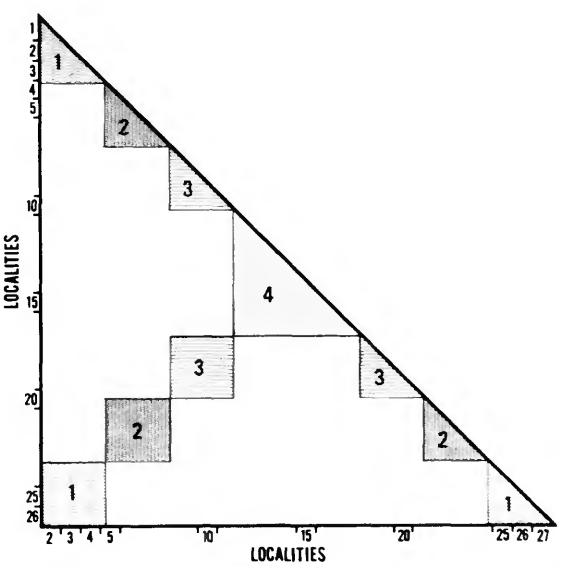


FIGURE 10.—Matrix of stations showing kind of situation that appeared in the analysis of *Anton Bruun Cruise* 13 fish data. See text for amplification.

a triangle of relationships at either end of the matrix, with a square of relationship shown for this group in the lower left-hand corner of the matrix, as well. If one assumes that the numbering of the stations was sequential in time, and also assumes that any pair of stations with a very low discrepancy value was situated in the same biotic area, the ship must have passed through the same biotic area twice, with the low station numbers representing an early visit, and the high station numbers a visit late in the cruise. Since the intermediate station numbers form a single group, it would seem likely that the ship turned to retrace its route while in that biotic area. From the information derived solely from the list of fishes as collected at each station, I was able to determine that the ship had left some specific place, probably a coastal port, traveled out along a path that cut across one extremely distinct biotic boundary, a second boundary of somewhat lesser significance, and a third that is rather poorly defined. These three boundaries are shown in Figure 11 as levels on a phenogram. This phenogram is constructed using the single-linkage method, using discrepancy values taken directly from the matrix. The lowest values in the table are used for the first combination of localities. Thus, localities 4, 47, 48, 49, 50, 51, 52, and 53 are all joined together at the level of zero discrepancies, because each of them has that value in combination with at least one of the others. The lowest value any of that group of eight localities has with any other locality is three, which number 47 has with 46, shown by a line connecting 46 to the group at the "3" level. This technique has been used throughout in constructing the phenogram. The first and second groups are not sharply distinct, but do seem to form groups within themselves, each of which shows stronger relationships within the group than with stations in the other group. The third group is more distinctly set off from the first two, with the lowest discrepancy value for any station within it compared with any station outside it indicating 14 discrepancies, while within-group discrepancy values are fairly uniformly low. The sharpness of the boundary distinguishing the fourth group is shown by the very low values within the group, with all stations joined in the phenogram before the level of four discrepancies is reached, and by the high value that must be reached for any relationship with stations outside the group, which is 41 discrepancies. The cruise track of the ship has been mapped by Mead (1966, fig. 1), and the stations have been plotted by Craddock and Mead (1970, fig. 1B). Using both of these as well as the list of stations in Mead (1966, Table 2), I have plotted the positions of their numbered stations, and shown the four groups derived from my analysis of the data (Figure 12). I conclude that the first two groups lie within the area of influence of the Humboldt Current, and possess a distinctive fish fauna as a consequence. The third group is either distinct in itself or transitional. The fourth group is completely distinct, and probably represents the fauna of a large, well-defined water mass.

Confirmation of some of my conclusions comes from Craddock and Mead (1970:3.39), who say, "The analyses confirmed existence of a faunal break at about 80 W, just west of the region of maximum motion and the high salinity, low-oxygen wedge . . ." This is also shown graphically in their Figure 11, which shows a sharp peak in the chi-square values between localities 16 and 17, as well as almost as sharp at localities 6 and 47. These points are identical with the ones identified in my analysis.

The Use of Numerical Taxonomic Methods in Biogeography

As was pointed out before, some authors (as, for example, Hagmeier and Stults, 1964) have used certain techniques derived from numerical taxonomy as the basis for analysis of biogeographic regions. One might wonder why I have not used these numerical taxonomic methods in my work. It would be simple to say that I have tried them and found them inadequate and unsatisfactory. I think, however, in light of the fact that many workers still feel uneasy about these techniques, without knowing exactly why, I should attempt to show why they do not work sufficiently well to be useful in biogeography.

To do this adequately, it will be necessary to review in some detail what is done in biogeographical data processing if numerical taxonomic methods are used. The original data consists of records of the presence or absence of taxa within some circumscribed geographical unit. In my analysis of the Chain Gang data, the unit was a single locality, or rather a single collection site. Hagmeier and Stults ran the earlier part of their analysis on 50 square-mile blocks, but when they used numerical taxonomic techniques,

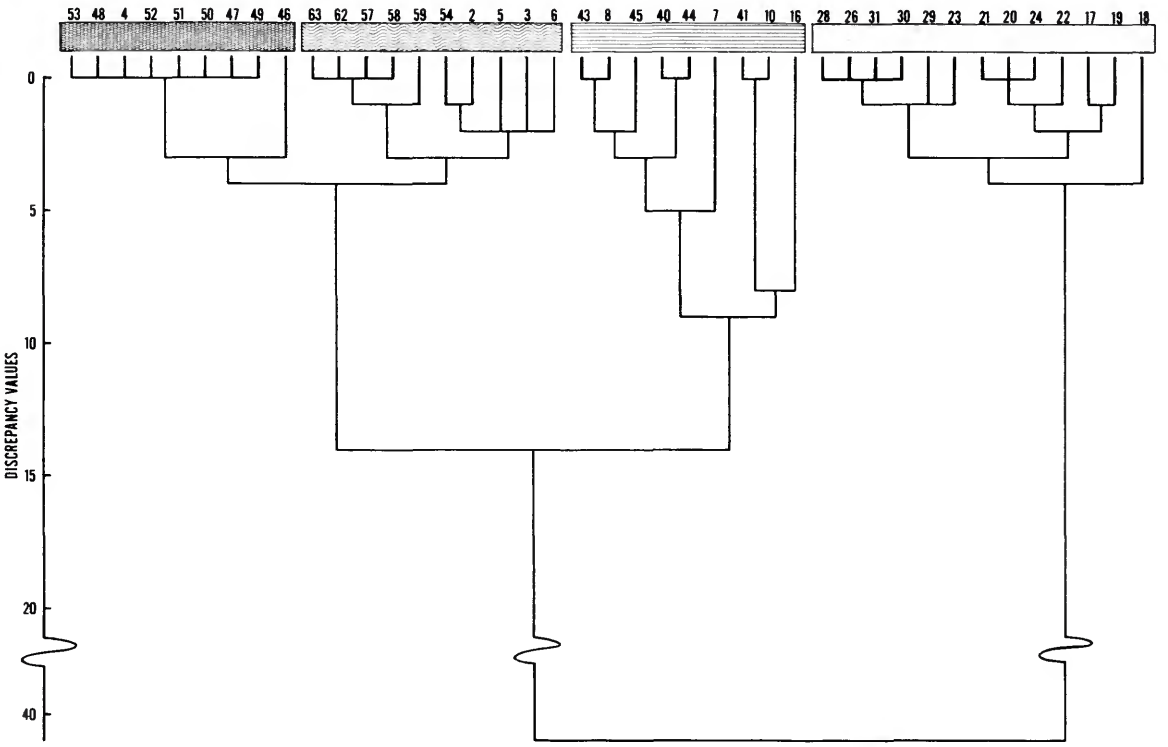


FIGURE 11.—Phenogram of relationships of stations from Cruise 13, *Anton Bruun*, based on single-linkage method. See text.

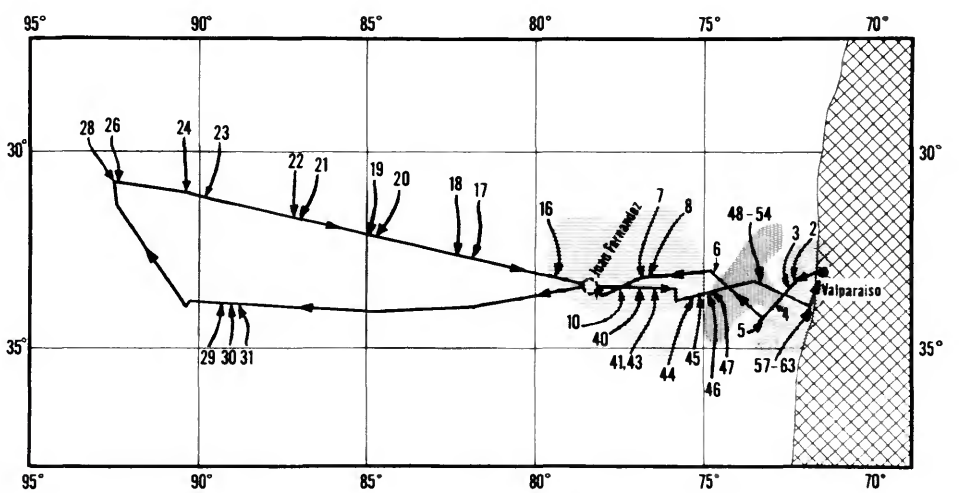


FIGURE 12.—Cruise track of *Anton Bruun*, Cruise 13, with stations indicated. The stations shown here are in the same groups formed in the matrix, Figure 10, and the phenogram, Figure 11.

they changed the basic unit to biotic provinces that had been defined earlier. The sequence of events involved in cluster analysis, which is the technique used by Hagmeier and Stults, is shown in Figure 13. One begins by constructing a matrix that will eventually show various interrelationships between the geographic units. In Figure 13, the geographic units are "R," "S," and "T." In the first three-by-three matrix (or table) we see, on the diagonal, the number of species that occur at each locality. This value is found by going through the original presence-absence data for each locality and counting every taxon marked "present." There are 40 species in locality "R," 55 in "S," and 50 in "T." Once these data are recorded, we reexamine our original data to see how many species are shared between each pair of localities. The second matrix has this information recorded in its upper right half, in the larger figures. These values are found by comparing the lists of taxa for any two localities, and any time a taxon is found as "present" in both, it is added to the accumulating total. The matrix shows that 30 taxa out of the 50 found in "S" are also found in "R," and also that 40 of the 50 known from "T" are included in the 55 found in "S." The third matrix shows the results of the first calculation involving these data. Using the formula shown under the second matrix (the Simpson Coefficient), a similarity coefficient is calculated. The "c" of the formula refers to the number of taxa in common, and the "n_1" refers to the number of species found at the locality with the smaller number of taxa. There are many other similarity coefficients that could be used (Peters, 1968; Cheetham and Hazel, 1969), but this is the simplest and, therefore, the easiest to follow through this summary. The similarity coefficients are the enlarged figures in the third matrix, which are calculated using the numbers in the rest of the matrix. Thus, the value of 75 is arrived at by using the number of species shared between "R" and "S" (30), dividing that by the number of species in the locality having the smaller number ("R," with 40), and multiplying the resultant .75 by 100. Of the three similarity coefficients calculated, the largest is 80, for localities "S" and "T," and is indicated by an arrow.

Up to this point, everything that has been done is common to the JPFRF program and to cluster analysis. One refinement of cluster analysis, commonly used in taxonomic studies, is to count the number of situations where an item is missing in both units of a pair, as well as when it is present, which will increase the value of a similarity coefficient considerably, but need not concern us at the present. The next step in cluster analysis is to "shrink the matrix," a technique commonly used in numerical taxonomic studies for phenogram construction. The value of "80" is recorded as the point at which "S" and "T" were joined, and the two columns devoted to those two localities in the three-by-three matrix are combined, using the procedure shown in the left-hand two-by-two matrix. The number of species shared by "S" with "R" (30) and the number shared by "T" with "R" (26) are added, and then divided by two for the average value, which is 28; this is the new value used in further work with the matrix, including additional shrinking. The "28" is an expression of the number of species shared between "R" and the combined localities "S" and "T." To find the number of taxa used in continuing cluster analysis for the combined localities, the number of taxa found at locality "S" (55) is added to the number found at locality "T" (50), and the average taken by dividing by two. This gives a new value of 53, and this new value will be used in further shrinking of the matrix. Were the matrix larger, this procedure would be followed for the entire set of mutual values between "S" and "T." Finally, then, a new similarity coefficient is calculated for the relationship between "R" and the combined "S–T" locality. Since the averaged value for shared taxa is 28, and the locality with the smaller number of species is "R" (40), the new similarity coefficient is 70. This is put in the lower left, where it would, in this example, represent the value used for the final horizontal connection in a phenogram.

When the matrix is larger than in my figure, which shows only the first shrinkage, from three to two in this case, there are two ways to combine values existing in the matrix. The first uses the averages calculated in earlier steps of shrinkage to find a new average, and is called the "weighted pair group method." If the values found in the original, full matrix are used, the method is called the "unweighted pair group method." These two methods produce what I consider to be the first failure of numerical taxonomic methods, because they do not retain the biological and geographical realities that exist in the original data. First of all, let us examine the figure that rep-

resents the number of taxa found in the combined localities "S" and "T." By averaging the known number of species in the two localities, we arrived at a new value of 53. But it is immediately clear that these two localities, after they have been combined, could not have fewer species than the larger number found at either independently, which is the 55 found at "S." No matter how many times "S" is combined with other spots on a map, it is biologically accurate to say that any geographic unit including "S" must be recognized as having at least 55 taxa within its limits. There are two ways to find out exactly how many taxa occur within the new geographic unit that includes both "S" and "T." Either the original data can be reexamined and the taxa recounted, or we can simply add together the two known values and subtract the known number of shared species. The latter procedure is carried out in the lower right of the matrix on the right side, with the resultant value of 65. This tells us, of course, that exactly 65 species are known from the geographic unit we are now considering, a value quite different from the 53 coming from the calculated average, and one that will have a sizable effect on further calculations of similarity coefficients.

Similarly, it is clear that the number of species shared by the new, combined geographic unit with "R" cannot be less than the larger number shared with "R" by either of the original localities. This is 30 taxa, and the upper right section of the matrix shows that the new value must be larger than that. In general, the maximum value possible for the number of species in common would be the sum of the species shared by "S" with "R" and by "T" with "R," or 56. This is a very improbable figure, since the two localities were combined because of a high

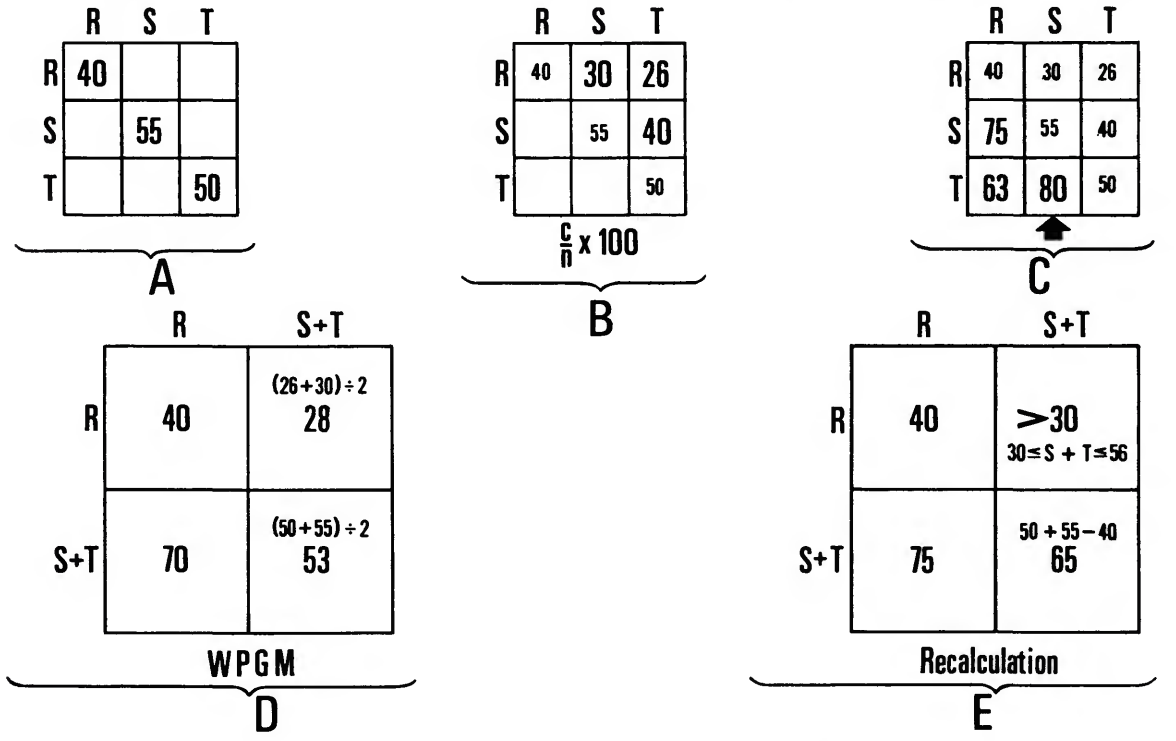


FIGURE 13.—The first three matrices, A, B, and C, indicate the method of calculating similarity coefficients. Matrix D shows a single step in shrinking the matrix in C, using the weighted pair group method (WPGM). Matrix E shows the same step if one returns to the original data, and establishes accurate figures for the combined localities.

similarity coefficient, and thus share many species. In this case, there is only one way to find out the actual number of species shared by the new geographic unit with "R." One must go back to the original data, consolidate presence-absence data for the two combined localities, and then retabulate the number of species shared with all other localities. In effect, one is saying that since these two localities are so similar as to have the highest similarity coefficient value within our matrix, we will henceforth treat them as a single, recognizable unit. This seems biologically defensible to me, and is certainly the assumption made when one works with the fauna of a "biotic province," as was done by Hagmeier and Stults.

While doing this by hand would be extremely tedious—and clearly is the primary reason no one has done it in previous analyses—the computer facilitates such data manipulations, doing them so rapidly that one is unaware of the magnitude of the reshuffling involved, if the number of localities is high. The two-by-two matrix on the right in Figure 13 shows the kinds of results to be expected if one recalculates both the number of species at a locality and the number of species that new locality shares with others by going back to the original data strings. The differences that result from calculation using the WPGM method and the method of recalculation from the original data are so great that I must consider the WPGM biologically indefensible, even though it may be demonstrably legitimate mathematically. There is some reason to doubt the legitimacy of the latter, however, in that it involves the use of averages to calculate further averages, a technique that results in increasing error as it is continued.

The data from the Chain Gang collections were run through cluster analysis, using both the WPGM and the recalculation method proposed here, to see whether there would be significant differences. Figure 14 is the phenogram constructed through use of the WPGM. The horizontal line connecting two or more vertical lines is based on the selection of the highest value in a matrix of similarity coefficients, followed by shrinkage of the matrix through combination of the two localities involved. If this phenogram means anything at all, it is that there is a series of fairly closely related localities, ranging from 802 to 813, with 800 and 801 quite distinct from the others and

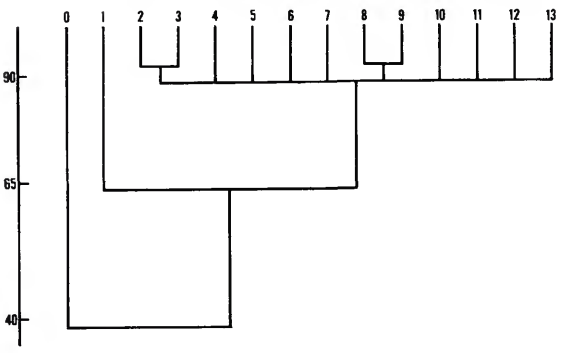


FIGURE 14.—Phenogram resulting from cluster analysis of Chain Gang data, using the WPGM.

from each other. This bears no resemblance at all to the faunal boundary picture derived both from the original analysis by the Chain Gang or from the JPFRF analysis.

If, however, the same cluster analysis program is modified to recalculate the data concerning presence-absence from the original information and to insert the recalculated values in the matrix where appropriate, several changes emerge when the same data are run through analysis. First of all, this cluster analysis now produces subsequent similarity coefficients higher than earlier values. This is impossible when using the WPGM, because that method is dependent upon averaging, and any new value based on averages of previous values must necessarily be smaller throughout the sequence. This, in fact, is vitally necessary to permit the construction of a phenogram, as is done in numerical taxonomy, since the reducing values are used to indicate a lower level of connection between any two units. Using the recalculation method, however, it is to be anticipated that higher similarity coefficients will result as more and more localities are combined, because the number of shared species can only get bigger. This means that the value of "c" in the equation shown in Figure 13 gets larger, and the fraction gets closer and closer to unity. As a consequence, the standard phenogram cannot be constructed. This is shown in Figure 15. The values given are taken directly from the computer readout, showing the numbers of the localities which are combined in each step (812 and 813, in the first case), and the new value assigned to the combination (15). This is shown in A as vertical lines from 12 and 13, joining at a level of 80, which is the value

of the similarity coefficient, to form a new locality 15. The same is true throughout the figure, except when a new similarity coefficient is larger than a previous one involving localities already in the phenogram. The first of these instances is the merger of 16 and 20 at a coefficient of 64. Since the merger of "0" and "1" had occurred at 60, it is no longer possible to show the further merger in the standard fashion. I have modified the phenogram, therefore, by combining 0 to 1 into a single line, doing the same with 4 and 6, and putting the new value of 21 at the horizontal line. As is seen in B, however, the next similarity coefficient is still higher, and the following higher again. In each case, I have combined all localities previously "clustered" into a single line, until the situation shown in C is reached. At this point, where 19 and 24 combine to form 25, there are two arms on the phenogram, one of which includes all localities from 0 to 7, except 3, and the other of which includes all localities from 8 to 13, except 10. In D, we see the final merging of these last two with the others. While this merging of the arms of the phenograms certainly obscures the picture and is quite unsatisfactory as a method of demonstrating the events taking place, it is worthwhile to point out that two distinct groups of localities have formed, and that the groups so formed are the same as those distinguished by the Chain Gang analysis and by the JPFRF program. I regard this as a third verification of the validity of the recognition of a faunal barrier at or near locality 807, and the total absence of any such indication in straight cluster analysis supplements my statements concerning the inadequancy of this method of analysis.

In addition to the remarks above, there is another fact that militates against the use of cluster analysis for biogeographic data. This lies in the fact that the similarity coefficients are calculated for the relationships of all localities with every other, but the only coefficients used in any one shrinkage of the matrix are the individual values produced by a single locality compared with another. This requires the assumption that a high similarity coefficient necessarily indicates a strong relationship between the two localities, i.e., "A" and "B," which may be true in almost every case. But it also requires the opposite assumption

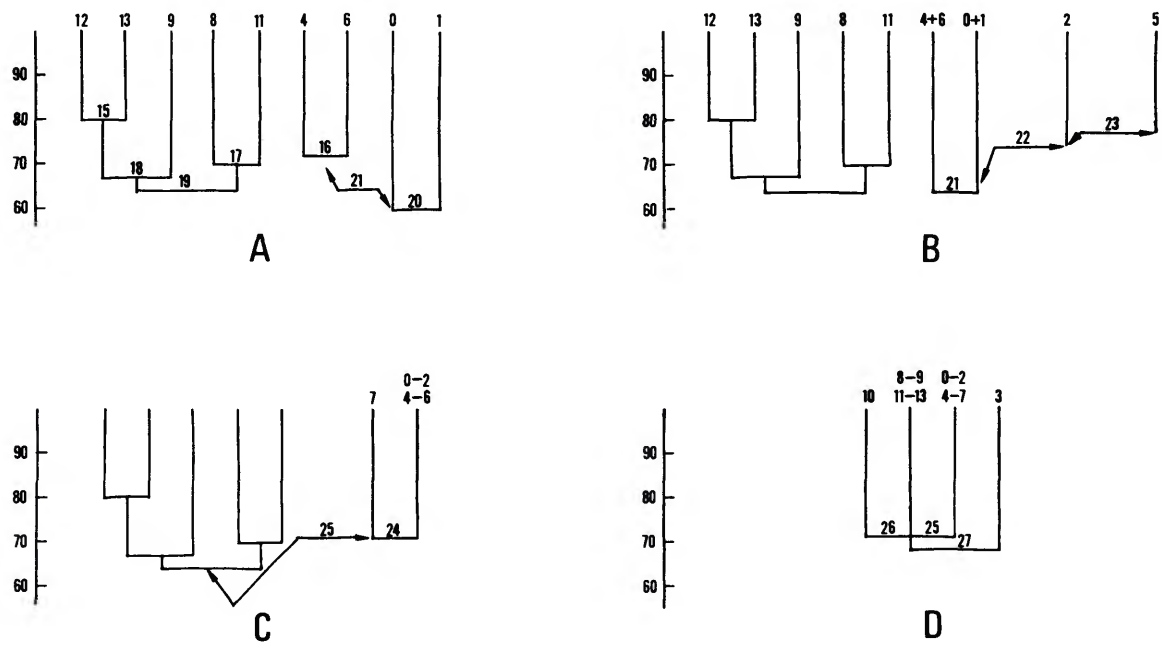


FIGURE 15.—The "phenogram" resulting from attempting cluster analysis on Chain Gang data, if data are recalculated, using original data after each step in matrix shrinkage. See text for details.

that a lower similarity coefficient between two other localities, i.e., "X" and "Y," means that this latter pair is less closely related, and that may not be true at all. When all localities are ranked by similarity coefficient values under A, B, X, and Y, and discrepancy counts made, it is entirely possible that X and Y will show no discrepancies at all, or that A and B comparison will produce a higher discrepancy count than X and Y comparison. This is the reason that the technique used in the JPFRF program takes advantage of the similarity coefficients known for all possible combinations of localities, and includes them all in the ranking procedure. Highest degree of relationship is correlated with number of discrepancies, not with the calculated value of the similarity coefficient. The failure to use all the available data in a cluster analysis adds to its unsatisfactory nature, since the computer is available to make it possible to use ranks.

The JPFRF Program

The computer program (JPFRF) used in this analysis is written in the language BASIC. The initial section of the program uses the short program published by me in 1968, and it is still possible to use that paper for the insertion of a choice of similarity coefficients. The line numbers and statements remain the same.

The data to be analyzed are stored in a separate file, to be read when called for by the program. The name of the file, which can be any combination of one to six letters or numbers (it is usually not advisable to start a file name with a number), is inserted in the program by the user, by typing the following:

22 FILES FRFDTA

The "FRFDTA" is the name I assigned to my file of data, but any name not used elsewhere in stored files of the user's computer system can be assigned. The data file consists of the following information: first, the number of localities; second, the total number of taxa recorded for all localities; and third, the identifying numbers assigned to the localities. These "identifiers" need not be in any numerical sequence, but must be from 1 to 99, inclusive. Then, for each locality, in the same sequence in which the identifying numbers were given, a string made up of ones and zeros, totaling the number of species recorded for all localities, is entered. The total species list must remain the same, in the same sequence, for all localities. If a species is present in a locality, a one is entered in the data string for that locality at the proper point for that species in the sequence. If a taxon is absent, a zero is entered. The data file is complete when the string of ones and zeros has been recorded for the last locality on the numbered list.

When the data file is properly constructed and the file is named in JPFRF1, the program can be run. At the present time the program JPFRF is in two parts, due to restrictions on its length imposed by the computer system we are using. The first half,

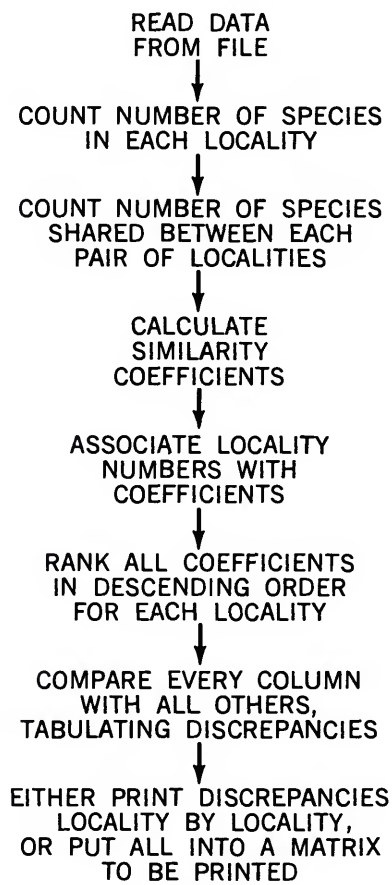


FIGURE 16.—The flow sequence taking place in a run of the JPFRF program.

JPFRF1, completes part of the data analysis and records the results in an output file. It gives the opportunity to have a listing of the matrix showing the number of species at each locality, the number of species shared between all pairs of localities, and the similarity coefficients that have been calculated, if the user wishes. If one is using Peters, 1968, to change the similarity coefficient to be used in the analysis, the revised lines should be added to JPFRF1 before it is run, exactly as instructed in the paper.

The second half of the program, JPFRF2, is then run. This section cannot be run prior to JPFRF1, because it reads its data from the output file (called FRFMAT) constructed by JPFRF1, and FRFMAT does not exist before that run. This half of the program will produce final output either as a listing of each possible pair of localities with the discrepancy value tabulated for them, or as a matrix with the localities identified across the top and down the side, depending upon what the user wishes to have.

The sequence of events taking place during a run of the program is shown in Figure 16. Each step in the sequence is ended by the completion of a new matrix based on the calculations made during that step, and it is possible to obtain a listing of that matrix if the user wishes. This is not built into the program at the present time, however, and the program will run to completion without intermediate output, unless the user modifies it.

The program itself follows:

```
19 DIMX(100),C(100)
20 DIM S(35,35), A(200,2), P(35)
21 LETC=O
22 FILES FRFDTA
23 READ#1,Z
24 READ#1,K
30 IFC>OTHEN39
31 FORI=1TOZ
32 LETJ=O
33 READ#1,S(I,J)
34 LETS(J,I)=S(I,J)
35 NEXTI
36 LETC=C+Z+2
37 FOR B1=1TOZ
38 FILES FRFDTA
39 FORH=1TOC
40 READ#1,X
45 NEXTH
47 LETC=C+K
55 LET R=O
60 FORI=OTOK-1
65 LETJ=O
70 READ#1,A(I,J)
75 IFA(I,J)=OTHEN85
80 LET R=R+1
83 LET P(B1)=R
84 LET S(B1,B1)=P(B1)
85 NEXT I
90 IF B1=ZTHEN250
95 LETB2=B1+1
96 LET Q=O
100 FORI=OTOK-1
105 LETJ=1
110 READ#1,A(I,J)
115 NEXTI
120 FORI=OTOK-1
121 LETJ=O
125 LETN=A(I,J)
130 IFN=OTHEN150
135 LETM=A(I,J+1)
140 IFN<>MTHEN150
145 LETQ=Q+1
150 NEXTI
153 LETS(B1,B2)=Q
155 LETB2=B2+1
160 IFB2=Z+1THEN175
165 GOTO96
175 NEXTB1
250 LETN=O
254 LETW=Z
257 FORI=1TOZ+1
260 LETN=N+1
263 IFN=ZTHEN331
264 LETB=P(N)
267 LETT=N
270 FORJ=I+1TOZ
273 IFI=JTHEN320
275 LETT=T+1
280 LETC=P(T)
300 LETV=S(I,J)/(B+C-S(I,J))*100
310 LETV=INT(V+.5)
315 LETS(J,I)=V
320 NEXTJ
325 LETW=W-1
330 NEXTI
331 FILES FRFMAT
333 WRITE#1,Z
335 FORI=OTOZ
```

```
340 FORJ=OTOZ
345 WRITE#1,S(I,J)
350 NEXTJ
355 NEXTI
505 PRINT"WANT PRINTOUT OF MATRIX?
    (0=NO, 1=YES)"
506 INPUT E
507 IFE=OTHEN10000
510 FOR H=O TO Z-1 STEP 10
520 FOR I=O TO Z-1
530 FOR J=H TO H+9
540IFI=JTHEN570
550 PRINTS(I,J);
560 GOTO580
570 PRINTP(I);
580 IFJ=Z-1THEN600
590 NEXTJ
600 PRINT
610 NEXTI
620 PRINT
630 NEXTH
10000 END
```

JPFRF2

```
15 DIMS(35,35)
18 FILES FRFMAT
19 READ#1,Z
20 FORI=OTOZ
21 FORJ=OTOZ
22 READ#1,S(I,J)
24 NEXTJ
26 NEXTI
32 FORI=1TOZ
34 FORJ=1TOZ
36 LETS(I,J)=S(J,I)
38 NEXTJ
40 NEXTI
42 LETI=O
44 FORJ=1TOZ
46 LETA(I,J-1)=100+(S(I,J)/100)
48 NEXTJ
50 FORI=1TOZ
52 LETH=O
54 FORJ=1TOZ
56 IFJ=ITHEN66
57 LETH=H+1
58 IFJ>ITHEN64
60 LETA(H,I-1)=S(I,J)+S(J,O)/100
62 GOTO66
64 LETA(H,I-1)=S(I,J)+S(O,J)/100
66 NEXTJ
68 NEXTI
70 FORI=OTOZ
72 FORJ=OTOZ
74 LETS(I,J)=A(I,J)
76 NEXTJ
78 NEXTI
90 FORJ=OTOZ-1
100 FORI=OTOZ-1
105 LETX1=I
110 LETP=S(X1,J)
120 LETM=X1
130 LETX1=X1+1
140 IFM+1=ZTHEN260
150 LETN=S(X1,J)
160 IFP>NTHEN240
170 IFP=NTHEN240
180 LETX1=X1-1
190 LETS(X1,J)=N
200 LETS(X1+1,J)=P
210 IFM=OTHEN250
220 LETX1=M-1
230 GOTO100
240 LETX1=M
250 NEXTI
260 NEXTJ
270 FORJ=OTOZ-2
280 LETX2=J
285 LETD=J
290 LETC=S(O,X2)
300 LETB=INT(C)
310 LETC=INT(((C-B)*100)+.5)
320 LETA=X2
330 LETX2=D+1
340 LETG=S(O,X2)
350 LETH=INT(G)
360 LETF=O
370 LETG=INT(((G-H)*100)+.5)
380 PRINT
400 FORI=1TOZ-1
405 LETX1=I
410 LETX2=A
420 LETN=S(X1,X2)
430 LETM=INT(N)
440 LETN=(N-M)*100
450 IFABS(N-G)<.001THEN610
460 LETT=X1
470 LETK=-X1
```

```
480 LETX1=T+K
490 IFX1>Z-1THEN580
500 LETX2=D
510 LETQ=S(X1,X2+1)
520 LETP=INT(Q)
530 LETQ=(Q-P)*100
540 IFABS(N-Q)<.001THEN570
550 LETK=K+1
560 GOTO480
570 GOSUB690
580 LETK=ABS(K)
590 LETF=F+K
600 LETX1=T
610 NEXTI
620 PRINT"LOCS";C;"AND";G;"---";F
630 LETD=D+1
640 IFD=Z-1THEN660
650 GOTO330
660 LETX2=A
670 NEXTJ
680 GOTO10000
690 IFK=OTHEN990
700 IFK>OTHEN770
710 LETR=INT(S(X1+1,X2+1))
720 IFABS(P-R)>.001THEN830
730 LETX1=X1+1
740 LETK=K+1
750 IFK=OTHEN990
760 GOTO710
770 LETR=INT(S(X1-1,X2+1))
780 IFABS(P-R)>.001THEN910
790 LETX1=X1-1
800 LETK=K-1
810 IFK=OTHEN990
820 GOTO770
830 LETX1=T
840 LETX2=A
850 LETU=INT(S(X1-1,X2))
860 IFABS(M-U)>.001THEN990
870 LETX1=X1-1
880 LETK=K+1
890 IFK=OTHEN990
900 GOTO850
910 LETX1=T
920 LETX2=A
930 LETU=INT(S(X1+1,X2))
940 IFABS(M-U)>.001THEN990
950 LETX1=X1+1
960 LETK=K-1
970 IFK=OTHEN990
980 GOTO930
990 RETURN
9000 FORI=OTOZ
9005 FORJ=OTOZ
9010 PRINTS(I,J);
9015 NEXTJ
9016 PRINT
9020 NEXTI
9025 RETURN
10000 END
```

Literature Cited

Backus, R. H., G. W. Mead, R. L. Haedrich, and A. W. Ebeling
1965. The Mesopelagic Fishes Collected during Cruise 17 of R/V *Chain*, with a Method for Analyzing Faunal Transects. *Bulletin of the Museum of Comparative Zoology,* 134:139–158.

Cheetham, Alan H., and J. E. Hazel
1969. Binary (Presence-Absence) Similarity Coefficients. *Journal of Paleontology,* 43:1130–1136.

Craddock, James E., and Giles W. Mead
1970. Midwater Fishes from the Eastern South Pacific Ocean. *Anton Bruun Report Number 3*:3.3–3.46.

Hagmeier, Edwin M.
1966. A Numerical Analysis of the Distributional Patterns of North American Mammals. II. Re-evaluation of the Provinces. *Systematic Zoology,* 15: 279–299.

Hagmeier, Edwin M., and C. Dexter Stults
1964. A Numerical Analysis of the Distributional Patterns of North American Mammals. *Systematic Zoology,* 13:125–155.

Heatwole, Harold, and Faustino MacKenzie
1967. Herpetogeography of Puerto Rico. IV. Paleogeography, Faunal Similarity and Endemism. *Evolution,* 21:429–438.

Holloway, J. D., and N. Jardine
1968. Two Approaches to Zoogeography: a Study Based on the Distributions of Butterflies, Birds and Bats in the Indo-Australian Area. *Proceedings of the Linnaean Society of London,* 179:153–188.

Huheey, James E.
1966. A Mathematical Method of Analyzing Biogeographical Data. I. Herpetofauna of Illinois. *American Midland Naturalist,* 73:490–500.

Kendeigh, S. C.
1961. *Animal Ecology*. Englewood Cliffs, N.J.: Prentice-Hall. 468 pages.

Kikkawa, J., and Kay Pearse
1969. Geographical Distribution of Land Birds in Australia—A Numerical Analysis. *Australian Journal of Zoology,* 17:821–840.

Mead, Giles W.
1966. Cruise Report, Research Vessel *Anton Bruun*, Cruise 13. *Marine Laboratory, Texas A&M University, Galveston, Special Report Number 3*:1–3, 4 tables, 4 figures, 2 appendices.

Peters, James A.
1955. Use and Misuse of the Biotic Province Concept. *The American Naturalist*, 89:21–28.
1968. A Computer Program for Calculating Degree of Biogeographic Resemblance between Areas. *Systematic Zoology*, 17:64–69.

Simpson, George G.
1960. Notes on the Measurement of Faunal Resemblance. *American Journal of Science, Bradley Volume*, 258–A:300–311.

Simpson, George G., Anne Roe, and Richard C. Lewontin
1960. *Quantitative Zoology—Revised Edition*. vi+440 pages. New York: Harcourt, Brace, & World.

Smith, Hobart M.
1949. Herpetogeny in Mexico and Guatemala. *Annals of the Association of American Geographers*, 39: 219–238.

Sokal, Robert R., and F. James Rohlf
1969. *Biometry*. xxi+776 pages. San Francisco: Freeman.

Webb, William L.
1950. Biogeographic Regions of Texas and Oklahoma. *Ecology*, 31:426–433.

* U.S. GOVERNMENT PRINTING OFFICE: 1971 O—437–339/10

Publication in Smithsonian Contributions to Zoology

Manuscripts for serial publications are accepted by the Smithsonian Institution Press, subject to substantive review, only through departments of the various Smithsonian museums. Non-Smithsonian authors should address inquiries to the appropriate department. If submission is invited, the following format requirements of the Press will govern the preparation of copy.

Copy must be typewritten, double-spaced, on one side of standard white bond paper, with 1½″ top and left margins, submitted in ribbon copy with a carbon or duplicate, and accompanied by the original artwork. Duplicate copies of all material, including illustrations, should be retained by the author. There may be several paragraphs to a page, but each page should begin with a new paragraph. Number consecutively all pages, including title page, abstract, text, literature cited, legends, and tables. The minimum length is 30 pages, including typescript and illustrations.

The *title* should be complete and clear for easy indexing by abstracting services. Taxonomic titles will carry a final line indicating the higher categories to which the taxon is referable: "(Hymenoptera: Sphecidae)." Include an *abstract* as an introductory part of the text. Identify the *author* on the first page of text with an unnumbered footnote that includes his professional mailing address. A *table of contents* is optional. An *index,* if required, may be supplied by the author when he returns page proof.

Two *headings* are used: (1) text heads (boldface in print) for major sections and chapters and (2) paragraph sideheads (caps and small caps in print) for subdivisions. Further headings may be worked out with the editor.

In *taxonomic keys,* number only the first item of each couplet; if there is only one couplet, omit the number. For easy reference, number also the taxa and their corresponding headings throughout the text; do not incorporate page references in the key.

In *synonymy,* use the short form (taxon, author, date:page) with a full reference at the end of the paper under "Literature Cited." Begin each taxon at the left margin with subsequent lines indented about three spaces. Within an entry, use a period-dash (.—) to separate each reference. Enclose with square brackets any annotation in, or at the end of, the entry. For *references within the text,* use the author-date system: "(Jones 1910)" and "Jones (1910)." If the reference is expanded, abbreviate the data: "Jones (1910:122, pl. 20: fig. 1)."

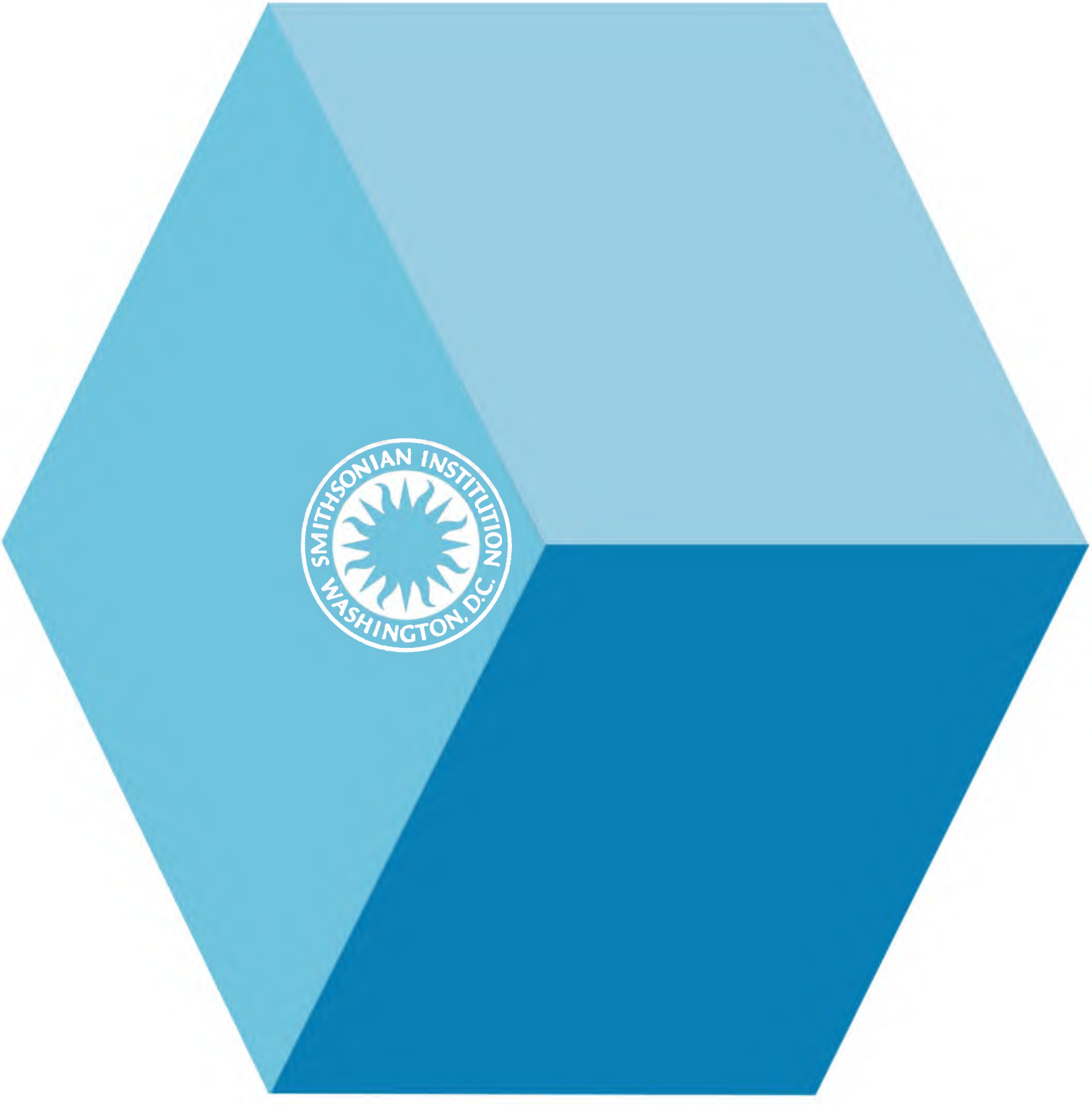
Simple *tabulations* in the text (e.g., columns of data) may carry headings or not, but they should not contain rules. Formal *tables* must be submitted as pages separate from the text, and each table, no matter how large, should be pasted up as a single sheet of copy.

Use the *metric system* instead of, or in addition to, the English system.

Illustrations (line drawings, maps, photographs, shaded drawings) can be intermixed throughout the printed text. They will be termed *Figures* and should be numbered consecutively; however, if a group of figures is treated as a single figure, the components should be indicated by lowercase italic letters on the illustration, in the legend, and in text references: "Figure 9*b*." If illustrations (usually tone photographs) are printed separately from the text as full pages on a different stock of paper, they will be termed *Plates,* and individual components should be lettered (Plate 9*b*) but may be numbered (Plate 9: figure 2). Never combine the numbering system of text illustrations with that of plate illustrations. Submit all legends on pages separate from the text and not attached to the artwork. An instruction booklet for the preparation of illustrations is available from the Press on request.

In the *bibliography* (usually called "Literature Cited"), spell out book, journal, and article titles, using initial caps with all words except minor terms such as "and, of, the." For capitalization of titles in foreign languages, follow the national practice of each language. Underscore (for italics) book and journal titles. Use the colon-parentheses system for volume, number, and page citations: "10(2):5-9." Spell out such words as "figures," "plates," "pages."

For *free copies* of his own paper, a Smithsonian author should indicate his requirements on "Form 36" (submitted to the Press with the manuscript). A non-Smithsonian author will receive 50 free copies; order forms for quantities above this amount with instructions for payment will be supplied when page proof is forwarded.


SMITHSONIAN INSTITUTION
WASHINGTON, D.C.