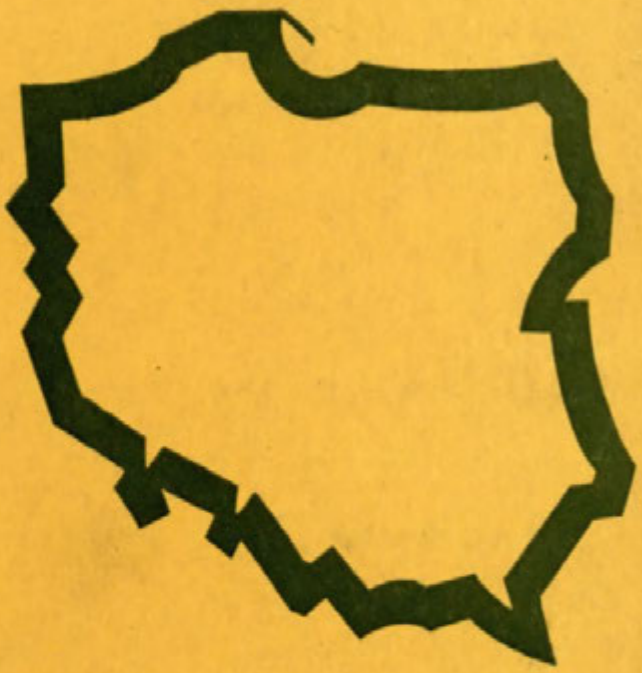
POLISH ACADEMY OF SCIENCES

GEOGRAPHIA POLONICA



25

PWN—POLISH SCIENTIFIC PUBLISHERS

Address of Editorial Board

KRAKOWSKIE PRZEDMIEŚCIE 30,
WARSZAWA 64
POLAND

Printed in Poland

POLISH ACADEMY OF SCIENCES
INSTITUTE OF GEOGRAPHY

GEOGRAPHIA POLONICA

25

PWN — Polish Scientific Publishers • Warszawa 1973

PERSPECTIVES ON SPATIAL ANALYSIS

Edited by

DUANE F. MARBLE and ZBYSZKO CHOJNICKI

CONTENTS

PREFACE

The papers in this volume represent part of the continuing program of the Commission on Quantitative Methods of the International Geographical Union. This program has a threefold intent: to improve international communications among scholars undertaking quantitative methods in geographic research; to promote a continuous series of "state-of-the-art" reviews of statistical and mathematical methods in forms useful to those seeking to improve their quantitative skills; and to stimulate new and highly original work on the frontiers of quantitative methodology.

This collection resulted from a conference in Poznań, Poland, held primarily to meet the first two goals. It assembles the mathematical papers presented. A companion publication will be published elsewhere containing the statistical papers delivered at Poznań. Draft of the papers were circulated, discussed, and then revised for this publication.

The Commission is indebted to each author and participant in the conference, to the Institute of Geography, Polish Academy of Sciences and to Professor Stanisław Leszczycki, to the Adam Mickiewicz University, and to the editors, Professors Zbyszko Chojnicki and Duane F. Marble, for their efforts on behalf of greater international understanding.

Brian J. L. Berry
Chairman of the Commission

GEOGRAPHIA POLONICA 25, 1973

CLASSIFICATION AND GEOGRAPHY

PHILIP M. LANKFORD AND R. KEITH SEMPLE

GENERAL CONSIDERATIONS

Much of the scientific methodology found in geography is common to other social, physical, and biological sciences. All sciences, for example, place emphasis on observation, classification, experimentation, as well as, theory and model building. Classification of observations is an important, but normally early, step in the development of a science. The physical sciences have passed this step while the social sciences have just begun to produce generally-acknowledged classifications. The biological sciences such as botany and zoology are reexamining the issue of classification presently and numerical taxonomy is of central concern in the debate.[1] When classification and regionalization are shown to be analogous procedures the links between formal logic and classification can be used to examine regionalization. The relations between regionalization and classification have been formally stated by Bunge (1962), DeJong (1962), and most extensively by Grigg (1965, 1967). As early as 1956 Reynolds (1956) noticed that "...the delineation of regions is essentially a classification process," but the obvious analogy with classification was missed because regionalization was carried out by drawing isarithms and not by classifying similar objects.

Classification may be defined as the grouping of objects into classes, based on some similarity of properties of, or relationships between, the objects. The object is an individual, with properties. All individuals together form the universe. To classify, one property possessed by all the individuals is selected as the differentiating characteristic. The universe of k individuals may be partitioned into one to k classes. The classes may be grouped into larger classes to form a hierarchy.

The inverse of classification is termed logical division. Beginning with the universe the group is divided according to a principle. A special case of logical division is dichotomous division; an example is the division of the United States into The South and Not-the South.

The analogy is made that regions are areal classes. The basic procedures of formal logic can be applied to the methods used by geographers for regionalization. The two methods of classification, classification and division, can be equated with inductive and deductive methods. Much has been written on the latter method in geography and in many other disciplines. Whittlesey (1954, p. 38) has declared that regional systems may be arrived at by either method; the same point was recognized by Gilbert (1960, p. 160). Implied by this distinction is that theoretically either method may be used with an area of any size even

[1] See R. R. Sokal (1962), R. R. Sokal and J. Camin (1965), R. R. Sokal and P. H. A. Sneath (1963).

though classification is usually used for small areas and divisions employed for world or national classification. The world classes are deductive and therefore assumed to exist *a priori*, which is NOT the case with classification. The basic approaches to classification are, therefore, parallel to the methods of regionalization.

The main problem in geography is the choice of the object or individual for classification. Bunge (1962, p. 16) suggests "place," but this does little to solve the problem, which is a direct consequence of the continuity of the earth's surface versus the ease of individual definition in zoology or botany. As an example of the problem, a regionalization based on natural vegetation, type of farming, and settlement pattern, may choose farms, plant communities, or settlements as the defined individual. Little attention has been given to this problem in English, but discussions in Russian and German have been profuse. The solution is simple: use what units of observation are readily available and label such individuals "Operational Taxonomic Units" (OTUs), as taxonomists have done.

Location is another problem for the analogy with classification. Contiguity has no parallel with the principles of classification. Bunge (n.d., p. 1) has criticized Berry for not including location as a category in classification. "Since the locational category is the category that distinguishes geographic classification (uniform regions) from all others, this omission is serious." This location problem has been essentially solved methodologically by adding location as a restraining characteristic during the classification process.[2]

Employing the analogy between classification and regionalization, there are several principles of classification highly relevant: [3]

(1) Classifications should be designed for a specific purpose; they rarely serve two purposes equally well. Purpose and use must be linked.

(2) The classification of any group of objects should be based upon properties which are properties of those objects; it follows that differentiating characteristics should be properties of the objects classed.

(3) The differentiating characteristics must be important for the purpose of classification or else the classification is trivial.

(4) Classifications are not final and must be changed as more knowledge is gained about the objects.

(5) Classification should proceed at every stage and as far as possible on one principle. If this principle cannot be used for the entire classification, the properties used at the higher class must be more important than those used in lower classes.

The established analogy between classification, the principles of logic, and regionalization allows a rigorous examination of methodology and results, but the corresponding development of a theoretical framework for regionalization is lacking. Rodoman (1967) declares that the theoretical problem is mathematical. Geographers essentially treat the earth's surface, a positively curved plane or Riemann suface in three dimensional space, as a euclidean surface in two space, and attempt to transform the surface into the one dimensional space of words in sentences in a text. Mathematically the necessary transformations involved do not allow the preservation of spatial ordering (unique location). The framework therefore must be developed in terms of set theory.

Whether old or new in method, the process of regionalization normally

---

[2] Such as employing a symmetric contiguity matrix to modify the grouping algorithm.

[3] D. Grigg (1967), pp. 486–489.

THE FORMULATION OF THE SET-THEORETIC ABSTRACTION

begins by examining the characteristics or attributes of various locations on the earth's surface. In practice these locations are not actually points, for points are zero dimensional. Instead very small areas are studied. For purposes of abstraction, consider infinitesimal areas as geographical locations or points. These geographic units can be equated with OTUs. Beginning with these units, the analysis considers either the attributes of the OTUs or the interactions between them.

It is obvious that each of the geographic units has many characteristics, such as temperature, altitude, and population potential. We can define a geospace, G, of dimension N containing all points on the earth's surface, where N is the number of attributes or characteristics of a unit. In theory there is an infinite number of characteristics of a location so the space G has infinite dimension. However, in practice only a few characteristics are selected for analysis, which defines a finite dimensional subset of G.

There are many types of interaction between two geographic units. Interaction can be considered to be a vector in that it has direction and magnitude. Measures of such dyadic interaction like migration flows and telephone calls assign a number to a pair of points. Dyadic interaction such as migration may be considered as mathematical many to one functions. Define a set of many to one functions F. If f is such a function, f maps two points, elements of G, into a vector or relation (interaction) space R. R has dimension M, the number of f's selected. The f chosen depends upon the measure of dyadic interaction being mapped. Again in practice only a small subset of F is chosen for investigation. These f's operating upon the subspace of G selected generate a subspace of R, so that if $A \subset G$, $f \in F$, $f(A) \rightarrow B$, $B \subset R$. That is, for a set of points, A, on the earth's surface, and for a given f, or dyadic interaction or relation, f operating upon A generates a set of relations, B, which is a subset of all possible relations between all possible point pairs.

Contiguity plays a dual role for not only is it a relation between two units, but it can be considered a characteristic of a unit. However, the abstraction considers units as points and therefore contiguity will be considered to be only a relation between two points.

Mathematically there are five possible relations between two sets, giving three types of regional designs:

(1) $A \cap B = \Phi$. The sets have no points in common. Such a regionalization would be a set of well-defined regions (high closure) of the same order.

(2) $A \cap B \neq \Phi$, $A \cap B \neq A$, $A \cap B \neq B$. The sets intersect, giving a regionalization of overlapping regions.

(3) $A \cap B \neq \Phi$, $A \cap B = A$, $A \cap B \neq B$. $A \subset B$. A is a subset of B.

(4) $A \cap B \neq \Phi$, $A \cap B \neq A$, $A \cap B = B$. $B \subset A$. B is a subset of A.
The third and fourth relations result in a hierarchy of regions (nested regions), such as political regions.

The identity relation:

(5) $A \cap B \neq \Phi$, $A \cap B = A$, $A \cap B = B$, $A = B$.
is of no interest here and will be ignored.

There are, therefore, three types of regional designs possible: well-defined, intersecting, hierarchy.

The abstraction also implies three types of regions. Regionalization within the space G would result in the usual descriptive set of homogeneous regions. Operating upon the set R gives the usual functional regions. A regionalization process could operate upon both G and R simultaneously resulting in a hybrid

region having no identity with either of its parents. These three types of regions are the usual types recognized by many.

THE SPATIAL FIELD THEORY

A recent development in the study of regions has been spatial field theory, which permits the examination of the set of f's between homogeneous and functional regions. Brian Berry (1966b, 1968), noting that the same procedures of numerical taxonomy are used to derive both formal and functional regions, developed the field theory to relate the two regional types, using techniques of systems analysis. The field theory, operating upon a spatial system, involves places, their attributes, and the interaction among them.

Choosing a subset of G, one can develop an n-place, a-attribute, matrix. The matrix describes spatial association and variation over the n places. Berry asserts that the infinite number of attributes of n places and their variation are actually indexed by a finite number of fundamental, independent concepts. Using principal components analysis the number of key factors underlying the total variation is identified. Each observation has a score on each factor, creating a $n \times s$ matrix of n-places and s-factors. This is the structure matrix.

A similar approach is used on interaction data. Choosing the various interactions, f's, for which data are available, and using the same n-places used above, the subset of R space is defined. For each f we can construct an $n \times n$ matrix. Such a matrix can be "unfolded" into an n^2-n array. The arrays for each f can be grouped to form an n^2-n by d interaction matrix. Again, the underlying factors of variation are identified. The factor scores define an n^2-n by b-factor behavior matrix.

The structure and behavior matrix are, however, not enough for formulating a field theory. There is no way to relate the matrices. Defining the similarity between two points as the euclidean distance in s-space, an n^2-n by s matrix of place similarity of structure is developed. This similarity matrix is row-wise comparable with the dyads of the interaction matrix. It is now possible to formulate the relation between structure and behavior.

Two views could be taken: (1) dyadic behavior is a function of characteristics of places, and that changes in characteristics will affect the dyadic interaction, or (2) place characteristics are dependent upon relationships with other places, and changes in relationships would change characteristics of the places. However, as mentioned, places, their attributes, and their interaction form a system. Instead of simple cause-effect relations spatial structure and spatial behavior must be discussed as in a state of mutual equilibrium with very complex interdependencies.[4]

In terms of the abstraction the field theory is the mapping of a subset of G into R and vice versa, or, simply a study of a set of f's and f^{-1}'s.

[4] Canonical correlation allows the field theory to have a mathematical statement. Canonical correlation is an extension of regression analysis. Simple regression has one independent and one dependent variable. Multiple regression has one dependent and several independent variables. The method maximizes the dependence of the two sets. The correlation on various levels are weighted by factors. This allows one to examine the relation of pairs of factors across the equation. If the factors are "paired off" at every level of correlation then we have the simple "Philbrickian world" of alternating functional and uniform regions. However as Berry found, the relations are very complex, and Philbrick's notion is a very special case of the field theory.

METHODS OF REGIONALIZATION

A REVIEW OF PAST METHODS

Considering regionalization as classification, many regionalization methods can be contemplated. Sebastyen (1962, p. 36) has demonstrated that classification may be considered as part of the decision-making processes of pattern recognition. Indeed, in the general sense, regionalization is the recognition of a pattern in the available data. A method, of course, is an orderly procedure; an algorithm is a simple computational process or procedure, and therefore is a special case of a method. Pattern recognition methods used by geographers for regionalization may be grouped into two categories: (1) the testing of an *a priori* classification, and (2) the development of a classification. The first category includes use of chi square tests, analysis of variance, and discriminant analysis. Factor analysis and grouping algorithms form the second category.

TESTING AN *A PRIORI* CLASSIFICATION

The chi square test has been used by Zobler (1957). His aim was to statistically test the implication that there is a relationship between the data used to draw boundaries and those used to describe the regionalization. The chi square was used to test the expected distribution in each of the regions with the actual distribution of field data with the null hypothesis that no relation exists. That regionalization was chosen that produces the highest significant chi square. In a later paper Zobler (1958) employs the same method in allocating states in a regionalization of the United States. Both Zobler's papers received much comment, notably by Berry (1958) and Mackay (1958) on the use of the chi square test. They objected to the use of relative frequencies. Berry demonstrated that such relative frequencies may be used to establish categories, but actual frequencies must be absolute, independent events, not transformable to other units. Variance analysis has also been used by Zobler (1957) as an extension of chi square analysis, testing both intra- and inter-regional variance. An F test is made with a null hypothesis that there is no significant difference between the two variances. If the chosen F value is not exceeded then the regionalization is not valid. The variance analysis approach, since it accounts for both intra- and inter-regional variance, is more rigorous than the simple chi square approach.

Both chi square and analysis of variance procedures have received very little attention in the last several years. Operationally the methods are very cumbersome, for to derive the "best" regionalization, all possible pairs of objects would need be tested and secondly, the choice of probability level is subjective. Other more rigorous methods have been sought.

Discriminant function analysis computes the vectors associated with the latent roots, λ, of $|W^{-1}A-\lambda I| = 0$, where W is the matrix of pooled within-group deviation scores cross products, A is the between-groups cross products of deviation from grand means weighted by group sizes, and I, the identity matrix. The method can only refine and test the goodness-of-fit of an existing classification. The best example of the method is represented by Casetti's study (1964) of climatic regions.

THE DEVELOPMENT OF A CLASSIFICATION

The second category contains three types of regionalization methods: (1) factor component analysis plus mapping, (2) factor component analysis plus grouping analysis, (3) grouping analysis.

Factor analytic methods reiteratively process the eigenequation $(R-\lambda_i I)a_i =$

$= 0$,where R is a correlation matrix, I the identity matrix, λ the latent root and a a column vector of factor loadings, to eliminate redundancies in the data by developing a new orthogonal basis for a space in which the observations are distributed. Gower (1966) has shown some of the relations between various latent root methods. The distances between points in a factor score space are shown to be the Mehalanobis D^2 with a singular dispersion matrix. The classic paper using this first method is Kendall's study (1939) of crop productivity in England, mapping results of coefficients greatly similar to those produced by factor theory. Hagood (1941) mapped factor loadings to derive a regionalization. Later work in such an approach has been by Thompson, Sufrin, and Buck (1952), who mapped factor loadings in their study of New York. A variant of this method which crosses over into the next branch is a paper by Goodall (1954), who mapped factor scores.

Berry is the chief contributor to the second branch with a large number of studies employing the same method of factor analysis to shrink the dimensionality of a variable space and then grouping the factor scores according to a selected algorithm. This method is outlined in several articles (1959, 1961, 1966a), with specific examples such as a study of India (1966b). Russett's use (1965) of R- and Q-mode factor analysis where the factors represent groupings of objects, represents a variant method similar to grouping analysis.

The success of the factor analytic method with mapping is limited by the subjectivity of the analysis. The second method has great objectivity and, operationally, the capacity to handle a large number of variables, but its success is greatly dependent upon the grouping algorithm chosen.

Grouping methods have been very successful in the development of classifications. Objects are grouped using some measurement of pairwise association. Any measurement that forms a metric space may be used.

A space X, under a measurement of association d, is called a metric space if for points x, y, z, in X: [5]

(1) $d(x, y)$ is a real valued function of two variables.
(2) $d(x, y)$ is greater than or equal to zero, and $d(x, y) = 0$ if and only if $x = y$.
(3) $d(x, y) = d(y, x)$.
(4) $d(x, y)$ is less than or equal to $d(x, z) + d(z, y)$ (the triangle inequality).

If the space formed is not metric results may be misleading if not contradictory.

The most common example of a measurement of association that generates a metric space is simple euclidean distance, defined as:

$$d_{ij} = \sum_{r} [(s_{ir} - s_{jr})^2]^{1/2}$$

for distance in r-space for points i and j members of s.

Sokal (1961) and others have discovered and defended the distance measurement of association as the best methodologically as well as conceptually. Besides euclidean distance, there are many other measurements of similarity including noneuclidean distance. McQuitty (1957), a sociologist, bases his grouping upon the correlation matrix. Boolean relations in matrix form also might be applied to regionalization, although no attempt yet has been made. [6]

Stone (1966) made use of a linking method in his study of the United Kingdom. Based on eleven variables, forming an eleven-space, distance between points was taken as a similarity measurement. Since the variables were not or-

[5] A complete discussion is given in L. Blumenthal (1953).

[6] The mathematics have been worked out by D. Rosenblatt (1967).

thogonal, he had to compute generalized distances to offset the affects of correlations. Several different analyses were attempted with a final determination that use of normalized data removed undesirable size effects. Berry has used the grouping technique for several years in a variety of studies,[7] making a variety of modifications to the simple linking method. In further developments Kaiser (1964) has constructed an algorithm to account for area of region at each step to seek similar and compact regions. This development was carried one step further by the Regional Science Research Institute's study of Pennsylvania (n.d.) by seeking nodality and compactness as well as homogeneity of the regions. Rubin (1965) has developed an inverse algorithm which divides the universe into similar regions.

SINGLE LINKAGE ALGORITHMS

The so-called "single-linkage" algorithms are by far the most important clustering methods. Pairs of points are joined by steps according to a rule until all points are in one group. The method is conceptually simple and easy to develop. Since only single links are possible, regions cannot overlap.

Mathematically the methods begin with a finite set S of k members which can be partitioned into one to k subsets. The ordered set of such partitions, a (complete) classification, is noted as $C(S)$.[8] Each partition, P_r is assigned a rank, rank one being k subsets, $r = k$ being the whole set and having rank k. If a classification is strictly hierarchical then for every group G of rank one there exists a sequence $G_1 \ldots G_{r+1} \ldots G_k$. The rank of classification then is the number of partitions.

For each point, s, to be classified, we can define a set of neighbors $N(s)$. A measurement of association exists for all neighbor pairs $f(g,h)$, and only for neighbors. The measurement f must form a metric space. A point is not its own neighbor. When points or groups are joined, their neighborhood is defined as the union of their neighborhood sets.

Given two groups G and H, if there is some $g \in G$, $h \in H$ such that $h \in N(g)$, then G and H are called neighbor groups. The set of all such pairs common to both sets is called the interface, $I(G, H)$.

A group may have an interior defined as the set of all the $g \in G$ such that $N(g)$ is a proper subset of G. The interior may be the empty set. Points not in the interior of the group can be considered to be the boundary of the group. Note that these definitions are not strict in the topological sense, but are unique to this discussion.

Two single linkage methods in common use are the centroid and Ward's grouping algorithms. Both algorithms, special cases of the general statement above, begin with the generation of a k by k matrix of associations between *all* pairs of points. Initially each point has $k-1$ neighbors, decreasing with each step. An ordered set, $C(S)$, is produced. The interface examined at each step in the grouping procedure comprises *only* the remaining pair of points which is most similar.

---

[7] The method is outlined in the papers referenced in the factor analytic section of this paper.

[8] This discussion is based, partly, on several discussions during 1967 with Peter Neely and on his two papers (n.d.).

The centroid and Ward's algorithms examined in this study seek the minimum squared euclidean distance as the measurement of association,[9] although many other measurements may be used. Since euclidean distance forms a metric space, the measurement is symmetric, $d(x, y) = d(y, x)$, and only the upper left triangle of the matrix, omitting the diagonal, need be examined. Contiguity can be introduced into the analysis at this point. If two areas are contiguous the appropriate entry is scored negative, otherwise it is left positive. In effect contiguous areas are made "more similar". The two methods begin by scanning the distance squared matrix. The results of the grouping algorithm may be displayed in the form of a tree graph, showing at each step which points are joined. Examples are given in the analysis portion of this paper.

The centroid method, the simplest algorithm of those studied here, joins at each step that pair of points (i,j) for which the measurement of association is a minimum. The appropriate i-th row and column and j-th row and column are deleted from the matrix and replaced by a new point which is the centroid of the pair (or group). The process is repeated until all the points are grouped into one set. This method would be expected to work well when the data are already well patterned, but when groups are highly dispersed the grouping may become unstable. This is due to what is termed the "chaining" problem. Suppose there are several points in a row, evenly spaced. If one point is introduced, closer to one point than the others are to each other, as with point eight, the conditions are set up for "chaining".

1 2 3 4 8 5 6 7

Points four and eight would be joined first, for they are most similar. However, their centroid lies between four and eight and closer to five than to three. Point five would then be joined to the group (4, 8) and a new centroid placed somewhere between (4, 8) and five. Point six would next be joined and so on down the line. Only after seven was joined to the group would it be possible to join points one, two, and three. This process of the centroid dragging a group in one direction is the "chaining" problem. On a more general scale it could force a group to be dragged to a larger group when actually the former group was more similar to a very different group. The centroid grouping works well on well-patterned or densely-packed data, but the chaining problem exists when data are dispersed.

In an attempt to develop a better algorithm Ward (1963) developed a routine that examines the entire matrix and joins that pair which makes the minimum increment to the pooled within-group sum squared distances. Within-group distances are checked at each step. The matrix is updated as above, but uses the group mean. By minimizing the sum of the squared deviations about the group mean, the procedure can: (1) maintain groups of nearly equal size, (2) maintain groups of high density, and (3) develop groups that are spherical in shape. The algorithm is better than the centroid grouping mathematically, for it can work in less patterned data and draw out a clean grouping. But since it employs the group mean for updating groups it suffers to some degree from the same chaining problem as the simpler centroid method.

Operationally the centroid and Ward's algorithms have several problems in common. Because of the larger size of the distance-squared matrix and the size of present computer storage, the sample size that can be handled operationally

[9] A possible variation is seeking the maximum, instead of the minimum, squared euclidean distance. Using the minimum criteria the most similar objects are grouped; the maximum criteria groups the most dissimilar.

is very limited.[10] Scanning the entire matrix at each step emloys a large amount of computer time. The size of groups, their number and shape, remain methodological problems. Also, since the grouping produced is an entire, ordered set of partitions, $C(S)$, there is no absolutely objective method to determine the level (s) at which to select regions, although the recent work by Hartigan (1967), and Friedman and Rubin's use of Wilk's criterion (1967), holds much promise for such an optimality measure.

One approach to the problem of the determination of the optimal number of groupings has recently been presented by Semple, Casetti and King (1969). The method they propose is designed to partition a finite set of items into an optimal number of groupings and then to achieve an optimal assignment of items to groupings. It differs from other grouping algorithms in that an optimal number of groupings is arrived at rather than being assumed *a priori*. The structure of the procedure is illustrated and some assumptions for its correct application are summarized.

Suppose that a set of items is given, and that each item is identified by the values of a number of variables. For simplicity, measurements on only two variables are assumed for each item, but no conceptual difficulty is introduced when a larger number of variables is used. Assume that the two variables are associated with the rectangular axes of a Cartesian diagram. Then each item is a point in the two-dimensional space and the similarity of any two items with respect to the variables involved, is measured by the distance between the two points. Hence, a cluster of points in the diagram identifies a subset of items which are more similar to one another than to other items outside of the cluster.

Grouping procedures aim at dividing a set of items into groups in such a way that the items in any one group are more similar to one another than they are to items in the other groups. These procedures, therefore, should be capable of identifying clusters of points such as were mentioned above.

The procedure outlined here aims at identifying clusters of point images and then determining which item belongs to which cluster. The method assumes that clusters of points do exist in the observation space. If the points, in fact, are distributed either randomly or uniformly in the space,[11] then an optimal grouping becomes meaningless.

The classification procedure proposed involves the following steps:

(1) The ranges of the variables are determined. Clearly, these values define the space which includes the point images of all the items. A fine rectangular grid is superimposed over this space and the number of points in each grid cell is recorded. The centers of the grid cells, therefore, can be associated with the frequencies of points in the cells. The cells which include part or all of clusters naturally will have higher frequencies than the other cells. These frequencies can be considered as measures of a two-dimensional spatial trend with clusters corresponding to points where relative maxima of the trend occur. Thus, the problem of identifying clusters can be translated into the problem of identifying the number and location of the relative maxima of this trend.

(2) The extremum points of the trend are located approximately by applying an adaptation of an algorithm developed in another paper.[12] This algorithm determines the following:

---

[10] A 32K computer can handle only 220 points.

[11] For the testing of randomness in point patterns see: M. F. Dacey (1963, 1964a, 1964b), and tests for clusters see C. Mack (1950) and J. Nans (1965).

[12] E. Casetti (1968).

(i) the distances from the centers of cells to each grid intersection point,
(ii) the transformation $d = 1/(d+1)$ of these distances,
(iii) the correlation, for each grid point, between these transformed distances and the frequency values for the cells,
(iv) the largest absolute value of these correlation coefficients and grid intersection to which this correlation coefficient relates. (This is taken to indicate, as a first approximation, the origin of the largest trend, and hence, the approximate location of the cluster),
(v) the regression of the cell frequencies on the transformed distances from the grid intersection identified in the previous step, and
(vi) for each cell the residual frequency resulting from this regression.

On the residual frequences, the operations (i) through (vi) are repeated to generate successive approximations for the locations of other relative maxima of the trend. The procedure is terminated when the fraction of the total frequency variance explained by an additional iteration is smaller than a predetermined threshold.

As a result of this second step the number and approximate cores of the clusters are identified.

(3) The point images of the items are then assigned to the nearest cluster core. Euclidean distance is used as the criterion for this assignment.

(4) An optimal assignment of the items to groups is obtained by applying a discriminant iterations technique.[18] Specifically, the grouping presented in the third step is improved as follows:

(i) the centroids of the point images of the items in each grouping are calculated,
(ii) all point images are assigned to the nearest centroid, so that a new grouping is obtained,
(iii) steps (i) and (ii) are repeated until two successive iterations generates the same grouping.

(5) Artificial groups have still to be eliminated. The steps 1 through 4 above may generate artifical groups in several ways, the first of which is illustrated in Fig. 1. Here the procedure has generated a centroid related to a cluster of clusters rather than to clusters of point images.

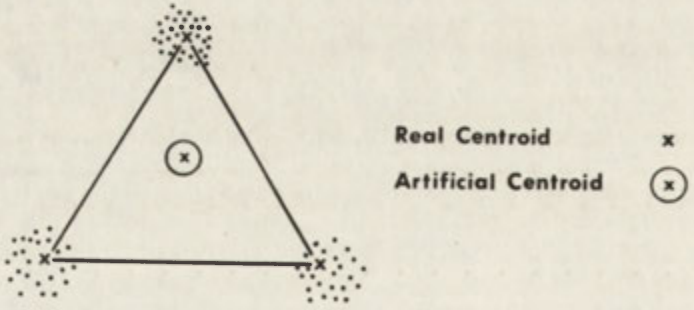


Fig. 1. Type one artificial group

Four cluster centroids have been generated, three located in clusters and the fourth is an artificial one located in a central position. Artificial centroids are usually generated if the natural groupings tend toward geometrical arrange-

[18] E. Casetti (1964).

ments in such a way that large negative trends are identified in the point images. For instance, in the diagram real clusters tend to have cores corresponding to the vertices of an equilateral triangle. Consequently, the first centroid to be identified would most likely be located central to the three clusters since this would be the core of a large negative trend. This implies, of course, that the further from this core the greater becomes the frequency of point images. This situation will require special attention in the present procedure.

A second type of artificial group may appear as a result of random trends being generated in steps 1 through 4. Assume a real cluster centroid has been identified and the regression analysis has yielded residuals which are to form the terms of a second spatial frequency series. This second spatial series must contain negative residuals, and these seem to imply some sort of negative frequencies. Suppose that these negative residuals accumulate after a number of iterations. It is possible that by chance a trend may appear that is identified to be significant, in the sense that the explained variance of the trend may be within the threshold limit imposed on the analysis. This trend also is artificial and the location of the trend-maximum purely random. Consequently, items may be assigned to this random centroid but the group thus formed is undesirable. This is illustrated in Fig. 2.

Fig. 2. Type two artificial group

Assume that two real groups have been identified with centroids lacated at *a* and *b* and a third significant group is identified with a centroid *c* located at random. Items that should be assigned to centroid *a* are actually closer to *c* and hence are assigned to *c* forming a third type of artificial group.

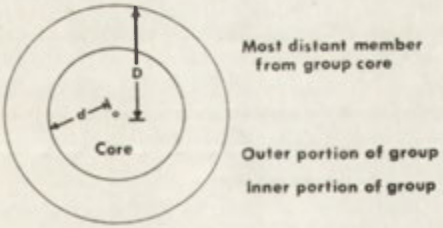


Fig. 3. Artificial group criterion

In order to identify and eliminate artificial groups a "decision rule" is defined in the following manner. An artificial group is simply a group that has more than one-half of the member items located in the outer portion of the group. For example, suppose that the most distant member of a group is thought to be located on the circumference of a circle, centered at the group core, and with radius *D* being the distance the member is located from the core (Fig. 3). Let a second circle be defined, centered on the group core and with

radius d, where $d = (D^2/2)^{0.5}$. This defines the area of the smaller inner circle, representing the inner portion of the group, to be one-half the area of the larger circle, representing the outer portion of the group. A group is now considered artificial if more than one-half of the group members are located in the outer portion of the group.

At this stage, the procedure indicates how many items have been assigned to the inner and outer portions of all groups. In this fashion real groups are identified and retained, while artificial groups are eliminated and their members reallocated.

The result of the procedure is the partitioning of a finite set of items into an optimal number of groupings, and the simultaneous optimal assignment of the items to the groupings.

AN EMPIRICAL APPLICATION

In the examples that follow, small towns in southern Ontario (Fig. 4) are grouped according to dimensions of viability.[14] Twentyfive variables all related to economic viability were collected for each town and a factor analysis was

Fig. 4. Selected towns in the study area

used to obtain three orthogonal dimensions of viability (Table 1). The three dimensions were identified as being related to (1) growth, (2) subsidization from the Federal and Provincial governments, and (3) level of spending on community services. The towns are grouped according to their relative position on these

[14] For a description of the study area see R. K. Semple (1966).

TABLE 1. Factor scores for small towns in southern Ontario

Factor Scores	1	2	3	Factor Scores	1	2	3
Towns				Towns			
1 Acton	−131	−158	139	2 Alexandria	093	−020	066
3 Alliston	−232	−054	089	4 Almonte	068	−108	−010
5 Amherstburg	−053	−138	097	6 Arnprior	−001	−151	−095
7 Arthur	275	066	−073	8 Aurora	−550	−118	097
9 Alymer	−037	−058	033	10 Bancroft	−451	414	−397
11 Barry's Bay	026	405	556	12 Beamsville	−132	−006	011
13 Beaverton	118	268	−083	14 Belle River	001	−252	−087
15 Blenheim	150	015	−012	16 Bobcaxgeon	098	189	054
17 Bracebridge	098	−005	085	18 Bradford	−040	017	046
19 Bridgeport	−142	−098	115	20 Brighton	−026	121	−063
21 Caledonia	019	−045	039	22 Campbellford	119	069	004
23 Cardinal	−064	−064	291	24 Carleton Place	150	−003	011
25 Casselman	091	194	193	26 Chesley	348	134	275
27 Chestervillle	071	017	158	28 Chippawa	−298	−198	209
29 Clinton	063	−129	−078	30 Colborne	127	141	−080
31 Crystal Beach	112	−171	−169	32 Delhi	103	−089	−167
33 Deseronto	169	−035	−049	34 Dresden	118	022	−088
35 Dunnville	021	−064	038	36 Durham	146	148	−004
37 Eganville	212	−007	049	38 Elmira	−013	−133	−022
39 Elora	213	−019	022	40 Essex	107	−188	−076
41 Exeter	076	−010	−020	42 Fenelon Falls	−086	476	−062
43 Fergus	001	−074	−028	44 Fonthill	−353	−304	119
45 Forest	183	142	105	46 Frankford	025	035	151
47 Gananoque	007	−146	−055	48 Georgetown	−576	−206	−056
49 Goderich	077	−100	−076	50 Gravenhurst	147	−071	−028
51 Grimsby	−213	−166	−030	52 Hagersville	096	020	071
53 Hanover	−026	−093	028	53 Harriston	169	126	−046
55 Harrow	129	−021	022	56 Havelock	182	089	006
57 Hespeler	−003	−204	064	58 Huntsville	156	−036	−034
59 Iroquois	−042	−122	002	60 Kemptville	021	035	074
61 Kincardine	170	039	−050	62 Kingsville	065	−064	−094
63 Lakefield	110	−044	−045	64 Listowel	105	−000	−056
65 Little Current	143	−001	−351	66 Madoc	090	121	323
67 Markdale	146	150	−066	68 Markham	−1348	223	151
69 Marmora	093	086	014	70 Mattawa	223	−059	−064
71 Meaford	092	021	−040	72 Milton	−320	−076	−003
73 Milverton	249	055	−058	74 Mitchel	113	194	−059
75 Morrisburg	135	−054	−067	76 Mount Forest	196	074	−082
77 Napanee	095	−076	−003	78 New Hamburg	082	−054	−076
79 Niagara	−078	−229	035	80 Norwich	260	162	−027
81 Orangeville	−076	−121	049	82 Palmerston	196	130	025
83 Penetanguishene	115	027	012	84 Petrolia	026	−017	132
85 Picton	114	−187	020	86 Port Credit	−357	−463	−034
87 Port Dover	075	−030	−149	88 Port Elgin	145	061	−027
89 Port Perry	−002	088	113	90 Port Stanley	230	006	−017

Cont. Table 1

Factor Scores	1	2	3	Factor Scores	1	2	3
91 Prescott	−159	−124	036	92 Richmond Hill	−2145	427	−374
93 Ridgetown	223	−045	−128	94 Rockland	−070	099	115
95 Seaforth	218	048	−057	96 Shelburne	260	363	−128
97 Southampton	213	060	−073	98 Stayner	219	128	−096
99 Stirling	177	056	−011	100 Stoney Creek	−801	−326	−033
101 Stouffville	−481	367	356	102 Strathroy	039	−077	−033
103 Streetsville	−995	−122	021	104 Sturgeon Falls	−077	−083	−006
105 Sutton	121	080	−099	106 Tecumseh	−038	−216	129
107 Tilbury	154	−086	−173	108 Tweed	165	005	029
109 Uxbridge	017	039	−074	110 Vankleek Hill	153	147	142
111 Walkerton	−015	−107	−035	112 Waterdown	−123	007	052
113 Waterford	180	093	007	114 Watford	139	241	105
115 West Lorne	165	140	087	116 Wheatley	147	−068	−073
117 Wiarton	137	057	022	118 Winchester	148	057	105
119 Wingham	221	−049	−152	120 Woodbridge	−171	−236	116

three orthogonal measures of viability by (1) the centroid, (2) Ward's and (3) Semple's optimal grouping procedures respectively. The same data bank was utilized in each case.[15]

THE CENTROID GROUPING

The results of the centroid grouping procedure are shown in Fig. 5. The tree is very complex with no definite major groupings until step 86. Steps 1 to 85 consist of the development initially of many small groups. For example, at step 50 there are 24 groups with an average membership of 2.1 points. The groups are very dense as shown by the value of the joins. After about step 50 the major group nuclei are well defined and points are slowly to these groups. The join values and average group membership slowly increase as the group centroids wander in the metric space.

A sharp jump in join values occurs at step 86. At step 85 there are 13 groups with an average membership of 6.6 points, with 7 relatively large groups, *A*, *B*, *F*, *H*, *I*, *D*, and *E*. (See Fig. 5) The average factor scores or group centroids of the 13 groups are in Table 2. Note that the 86 cities grouped at step 85 form 13 distinct groups in the 3 factor space. If the investigator is interested in the identification of the relatively large number of small distinct core groups, he would stop the analysis at this point and allocate the ungrouped cities to the defined groups.

[15] The actual groupings for procedures (1) and (2) were performed by the IBM 7094/360-65 facilities of the University of Chicago's Institute for Computer Research. Procedure (3) was performed by the IBM 7094/360–75 facilities at the Ohio State University. The centroid and Ward's algorithms are contained in a program written by Neely. The optimal grouping algorithm written by Semple may be found in Ohio State's Discussion Paper Number 10 footnoted previously. All of the algorithms are available on UCLA's Campus Computer Network IBM 360–91 facility.

TABLE 2. Group Centroids at Step 85

Group	Factors			Number of members
	I	II	III	
A	−0.74	−1.66	0.90	7
B	0.32	0.23	1.01	10
C	−2.76	−0.65	0.43	2
D	1.06	−0.55	−0.50	15
E	−0.10	−0.96	−0.11	12
F	1.84	1.04	−0.65	14
G	−1.27	0.00	0.31	2
H	1.58	0.34	0.04	17
I	1.40	1.59	1.13	7
J	−0.05	0.80	−0.68	2
K	1.10	−1.13	−1.47	5
L	−1.86	−1.45	0.03	2
M	2.22	−0.51	−1.15	3

TABLE 3. Group membership at step 85

Group	Towns by number															
A	1	5	19	57	79	81	106									
B	2	17	18	27	46	52	60	84	89	94						
C	3	72														
D	4	29	33	41	49	50	58	62	63	64	71	75	77	78	116	
E	6	9	21	35	38	43	47	53	59	102	104	111				
F	7	30	34	36	54	67	73	74	76	80	95	97	98	105		
G	12	112														
H	15	22	24	37	37	39	55	56	61	69	82	83	88	90	99	
	108	113	117													
I	16	25	45	110	114	115	118									
J	20	109														
K	31	32	40	87	107											
L	51	91														
M	70	93	119													

Grouping continues at step 86 by joining groups *F* and *H*. Even though the two groups are distinct, particularly on the second and third factors, the *F* and *H* group centroids are the close pair of points at this step. The grouping enlarges existing groups then, by adding additional points.

The next distinct change in join value occurs at step 97 when the enlarged *D* and enlarged *E* groups are joined. At step 96 there are 11 groups with an average membership of 8.8. The three large groups, $D+K+M$, *E*, and $F+H$, have 23, 14, and 31 points each respectively. At step 97 there are two large groups

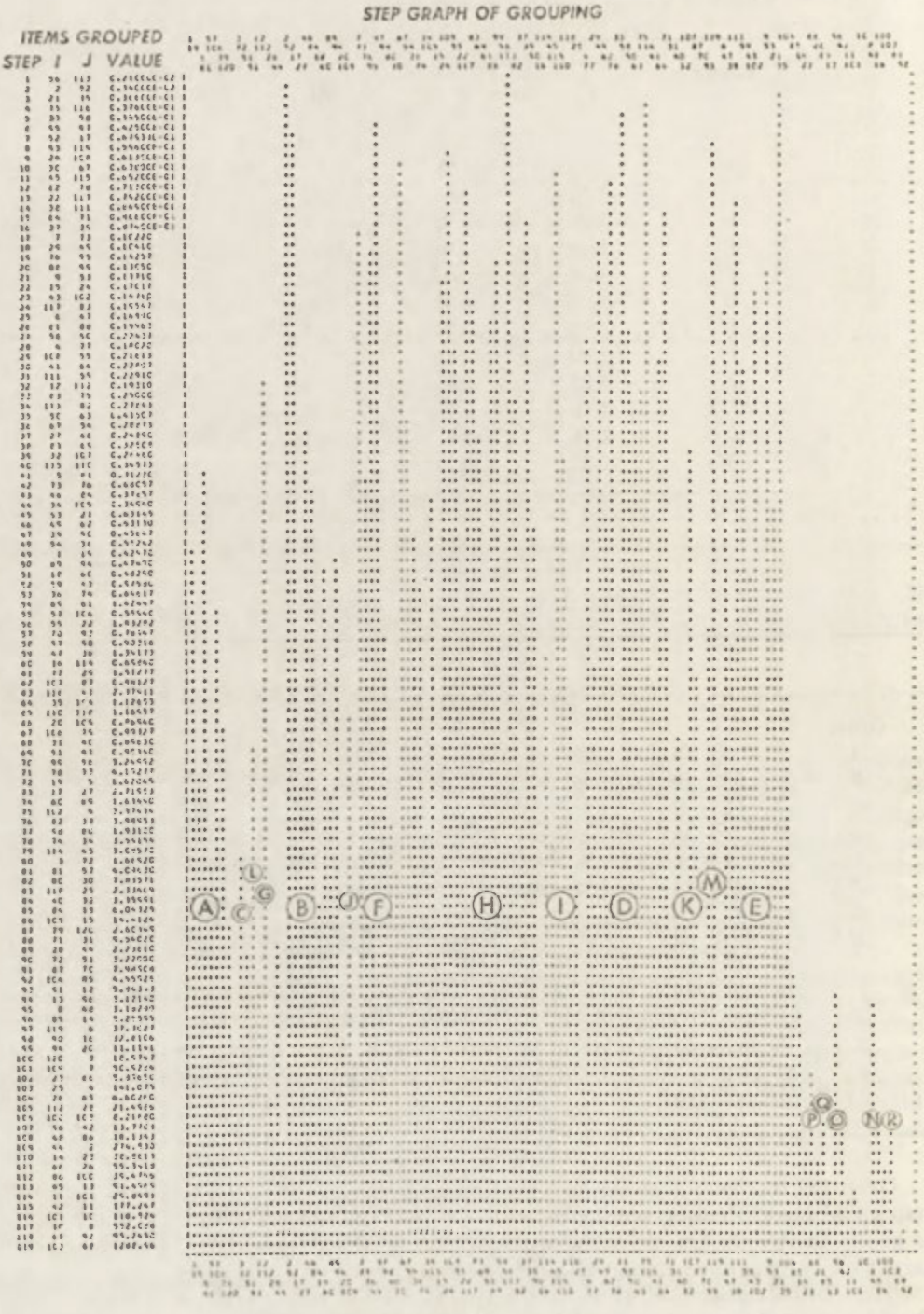


Fig. 5. Centroid grouping

about equal in size and ten small groups. Other "hills" in join values occur at step 101 when $B+J$ is joined to $F+H+I$ and at step 103 when $B+J+F+H+I$ is joined to $D+K+M+E$. There are three major groups at step 101 and two major groups at 103.

The next relatively large hill is reached at step 109 when the two major groups, $A+C+L+G$ and $B+J+G+H+I+D+K+M+E$ are joined. At the previous step there are seven groups, two of them major.

TABLE 4. Group centroids at step 108

Group	Factors			Number of members
	I	II	III	
enlarged *A*	−1.58	−1.45	0.77	16
enlarged *B*	1.08	0.06	−0.14	87
N	−4.94	−2.62	0.02	3
O	0.97	3.69	−0.91	3
P	0.13	0.28	3.07	2
Q	2.45	0.66	−3.13	2
R	−8.98	−2.24	−0.06	2

Note that the small groups represent strongly isolated groups in the three factor space.

As the grouping process continues *P*, *Q*, and *O* are successively joined to the extremely large $A+B$ major group. At step 117 there is an extremely sharp rise in join values as the isolate group $N+R$ is joined to the one large group. The final large leap in values comes at the final step when the strongly isolated points (68 and 92) are added to the large group. The final group of two isolated points have average factor scores of −17.40, 3.25 and −1.11 on each factor respectively.

As with all single linkage algorithms there is no totally objective criteria available to decide where to "cut" the tree. If the investigator is interested in a large number of distinct core groups he would stop the grouping at step 85. If generality is important, few groups with nearly all points allocated, the tree could be cut at step 108 where 109 points are grouped. Studying the entire tree to allocate all points there are actually four groups at step 108; enlarged *A*, enlarged *B*, $N+R$ plus other strongly negative isolates, and $O+P+Q$.

It is necessary to refer to the original factor structure to interpret the partition suggested above based on step 108. Group A towns are moderately low on factor one, suggesting better than average growth rates. The same group is low on the subsidy or second factor, and offers average spending on community services indicated by the average score on the third factor. Spatially these viable rural truck garden towns are located just beyond convenient daily commuting distance from those metropolitan centers extending from Windsor, Niagara Falls and Kingston to Toronto. Most of the towns in the *B* group have stagnant growth, indicated by a moderately high score on the first factor, and have average government subsidy and average level of community services, both of the latter being indicated by average scores. These towns in general occupy the more rural farming areas of southern Ontario. The *N* and *R* isolates are the fast growing dormitory suburban towns located within easy commuting distance of their respective metropolitan center. Group *O* clearly shows the

towns receiving heavy subsidies; *P* consists of those towns with extremely high level of local services. Group *Q* is composed of towns with low growth and a poor level of services and a moderate level of subsidization. Groups *O*, *P* and *Q* occupy cores of some of the poorest farming areas in southern portions of the province.

TABLE 5. Group membership at step 105

	Towns by number											
Group												
A	1	3	5	12	91	106	112	120				
B	2	4	6	7	9	14	15	16	17	18	20	21
	22	24	25	27	29	30	31	32	33	34	35	36
	37	38	39	40	41	43	45	46	47	49	50	52
	53	54	55	56	58	59	60	61	62	63	64	67
	69	70	71	73	74	75	76	77	78	80	82	83
	84	85	87	88	89	90	93	94	95	97	98	99
	102	104	105	107	108	109	110	111	111	113	116	117
	118	119										
N	8	48	86									
O	13	42	96									
P	23	66										
Q	26	65										
R	101	103										

THE WARD'S GROUPING

The complicated tree of Fig. 6 is the result of the Ward's algorithm. Group size is very stable and persists until very late in the grouping. The join values increase monotonically as is characteristic of the method with no abrupt change. The investigator could stop the grouping at almost any step if he is interested in a certain number of partitions. However, from the tree, note that step 107 is the last step that preserves the identity of the several large core groups, *A*, *B*, *C*, *D*, and *G*.

The major groups are about equal in membership. As the grouping continues, the large groups are linked until at step 116 there are three partitions, *A*+*B*+*C*,

TABLE 6. Group centroids at step 107

Group	Factors			Number of members
	I	II	III	
A	−1.35	−1.52	0.70	12
B	0.05	0.28	0.97	16
C	0.21	−1.19	−0.25	18
D	1.83	0.53	−0.16	24
E	−4.27	−2.58	0.67	5
F	1.31	1.98	0.18	15
G	1.37	−0.33	−1.04	23
H	−8.98	−2.24	−0.06	2

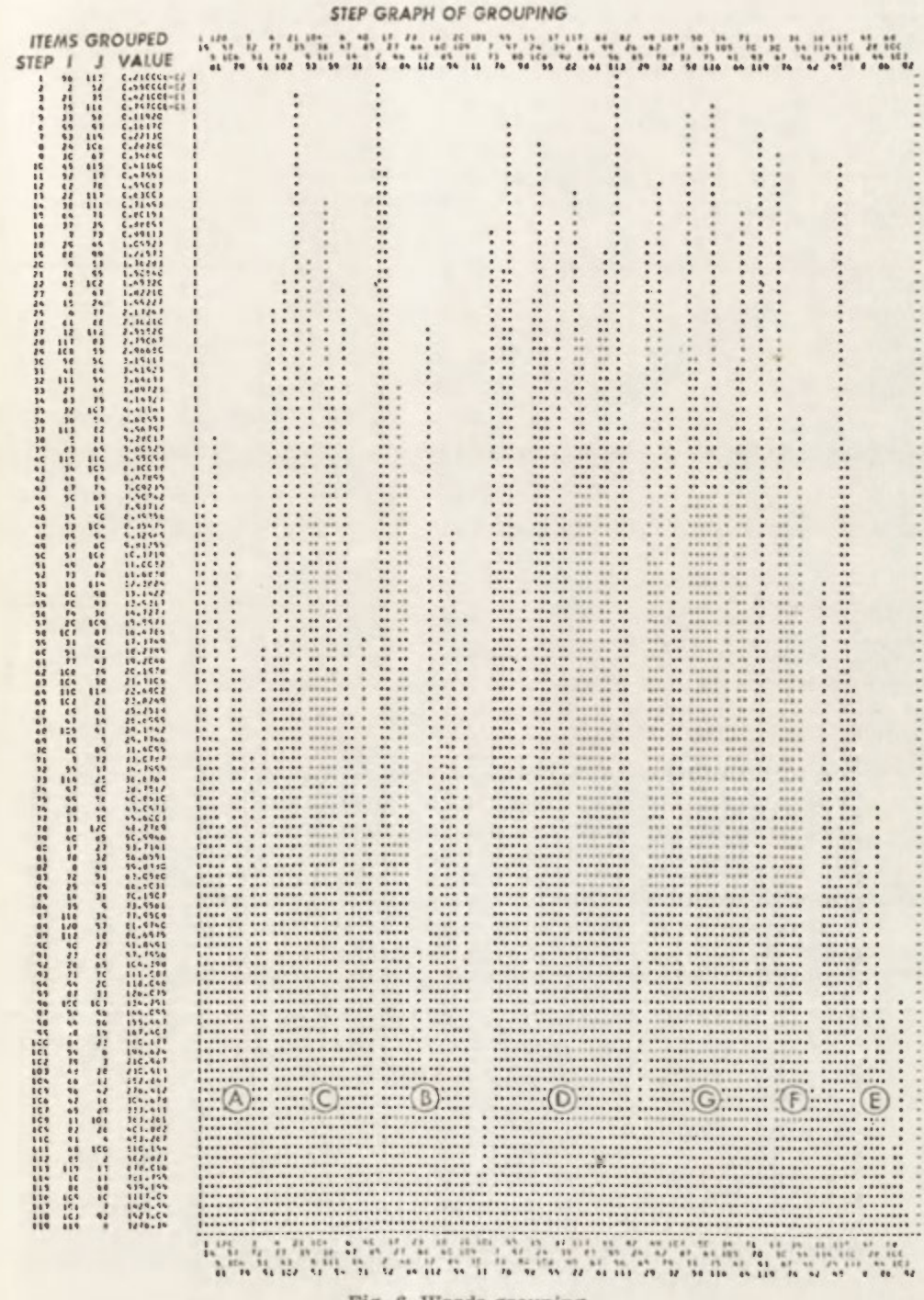


Fig. 6. Wards grouping

TABLE 7. Group membership at step 107 — Ward's

A	1	3	5	19	51	57	72	81	91	106	120														
B	2	12	17	18	20	23	27	46	52	60	66	84	89	89	94	109	112								
C	4	6	9	14	21	31	35	38	40	43	47	53	59	77	85	102	104	111							
D	7	15	22	24	37	39	55	56	61	69	73	76	80	82	83	88	90	95	97	98	99	108	113	117	
E	8	28	44	48	86																				
F	13	16	25	30	36	42	45	54	67	74	96	110	114	115	118										
G	26	29	32	33	34	41	49	50	58	62	63	64	65	70	71	75	78	87	93	105	107	116	119		
H	100	103																							

D+*G*+*F*, and isolates. The linked large groups are stable through even the final few steps till step 117. Step 116 represents the final stage preserving any core group identity; after that step there is only one very large group and isolates.

The first group of towns has a factor structure that indicates average growth, moderately low subsidization, and an average level of community services. The membership is essentially an enlarged group *A* of the centroid algorithm. The spatial arrangement is therefore somewhat similar, except that the average distance of the centers from the metropolitan areas is greater.

TABLE 8. Group centroids at step 116

Group	Factors			Number of members
	I	II	III	
A+*B*+*C*	−0.25	−0.76	0.42	49
D+*E*+*F*	1.53	0.56	−0.40	62
Isolates	−6.60	−1.89	0.59	8

TABLE 9. Group membership at step 116 — Ward's

A+*B*+*C*																									
1	3	4	5	9	19	21	35	38	43	51	53	57	59	72	77	79	81	91	102	104	106	111	120		
2	6	10	11	12	14	17	18	20	23	27	31	40	46	47	52	60	66	84	85	89	94	109	101		
D+*E*+*F*																									
7	15	22	24	37	39	55	56	61	69	73	76	80	82	83	88	90	95	97	98	99	108	113	117		
13	26	29	32	33	34	41	49	50	58	62	63	64	65	70	71	75	78	87	93	105	107	116	119		
16	25	30	36	42	45	54	67	74	96	114	115	110	118												
Isolates																									
8	28	44	48	68	86	100	103																		

The second partition shows high scores on the first factor, indicating extremely low growth; average subsidies, and average community service. The large group has nearly the same membership and spatial structure as the *B* group of the centroid method.

The isolates, from this method, contain only those towns with extremely rapid growth. Virtually all are dormitory centers.

THE OPTIMAL GROUPING

This algorithm identified three significant groups of towns. Each group is identified by its centroid (Table 10) and by the towns which are assigned to it (Table 11). Group *A* contains seventy-two towns, *B* forty, and *C* only eight. These three groups accounted for 89 percent of the variability on the three viability dimensions.

TABLE 10. Group centroids

Group	Factors			Number of members
	I	II	III	
A	1.38	0.62	−0.12	72
B	−0.65	−1.26	0.29	40
C	−9.19	0.83	−0.29	8

Group *A* is associated with those towns that have a high positive score on factor one, and an average score on factor two and factor three. This group is essentially the same as the second grouping of both the previous methods except that it is of intermediate size. Consequently, towns of Group *A* are characterized by slow growth, moderate subsidization and a moderate level of spending on recreation and community services. The four towns of this group which are most similar to the group centroid and hence most representative of the group

TABLE 11. Group membership

Group	Towns by number											
A	2	7	11	13	15	16	17	20	22	24	25	26
	27	30	32	33	34	36	37	39	41	42	45	50
	52	54	55	56	58	60	61	62	63	64	65	66
	67	69	70	71	72	74	75	76	77	78	80	82
	83	87	88	89	90	93	95	96	97	98	99	105
	107	108	109	110	113	114	115	116	117	118	119	
B	1	3	4	5	6	9	12	14	18	19	21	23
	28	29	31	35	38	40	43	44	47	49	51	53
	57	59	72	79	81	84	85	86	91	94	102	104
	106	111	112	120								
C	8	10	48	68	92	100	101	103				

are Wiarton, Port Elgin, Campellford and Stirling. It is noteworthy that these four towns occupy the core regions of the two poorest and most rural farming areas of southern Ontario. Port Elgin and Wiarton in Grey County represent marginal farming areas in a zone of poor quality soils and limestone outcrops. They are the farthest removed from the large urban markets to the south and represent at the present time an area of rural out-migration. Cambellford and Stirling are located in a zone of marginal farming in Northumberland and

Hastings Counties in eastern Ontario. These centers are mid-way between Toronto and Ottawa and are isolated by distance from the industrial area to the west.

Group *B* is associated with those towns that have an average index on factor one, a low negative index on factor two, and a high positive index on factor three. Consequently, towns in group *B* can be considered to have moderate growth rates, low aid from grants and subsidies, and low levels of expenditures on recreation and community services. This group contains essentially the same type of town as the initial groups of the previous two methods. For the most part the towns in this group are prosperous rural central places and retirement towns associated with the market gardening and fruit farming zones located in close proximity to the urbanized zone stretching from Toronto to the Niagara Peninsula. Secondary clusterings are found in the cornbelt area near Windsor and a third group along the St. Lawrence Seaway.

The towns of group *C* are associated with large negative scores on factor one, high positive scores on factor two and low negative scores on factor three. This group corresponds closely to the isolates of Ward's algorithm. Towns in this group are fast growing, highly subsidized, and they have rapidly expanding recreation and community services. The town that is most representative of this small group is Streetsville. It is a dormitory town located no more than twenty-five minutes by expressway from either Toronto or Hamilton and less than fifteen minutes from the three cities of Burlinton, Oakville and Mississauga. The towns in this group, with the exception of Bancroft (a town with a fluctuating mining economy), are located in predominantly rural settings but have convenient access to large urban areas.

CONCLUSION

The different groupings identified by the test algorithms established an important point. No matter how objective the algorithm, some subjectivity always exists. Not only is the choice of data, parameters, and method of grouping important, but the investigator should be aware of the consequences of his decisions. One example would be the awareness of the inability of the centroid and Ward's algorithms to identify certain data patterns.[16] Another would be the inability of Semple's algorithm to determine meaningful group unless the data are naturally clustered. Warnings such as these have been pointed out well by Johnston (1968).

University of California, Los Angeles
The Ohio State University, Columbus

BIBLIOGRAPHY

Berry, B. J. L., 1958, A note concerning methods of classification, *Ann. Ass. Amer. Geogr.*, 48, 330–303.

Berry, B. J. L., 1959, Methods and problems of taxonomy (unpublished).

Berry, B. J. L., 1961, A method for deriving multi-factor uniform regions, *Przegl. Geogr.*, 33, 273.

Berry, B. J. L., 1966a, Mathematical methods of economic regionalization, in *Proceed-*

[16] P. M. Lankford (1969).

ings of the Brno Conference on Economic Regionalization, Brno; Czechoslovakian Academy of Sciences.

Berry, B. J. L., 1966b, *Essays on commodity flows and the spatial structure of the Indian economy*, University of Chicago, Dept. of Geography, Research Paper 111.

Berry, B. J. L., 1968, A synthesis of formal and functional regions using a general field theory of spatial behavior, in *Spatial Analysis*, ed. Brian Berry and Duane Marble, Englewood Cliffs, N. Y., Prentice-Hall.

Blumenthal, L. M., 1953, *Theory and application of distance geometry*, London, Oxford Clarendon Press.

Bunge, W., 1962, *Theoretical geography*, Lund, Royal University of Lund, Department of Geography.

Bunge, W. n.d., Regional taxonomy and density transformation (unpublished).

Casetti, E., 1964, Classifications and regional analysis by discriminant iterations, *Technical Report 12*, Office of Naval Research, Evanston.

Casetti, E., and Semple, R. K., 1968, A method for the stepwise separation of spatial trends, Michigan Inter-University Community of Mathematical Geographers, *Discussion Paper 11*.

Dacey, M. F., 1963, Order neighbor statistics for a class of random patterns in multidimensional space, *Ann. Ass. Amer. Geogr.*, 53.

Dacey, M. F., 1964a, Two-dimensional random point patterns: a review and interpretation, *Papers, Reg. Sci, Ass.*, 13, 44–51.

Dacey, M. F., 1964b, Modified poisson probability law for point patterns more regular than random, *Ann. Ass. Amer. Geogr.*, 54, 559–65.

DeJong, G., 1962, Chorological differentiation as the fundamental principle of geography, Groungen, J. B. Walters.

Friedman, H. P., and Rubin, J., 1967, On some invariant criteria for grouping data, *Amer. Statist. Ass. J.*, 62, 1159–1178.

Gilbert, E. W., 1960, The idea of a region, *Geography* 45, 160.

Goodall, D. W., 1954, Objective methods for the classification of vegetation, *Austr. J. Botany*, 2, 41–50.

Gower, J. C., 1966, Some distance properties of latent root and vector methods used in multivariate analysis, *Biometrika* 53, 325–338.

Grigg, D., 1965, The logic of regional systems, *Ann. Ass. Amer. Geogr.* 55, 465–491.

Grigg, D., 1967, Regions, models, and classes, in *Models in geography*, ed. R. J. Chorley and P. Haggettt, London, Methuen.

Hagood, M. J., Danilevsky, N. D., Beum, C. O., 1941, An examination of the use of factor analysis in the problem of sub-regional delineation, *Rur. Sociol.* 6, 216–233.

Hartigan, J. A., 1967, Representation of similarity matrices by trees, *Amer. Statist. Ass J.*, 62, 1140–1158.

Johnston, A., 1968, Choice in classification: the subjectivity of objective methods, *Ann. Ass. Amer. Geogr.*, 58, 575–589.

Kaiser, H. F., 1964, An objective method for establishing legislative districts, *J. Midwest Polit. Sci. Ass.* 20, 75–100.

Kendall, M. G., 1939, The geographical distribution of crop productivity in England, *J. Roy. Statist. Soc.* 102, 21–48.

Lankford, Philip M., 1969, Regionalization: Theory and alternative algoritms, *Geographical Analysis* 1, 196–212.

Mack, C., 1950, The expected number of aggregates in a random distribution of n points, *Proc. Camb. Phil. Soc.*, 46, 285–292.

Mackay, J. R., 1958, Chi square as a tool for regional studies, *Ann. Ass. Amer. Geogr.*, 48, 164–166.

McQuitty, L. L., 1957, Elementary linkage analysis, *Educ. Psychol. Meas.*, 17, 207–229.

Nans, J., 1965, Clustering of random points in two dimensions, *Biometrika* 52, 263–267.

Neely, P., n.d., General discussion re: neighborhood limited classification (unpublished).

Neely, P., n.d., Towards a theory of classification (unpublished).

Regional Science Institute, Regionalization of Pennsylvania counties for development planning (unpublished).

Reynolds, R. B., 1956, Statistical methods in geographical research, *Geogr. Rev.*, 46, 129–132.

Rodoman, B. B., 1967, Mathematical aspects of the formalization of regional geographic characteristics, *Vestn. Mosk. Univ.*, 23, 28–44.

Rosenblatt, D., 1967, Aggregation in matrix models of resource flows, *Amer. Statistician* 21, 32–38.

Rubin, J., 1965, Optimal classification into groups: an approach for solving the taxonomy problem, IBM, Mathematics and Application Department, New York (unpublished).

Russett, B. M., 1965, Delineating international regions, in *Empirical studies in international relations*, ed. J. David Singer, New York, n.p.

Sebastyen, G. S., 1962, *Decision-making processes in pattern recognition*, New York, Macmillian Company.

Semple, R. K., 1966, A quantitative separation and analysis of spatial trends and the viability of small urban centers in Southern Ontario. Unpublished M. A. thesis, University of Toronto.

Semple, R. K., Casetti, E., and King, L. J., 1969, The determination of the optimal number of groupings in classification problems, *Discussion Paper No. 10*, Department of Geography, The Ohio State University, Columbus.

Sokal, R. R., 1961, Distance as a measure of taxonomic similarity, *Systematic Zoology* 10, 70–79.

Sokal, R. R., 1962, Typology and empiricism in taxonomy, *J. Theoret. Biology*, 3, 230–267.

Sokal, R. R., Camin, J., 1965, The two taxonomies: areas of agreement and conflict, *Systematic Zoology*, 14, 176–195.

Sokal, R. R., Sneath, P. H. A., 1963, *Principles of numerical taxonomy*, San Francisco, Freeman Press.

Stone, R., 1966, Comparison of the economic structure of regions based on the concept of distance, *J. Reg. Sci.*, 2, 1–20.

Thompson, J. H., Sufrin, S. C., Gould, D. R., Buck, M. A., 1952, Towards a geography of economic health: the case of New York state, *Ann. Ass. Amer. Geogr.*, 42, 1–20.

Ward, Joe H., Jr., 1963, Hierarchical grouping to optimize an objective function, *Amer. Statist. Ass. J.*, 58, 236–243.

Whittlesey, D., 1954, The regional concept and the regional method, in *American geography: inventory and prospect*, ed. P. James and C. F.

Jones, B., Syracuse, N. Y., Syracuse University Press.

Zobler, L., 1957, Statistical testing of regional boundaries, *Ann. Ass. Amer. Geogr.*, 47, 83–85.

Zobler, L., 1958, Decision making in regional construction, *Ann. Ass. Amer. Geogr.*, 48, 140–148.

GEOGRAPHIA POLONICA 25, 1973

STRUCTURAL CHANGES OF THE ECONOMIC REGIONS IN POLAND: A STUDY BY FACTOR ANALYSIS OF COMMODITY FLOWS

ZBYSZKO CHOJNICKI AND TERESA CZYŻ

INTRODUCTION

Between the elements of spatial economic structure there are various types of linkage. Among these, of particular areal significance, are those revealing the spatial links which occur between various phases of the production process as well as between production and consumption. These are expressed above all in the exchange of all kinds of goods and services. Such exchanges are reflected most strikingly in commodity flows. These flows establish a basic measure of the links, i.e., interregional links binding together the fundamental structural elements of space economy; these elements are the economic regions. That the phenomenon of commodity flows is a measure of inter-regional connections is substantiated by the fact that such flows reveal the magnitude of goods exchanged which, in turn, expresses a geographical division of labour seen in the specialization and complexity of individual economic regions.

The inter-regional exchange is deeply rooted in the chain of basic relations of economic processes. Essentially, it is the inequality within the regions between the level and structure of production and the level and structure of consumption which forms the basis for inter-regional exchange.

The breakthrough in research on inter-regional connections based on commodity flows was achieved by E. Ullman (1957) who worked out for the United States the pattern of commodity flows between states, and presented the characteristics of certain states from an interpretation of flow phenomena. However, it was only later through the efforts of W. Isard (1954, 1961) that the theoretical conclusions resulting from such analyses were applied to the investigation of regional patterns. According to W. Isard, investigations of commodity flows establish the essential contents of inter-regional dependence which are not taken into account in the Lösch's (1940) regional model. Commodity flows also throw light on the existence of regions of different order in a hierarchical arrangement of regional structure.

This type of research was undertaken in Poland by Z. Chojnicki (1961, 1964) and W. Morawski (1968 a, b).

Z. Chojnicki determined the degree of integration and differentiation of the nation's spatial structure based on the rail traffic flows between the voivodships for 1958. This study revealed that Poland is one region, its economic centre being Upper Silesia. Only within this primary transport region some additional subareas can be distinguished. Within the core area of industrial production conceived on a national scale there are — outside the Upper Silesian

conurbation — two subcentres: Wrocław strongly related to the north-western part of the whole country and Cracow related to the south-eastern part. Moreover, there are several subregions characterized by more intensive exchange of some products within them than with other areas. To these belong the north-eastern part of the country with Warsaw at its main economic centre and the west-northern part with Poznań and the main seaports. The importance of this is however reduced because the author, having limited statistical data at his disposal, discusses these inter-regional flows only in terms of tonnage and not of monetary value.

W. Morawski continued the research of inter-regional flows using value data for 1962. The results confirm that the whole regional system of Poland exhibits a conspicuous orientation towards the region of Upper Silesia.

A somewhat different approach, but perhaps the most promising for the structure of flow patterns, was adopted by B. J. L. Berry (1966, 1967, 1968). His methods is based on the extraction of redundancies in the $m \times m$ correlation of commodity flows using factor analysis. In *R*-mode analysis, the (column) correlation matrix is factored, yielding groups of destinations (factor loadings) similar in terms of the manner in which their needs are assembled. The factor scores identify those origins important in shipping to each group. *Q*-mode analysis results in essentially the same information for origins. Berry's analysis of Indian commodity flows between 36 trade blocks follow this methodology.

The concept of the flow matrix is further employed by B. J. L. Berry (1966, 1968) in his general field theory of spatial structure and spatial behaviour. This theory considers a system that consists of places, attributes of places, and interactions between places, all seen through time. Factoring the $n \times a$ attribute matrix yields a structural dimension, and an $n \times s$ structure matrix can be created. Similarly, various forms of interaction, including commodity flows of different kinds, can be used to build an $(n^2-n) \times y$ interaction matrix, where (n^2-n) dyads are treated as individual observations. This matrix can be reduced to an $(n^2-n) \times b$ behaviour matrix, again by factor analysis. Cannonical correlation analysis provides the means of observing the similarity between places and groups of places in terms of their scores on the structural and behavioral dimensions.

THE SCOPE OF THE STUDY

This study will analyse the structural changes of economic regions in Poland based on railroad commodity flows during the period 1958–1966.

Railroad transport in Poland plays a major role in the inter-regional exchange of goods. In Poland the railways share the largest part of the total freight tonnage moved (82,1%) and of all transportation movements (95,3%). This justifies to a high degree the representative character of railway transport as an indicator of commodity flows.

Data from the official state statistics of commodity flows by railways between 17 voivodships in 1958 and 1966 served as the starting-point. These data are published in the form of matrices, the volume of the flows being recorded in physical units of measurement, i.e., in tons. The matrices contain commodity flows for the following 17 freight groups:

(1) bituminous coal,
(2) brown coal and coke,
(3) ores and pyrites,
(4) stones,

(5) sands and gravels,
(6) crude and refined petroleum,
(7) metals and metal manufactures,
(8) bricks,
(9) cement,
(10) artificial fertilizers,
(11) chemical products,
(12) grains,
(13) potatoes,
(14) sugar beets,
(15) other crops and processed agricultural products,
(16) timber and timber manufactures,
(17) other freight.

However, there are obvious limitations to the scope of the conclusions and estimates resulting from the regional implications of the physical volume of commodity flow. Thus, those data on the physical volumes of the flows have been processed so as to achieve their (estimated) value size. This processing has been completed on the basis of a value index of the particular 17 groups of commodities, which was estimated by W. Morawski (1967). These indices are presented in Table 1.

TABLE 1. Index of value of one ton of commodities dispatched by railways based on the 1962 structure of production and dispatches

Group number	Categories of commodities	Value of one ton in zl (in factory prices)
(1)	Bituminous coal	350
(2)	Brown coal and coke	555
(3)	Ores and pyrites	450
(4)	Stones	95
(5)	Sands and gravels	45
(6)	Crude and refined petroleum	1985
(7)	Metals and metal manufactures	4580
(8)	Bricks	235
(9)	Cement	450
(10)	Artificial fertilizers	1060
(11)	Other chemical products	5310
(12)	Grains	3200
(13)	Potatoes	837
(14)	Sugar beets	505
(15)	Other crops and processed agricultural products	3800
(16)	Timber and timber manufactures	2040
(17)	Other freight	7540

The value of commodity flows based on the statistics of railway freight haulage, from the point of view of their application to regional analysis, is limited with respect to the following:

(1) The 17 voivodships as the consigning-receiving units provide too little spatial detail and permit an analysis of commodity flows only on a macro-regional scale. This limits analysis to higher order regions only.

(2) There is insufficient differentiation in generic grouping of freight in the 16 classified groups. From the economic point of view these do not have an homogeneous character and this makes impossible any differentiation in the individual types of raw materials and finished products. This also applies to any introduction of economic accounting in terms of monetary value.

(3) Other limitations result from the existence of crosshauls, extenuated hauls and back-hauls which do not represent true economic links.

Despite this, however, a comparison of railway freight flows on the inter-regional scale does show the existence of basic regional contrasts which, from the point of view of regional analysis, possess fundamental significance: they permit one to grasp the chief inequalities in the distribution of the output of raw materials and mass products, and they reflect the major elements of the geographical division of labour.

The definition of Poland's regional structure on the basis of the statistical material characterized above is limited to the existing voivodship framework. There is no possibility of achieving a correction of this division and as a result, one can only approximate reality.

Recognition of this fact limits the investigation of regional structure to the voivodship as the basic element, therefore establishing the administrative-economic units as the economic regions. It must be emphasized that the degree to which such an analysis is adequate is closely defined by the suitability of this initial system; only to that extent can one accept this analysis of the regional economic structure of the country.

Analysing the structure of the system of economic regions in this form is an exercise in definition based on flows, types of commodities of the economic regions, as well as on the links occurring between them. Investigation of the system's structure depends on the elaboration of the kind of relationships arising between the system's elements. The complex of these relationships can be named according to the nature of the connecting elements. This establishes a substitute for research on the regional structure because it permits the recognition of the whole feature of these structural elements as well as the existing relations between them. This emerges only from the investigation of regional peculiarities, and results from the individual features which distinguish one region from other regions.

Referring the investigation of regional structure to that of the spatial regional structure as given, the analysis can proceed to the first important problem, that of the complexity of the system of economic regions regarding their character as elements of that system, and the links between them.

THE ANALYSIS

The analysis of regional structure of Poland in this paper is based on the application of two methods:

(1) principal factor method introduced by H. Hotelling (1933),

(2) grouping algorithm presented by J. D. Nystuen and M. F. Dacey (1961).

The mathematical procedure starts from an interaction matrix of the order 272×17, in which the $(17^2 - 17)$ possible pairs of voivodship-regions (dyads) occupy the rows and 17 kinds of interaction (commodities) occupy the columns. Dyads are treated as individual observations. The types of commodity become the variables in this analysis.

This matrix is transformed into a matrix of standardized data, also of type 272×17, which consists of the values of the particular standardized variables expressed in units of standard deviation.

Normalization is completed on the basis of the formula:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \qquad \begin{matrix} i = 1, 2, \ldots N, \\ j = 1, 2, \ldots n, \end{matrix} \tag{1}$$

where:

x_{ij} = value of variable j of dyad i,

$\bar{x}_j$ = mean of N values of variable j (N denotes the number of dyads),

s_j = standard deviation of variable j.

The relationships between variables are expressed by help of the coefficient of correlation:

$$r_{jk} = \frac{\sum_{i=1}^{N} z_{ij} z_{ik}}{N}. \tag{2}$$

The correlation matrix of order n is a symmetrical matrix.

Multiple factor analysis extracts the factor (hypothetical variables), which constitutes the basis of correlations observed in a given set variables (x_1, x_2,... x_n). These factors may be treated as causes of the variation observed; it is then possible to interpret them as being of considerable importance in the measurement and explanation of variation. Factor analysis helps to reduce a primary set of variables that are characteristic of the objects under observation to a considerably smaller number of factors. In this manner, the number of dimensions of the objects diminishes and analysis becomes simpler.

In factor analysis n observed variables characterizing a set of N dyads is linear function of m unknown "common factors" (F_1, F_2, ... F_n), where $m < n$ and a "unique factor" for each of the variables (U_1, U_2,... U_n):

$$z_j = a_{j1} F_1 + a_{j2} F_2 + \ldots a_{jm} F_m + a_j U_j, \tag{3}$$

where a' s are called factor loadings.

If we assume that both the observed variables and the factors are at standard form (i.e. with the mean equal to zero and the variance equal to unity) and if we further assume that the factors are uncorrelated, then the variance of the observed variables, z_j, can be computed from

$$s_{zj}^2 = 1 = a_{j1}^2 + a_{j2}^2 + \ldots + a_{jm}^2 + a_j^2 = h_j^2 + a_j^2; \tag{4}$$

h_j^2 is called the communality and it is that part of the variance of the observed variable, which is due to the common factors, while a_j^2 the uniqueness is that part of the variance, which is due to the unique factor.

Factor analysis, as D. N. Lawley and A. E. Maxwell (1963) emphasize, usually implies some hypothesis as to the number of common factors underlying the set of variables in the research problem.

Factor analysis, which consists in examing the communality of features resulting from the operation of common factors, is carried out on reduced correlation matrix in the form:

$$R = \begin{bmatrix} h_1^2 & r_{12} & \cdots & r_{1n} \\ r_{21} & h_2^2 & \cdots & r_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ r_{n1} & r_{n2} & \cdots & h_n^2 \end{bmatrix} \tag{5}$$

h_j^2 denotes the communality of variable j and is approximated from the formula:

$$h_j^2 = \frac{r_{jk} r_{jl}}{r_{kl}} \qquad j = 1, 2, \ldots n, \tag{6}$$

where r_{jk} and r_{jl} are maxima coefficients of correlation of variable j.

The basic problem of factor analysis is to determine the coefficients $a_{j1}, \ldots a_{jm}$ of the common factors. This determination can be made by principal factor method.

The principal factor method makes possible the extraction of factors, which explain the maximum communality and give the smallest possible residuals in the correlation matrix. This means, that the sum of squares of the factor loadings is the largest possible for each variable.

The analysis begins with a factor F_1 whose contribution to the communality of the variables has as great a total as possible. Then the first — factor residual correlation is obtained, including the residual communalities. A second factor F_2, independent of F_1, with a maximum contribution to the residual communality is next found. This process is continued until the total communality is analysed.

If the composition of a statistical variable is taken to be

$$z_j = a_{j1} F_1 + a_{j2} F_2 + \ldots + a_{jm} F_m \qquad j = 1, 2, \ldots n, \tag{7}$$

with the unique factor omitted, the communality of z_j is then given by:

$$h_j^2 = a_{j1}^2 + a_{j2}^2 + \ldots a_{jm}^2. \tag{8}$$

The sum of the contribution of factor F_1 to the communalities of the n variables is

$$A_1 = a_{11}^2 + a_{21}^2 + \ldots + a_{n1}^2. \tag{9}$$

The solution of the problem consists in finding such values of the coefficients a_{j1} for which A_1, assumes the maximum value, the following condition being fulfilled:

$$r_{jk} = r'_{jk} = \sum_{t=1}^{m} a_{jt} a_{kt} \qquad j, k = 1, 2, \ldots m. \tag{10}$$

We have here a problem involving the maximization of A_1, a function of several variables which in turn are connected by a set of relationships. The mathematical procedure as outlined in H. H. Harman (1960) involves the use of Lagrangian multipliers to obtain a set of n equations of the form

$$\begin{bmatrix} h_1^2 - \lambda & r_{12} & \cdots & r_{1n} \\ r_{21} & h_2^2 - \lambda & \cdots & r_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ r_{n1} & r_{n2} & \cdots & h_n^2 - \lambda \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{21} \\ \cdot \\ \cdot \\ \cdot \\ a_{n1} \end{bmatrix} = 0 \tag{11}$$

These equations constitute the bass for the calculation of the unknown coefficients a_{j1}.

A necessary condition for the solution of this set of equations is that the determinant of the coefficients a_{j1} must be equal to 0.

$$\begin{vmatrix} h_1^2-\lambda & r_{12} & \cdots & r_{1n} \\ r_{21} & h_2^2-\lambda & \cdots & r_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ r_{n1} & r_{n2} & \cdots & h_n^2-\lambda \end{vmatrix} = 0 \tag{12}$$

This is a characteristic equation, in which all roots are real.

Corresponding to the first root or eigenvalue of this equation is a column vector or eigenvector (a_{11}, a_{21}, ... a_{n1}), which when scaled by the factor yields the coefficients a_{11}, a_{21}, ... a_{n1}.

$$\left(\frac{\lambda_1}{a_{11}^2+a_{21}^2+a_{n1}^2}\right)^{1/2} \tag{13}$$

The residual correlation matrix $[R']$ can then be computed as and the solution could proceed with finding the largest eigenvalue of this residual matrix, and so on.

$$[R'] = [R]-[a_{j1}][a_{j1}]^T \tag{14}$$

H. Hotelling introduced a simplified method of calculating factor loadings in solving the main factor. He used an approximate determination of the characteristic roots by the iteration process method without the previous unfolding of the characteristic determinant (H. H. Harman, 1960).

In this paper H. Hotelling's iterative method is used. The solution was based on a programme in Gier Algol IV language using the Gier computer.

The computer-derived solution in our example yields the following eigenvalues:

for 1958 $\lambda_1 = 7,9695$, $\lambda_2 = 2,8342$,

for 1966 $\lambda_1 = 5,2469$, $\lambda_2 = 3,1879$.

Each eigenvalue accounts for a percentage of the total common variance.

The question of how many factors should be interpreted is difficult. A convenient rule of thumb seems to be to evaluate all factors with an eigenvalue equal to or greater than one or, alternately to evaluate each one which accounts for a sufficiently high proportion of this communality.

In this example, factor analysis carried out by the principal factor method yields the factorial matrices of type 17×2 for 1958 and 1966, which contain the loadings of two factors in 17 variables (Table 2 and 3). Two factors accounted for 95% of a total common variance in 1958 and 75% in 1966.

The interpretation of the factors is usually important in a research problem. This interpretation is done mainly with reference to the factor loadings, which have the form of a coefficient of correlation between the variable and a given factor.

On any factor some variables will have low loadings and consequently will be ignored in the process of giving an interpretation to the factor.

We assume, that the regional structure is a linear function of some simple patterns and the factors in the linear model should illustrate the simple structure.

In 1958 an underlying two-factor structure was revealed. Factor I, accounting for 70.32% of common variance, consist of three groups: (1) raw materials of mineral origin (bituminous coal, brown coal and coke, ores, stones,

TABLE 2. Factor structure
Dyadic analysis of 17 commodities in Poland, 1958

Group number	Categories of commodities	Factor loadings	
		I	II
(1)	Bituminous coal	0.6958	−0.4241
(2)	Brown coal and coke	0.8649	−0.3570
(3)	Ores	0.8221	−0.3998
(4)	Stones	0.5925	0.0087
(5)	Sands and gravels	0.9033	−0.1220
(6)	Crude and refined petroleum	0.4266	0.0561
(7)	Metals and metal manufactures	0.7814	−0.3945
(8)	Bricks	0.7623	0.3239
(9)	Cement	0.5963	−0.2578
(10)	Artificial fertilizers	0.4901	0.1274
(11)	Other chemical products	0.8900	−0.3023
(12)	Grains	0.4901	0.1274
(13)	Potatoes	0.2709	0.5307
(14)	Sugar beets	0.4304	0.7131
(15)	Other crops and processed agricultural products	0.3144	0.4502
(16)	Timber and timber manufactures	0.7683	0.4523
(17)	Other freight	0.9477	0.0455
	λ	7.9695	2.8342
Per cent of common variance explained by the factor		70.32	25.01

sands and graves), (2) industrial goods (metals and metal manufactures, bricks, cement, artificial fertilizers, other freight), (3) timber and timber manufactures. Accounting for 25% of communality, Factor II represent agricultural products. Strong loadings are recorded by the commodities: grains, potatoes, sugar beets.

In 1966 situation changed very much. The identification of factors is not so clear. Factor I explains only 46% of the total common variance of the variables and comprises mainly industrial products and ores (ores, metals and metal manufactures, other chemical products, other freight), agricultural products (grains, sugar beets, other crops and processed agricultural products), timber and timber manufactures. Factor II is based primarily on the loadings by the raw materials for fuel and building (brown coal and coke, stones, bricks). This factor explains about 28 per cent of the communality of features.

Then the factor scores for dyads were evaluated according to the equation

$$[F] = [Z][A], \tag{15}$$

where

$[F]$ = matrix of factor score,
$[Z]$ = an observation matrix,
$[A]$ = matrix of factor loadings.

This factor scores matrix of type 272×2 was transformed into two matrices for every year (1958 and 1966) of order 17, being a starting-point for the spatial

TABLE 3. Factor structure
Dyadic analysis of commodities in Poland, 1966

Group Number	Categories of commodities	Factor loadings	
		I	II
(1)	Bituminous coal	−0.0003	0.1868
(2)	Brown coal and coke	0.4951	0.7492
(3)	Ores	0.6580	0.6498
(4)	Stones	0.5530	0.7135
(5)	Sands and gravels	0.4488	0.0563
(6)	Crude and refined petroleum	0.1928	−0.1293
(7)	Metals and metal manufactures	0.5610	−0.2515
(8)	Bricks	0.4700	0.7355
(9)	Cement	0.3290	−0.0017
(10)	Artificial fertilizers	0.3993	0.0961
(11)	Other chemical products	0.5629	−0.3587
(12)	Grains	0.8197	0.0882
(13)	Potatoes	0.3303	0.0704
(14)	Sugar beets	0.6983	−0.4895
(15)	Other crops and processed agricultural products	0.7201	−0.5430
(16)	Timber and timber manufactures	0.7392	−0.2677
(17)	Other freight	0.7547	−0.5283
λ		5.2469	3.1879
Per cent of common variance explained by the factor		46.07	27.99

grouping, which we can call "latent structure matrix" or using the term of B. J. L. Berry "the behaviour matrix".

Each cell of the matrix corresponds to a different element of interregional exchange, i.e., to a different inter-regional connection. The cells on the main diagonal referring to connection within each of the particular regions were omitted.

In the rows of the matrix for every factor we read outflows in the term of factor score from the particular regions i.e. their active connections, whereas in the colums we read the inflows, i.e., the passive connections (Tables 4—7). *

FACTOR INTERPRETATION

The second step of our analysis is associated with the problem of generalizing two basic factors into a system of regional structure, changing in time. This analysis requires the grouping together of voivodships on the basis flows in the term of dyad factor scores.

As the method of grouping dyads for each factor we used the method described by J. D. Nystuen and M. F. Dacey (1961), applied originally to telephone traffic in Washington. The application of basic theorems of graph theory interpretation by J. D. Nystuen and M. F. Dacey, permits hierarchical relations between voivodships to be established in two aspects: outflows (active connections) and inflows (passive connections). If the connections in terms of factor

* Tables 4—7 at the end of the volume

scores are ranked according to their magnitudes in the rows and columns, it is possible to determine the dominant and subordinate voivodships. The dominant voivodship is one which records its largest flow to a lower order voivodship. The subordinate voivodship is one for which the largest flow is to a higher order voivodship (Fig. 1).

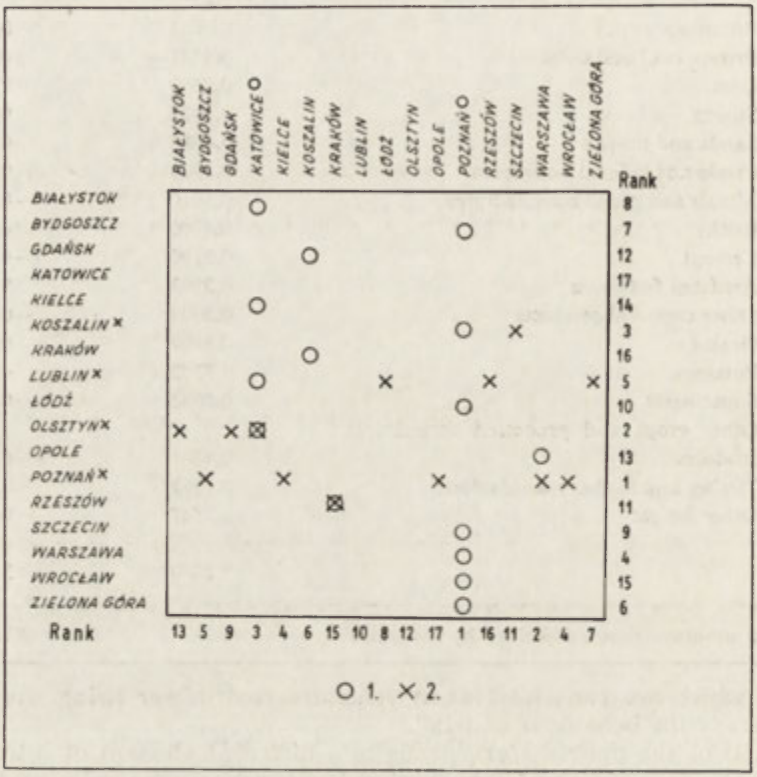


Fig. 1. Adjacenty matrix of graph F_2 (1958)
1 — largest outflow; 2 — largest inflow

The resulting hierarchy structure describing the regional pattern for each factor in both years is presented on 8 graphs for passive and active connections (Figs. 2–9).

The structure established by isolating the largest flows in the same manner as was described on graphs permit maps to be drawn of regional structure.

The pattern of connection presented on maps establishes a synthetic description of the complexities of the country's regional structure. That complexity is expressed in the differentiation of various forces integrating the inter-regional links.

The main descriptive conclusions concerning regional structure, can be drawn from a comparative analysis of changes in time of factor one, which identified the mining and manufacturing industry. First of all the whole regional system of country exhibits the most intensive connections with Katowice. The connections with Katowice occupy first place in the inter-regional flows of all other regions, endowing Katowice with a focal character on the national scale. This defines the role of Katowice (The Upper Silesian Industrial District) as that area upon which are focussed the productive-industrial activities of the country, the basic sections of heavy industry: coal-mining, metallurgy, engineering and

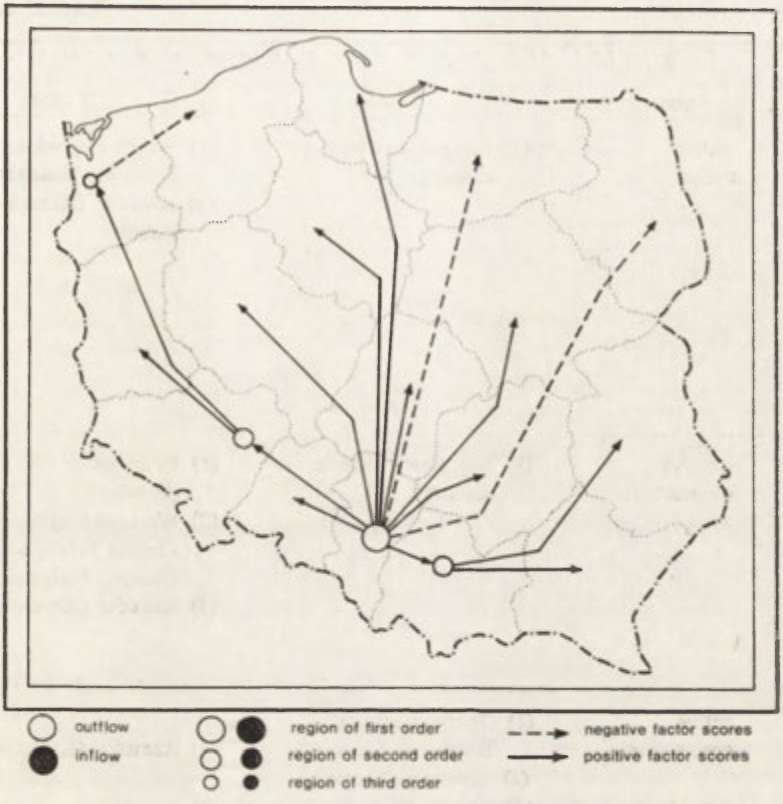


Fig. 2. Factor I. Interregional active connections, 1958

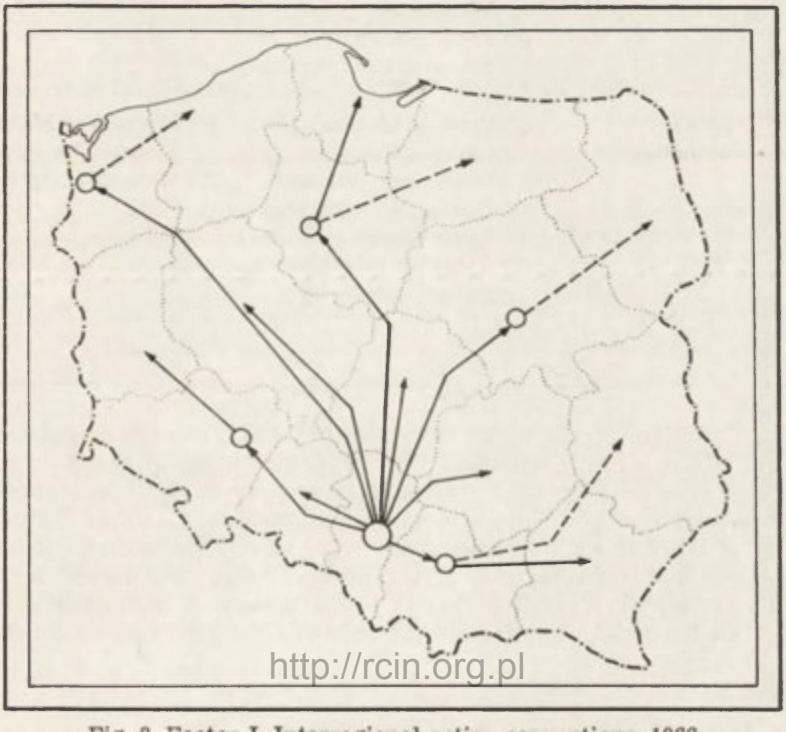


TABLE 8. Regional

Factor	Kind of connections	1958	
		I order	II order
	active connections	(1) Katowice (whole country)	(1) Wrocław (Zielona Góra, Szczecin, Koszalin) (2) Kraków (Rzeszów, Lublin)
Factor I	passive connections	(1) Katowice (whole country)	(1) Bydgoszcz (Gdańsk) (2) Warszawa (Poznań, Zielona Góra, Szczecin, Olsztyn, Białystok, Lublin) (3) Kraków (Rzeszów)
	active connections	(1) Olsztyn (Gdańsk, Białystok, Katowice) (2) Koszalin (Szczecin) (3) Poznań (Wrocław, Opole Kielce, Bydgoszcz, Warszawa) (4) Lublin (Zielona Góra, Rzeszów, Łódź, Kraków)	(1) Rzeszów (Kraków)
Factor II	passive connections	(1) Poznań (Zielona Góra, Szczecin, Koszalin, Bydgoszcz, Warszawa, Gdańsk, Wrocław, Opole, Łódź, Kraków, Rzeszów) (2) Katowice (Lublin, Kielce, Olsztyn, Białystok)	(1) Koszalin (Gdańsk, Kraków, Rzeszów) (2) Warszawa (Opole)

chemicals. The high degree of its specialization links it with a wide area, and as a result, gives a unity which is the functional basis of its ability for full complex economic development; thus simultaneously it also establishes its own inner coherence. The high intensity of the commodity flows of Katowice, the uniformity of links, the active and passive type of dependence and its character as an open economic region reflect the predominant role played by the raw materials and industry of this region in the structure of the national economy. As a result of its nodal organization, therefore, Katowice can be considered as

structure of Poland

1966			
III order	I order	II order	III order
(1) Szczecin (Koszalin)	(1) Katowice (whole country)	(1) Wrocław (Zielona Góra) (2) Szczecin (Koszalin) (3) Bydgoszcz (Gdańsk, Olsztyn) (4) Warszawa (Białystok) (5) Kraków (Lublin, Rzeszów)	
(1) Poznań (Zielona Góra)	(1) Katowice (Opole, Wrocław, Zielona Góra, Bydgoszcz, Gdańsk, Warszawa, Białystok, Kielce, Kraków, Rzeszów, Lublin) (2) Poznań (Koszalin, Łódź)	(1) Wrocław (Zielona Góra) (2) Bydgoszcz (Gdańsk) (3) Warszawa (Białystok) (4) Kraków (Rzeszów, Lublin)	(1) Rzeszów (Lublin)
	(1) Katowice (whole country)	(1) Wrocław (Zielona Góra)	
(1) Kraków (Rzeszów)	(1) Wrocław (Warszawa, Zielona Góra, Katowice, Poznań, Szczecin, Rzeszów) (2) Kraków (Koszalin, Białystok) (3) Łódź (Lublin, Bydgoszcz) (4) Kielce (Gdańsk, Opole)	(1) Rzeszów (Poznań) (2) Warszawa (Szczecin) (3) Koszalin (Białystok) (4) Lublin (Bydgoszcz) (5) Opole (Olsztyn)	

the focal economic region in the national system with no changes in active connections in time. (Table 8).

Second order pattern is different for active and passive connections. The active connections constitute two regions: Wrocław and Kraków voivodship, the passive connections — three: Bydgoszcz, Warszawa, Kraków voivodship. The changes in time in the second order patterns show the further differentiation and origin of new regional centres: active — Szczecin, Bydgoszcz, Warszawa voivodship; passive — Wrocław voivodship.

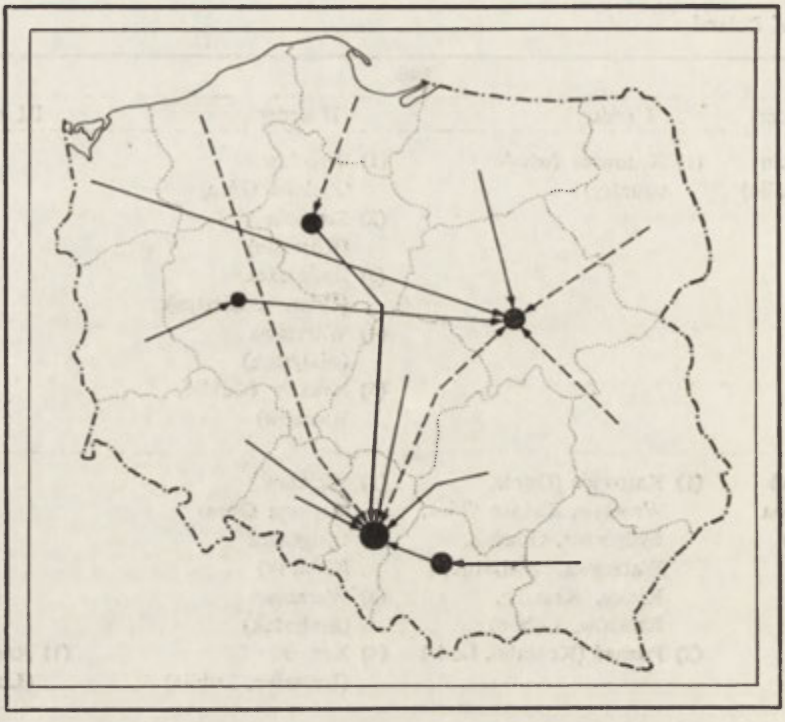


Fig. 4. Factor I. Interregional passive connections, 1958

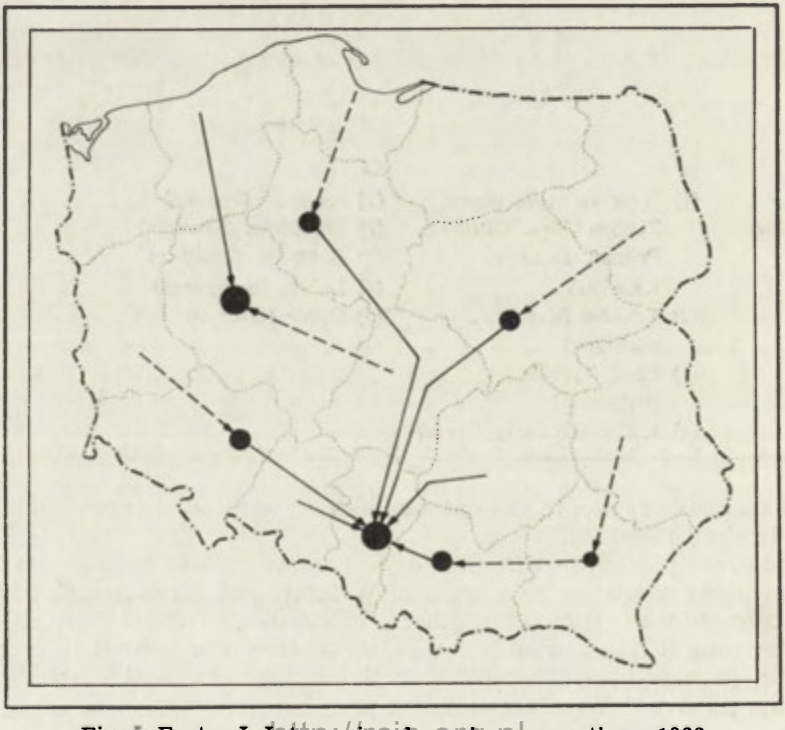


Fig. 5. Factor I. Interregional passive connections, 1966

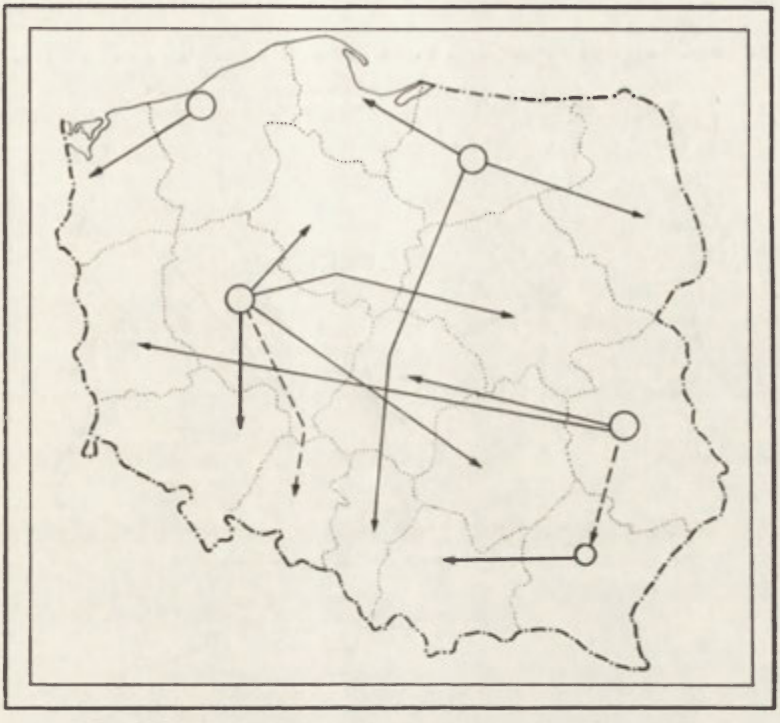


Fig. 6. Factor II. Interregional active connections, 1958

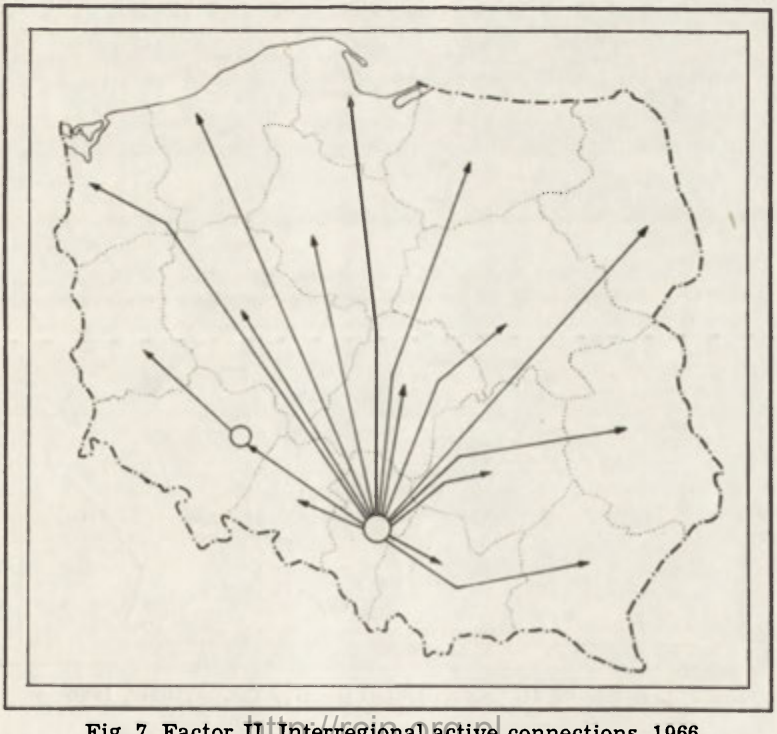


Fig. 7. Factor II. Interregional active connections, 1966

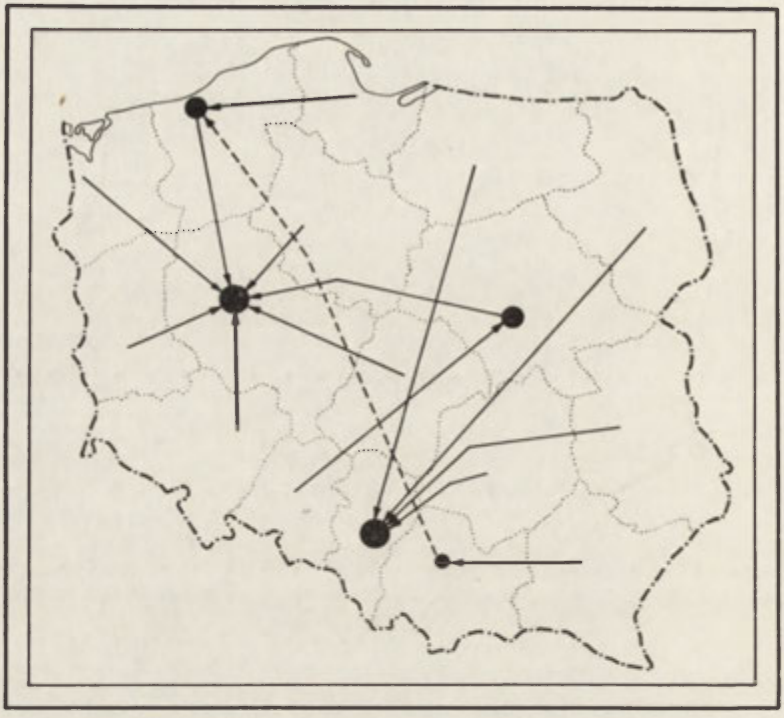


Fig. 8. Factor II. Interregional passive connections, 1958

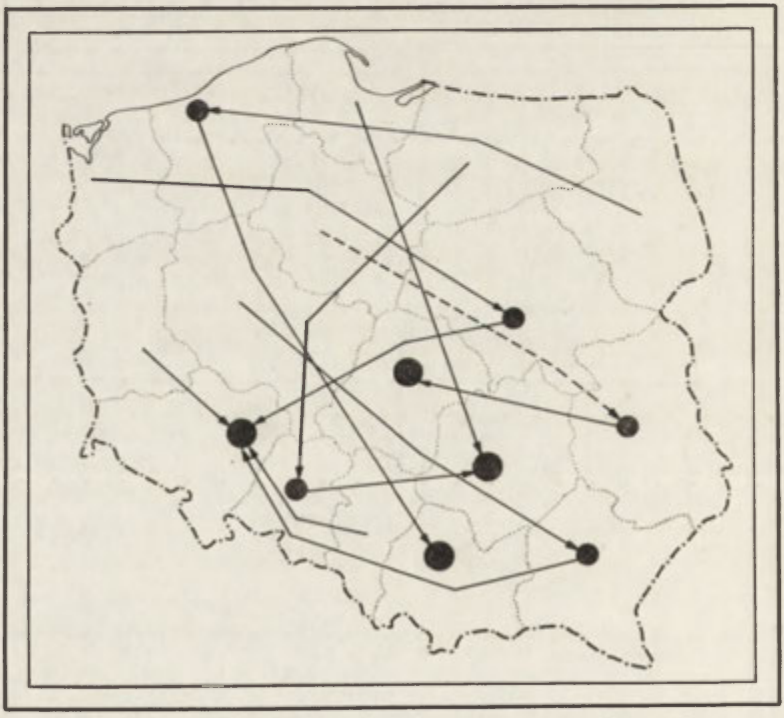


Fig. 9. Factor II. Interregional passive connections, 1966

The system is much more complicated and some subordinated regions are not continuous to its superior regions. This is probably partly attributable however to the some changes in the nature of the factor including also agricultural flows.

Factor two in 1958 picks out mainly agricultural patterns. These relations permit one to find certain elements for division into structure of more uniform regional organization. The nature of the second factor is not the same in 1966. This is why we can not compare the resulting structure in time. In 1966 second factor identifies the raw materials for fuel and building.

In the analysis of commodity flows for the purpose of organization of regions into a hierarchy one must emphasize that the different types of connections give varied organization, which is insufficiently integrated to establish the clear functional regional system.

Adam Mickiewicz University, Poznań

BIBLIOGRAPHY

Berry, B.J.L., 1966, *Essays on commodity flows and the spatial structure of the Indian economy*, University of Chicago, Dept. of Geography, Research Paper 111.

Berry, B. J. L., 1967, The mathematics of economic regionalization, in *Proceedings of the 4th General Meeting of the Commission on Methods of Economic Regionalization of the International Geographical Union*, September 7–12, 1965 in Brno, Prague, 77–106.

Berry, B. J. L., 1968, Interdependence of spatial structure and spatial behavior: A general field theory formulation, *Papers, Reg. Sci, Ass.*, 21, 205–227.

Chojnicki, Z., 1961, *Analiza przepływów towarowych w Polsce w układzie międzywojewódzkim* (The analysis of commodity flows in Poland in an intervoivodship pattern), Studia KPZK PAN, 1, Warszawa.

Chojnicki, Z., 1964, The structure of economic regions in Poland analysed by commodity flows, *Geogr. Pol.*, 1, 213–230.

Harman, H. H., 1960 *Modern factor analysis*, The University of Chicago, Chicago.

Hotelling, H., 1933, Analysis of a complex of statistical variables into principal components, *J. Educ. Psychol.*, 24, 417–441, 498–520.

Isard, W., Freutel G., 1954, Regional and national product projections and their interrelations, in *Studies in income and wealth*, 16, 427–471.

Isard, W., 1961, *Methods of regional analysis: An introduction to regional science*, New York, 132–181.

Lawley, D. N., Maxwell, A. E., 1963, *Factor analysis as a statistical method*, London.

Lösch, A., 1940, *Die räumliche Ordnung der Wirtschaft*, Jena, 64–142.

Morawski, W., 1967, Studium wartości jednej tony towarów przemieszczanych transportem kolejowym i problem integracji klasyfikacji (A study of the value of one ton of commodities transported by railways and the problem of integration of classification) (typescript).

Morawski, W., 1968 a, *Przepływy towarowe i powiązania międzyregionalne na obszarze Polski* (Sum.: Commodity flows and interregional connections in Poland. A value approach), Studia KPZK PAN, 25, Warszawa.

Morawski, W., 1968 b, Balances of interregional commodity flows in Poland: A value approach, *Papers, Reg. Sci. Ass.*, 20, 29–41.

Nystuen, J. D., Dacey, M. F., 1961, A graph theory interpretation of nodal regions, *Papers, Reg. Sci. Ass.*, 7, 29–42.

Ullman, E., 1957, *American commodity flow*, Seattle.

CANONICAL CORRELATION IN GEOGRAPHICAL ANALYSIS

D. Michael Ray and Paul R. Lohnes

NEW INTERPRETIVE DEVICES IN CANONICAL CORRELATION ANALYSIS

The potential contribution to geographic research of canonical correlation analysis as a powerful multivariate tool to investigate spatial interrelationships between two data sets has been demonstrated in a number of recent studies (L. J. King, 1969, pp. 217–222 and P. R. Gould, 1969, pp. 13–14). Berry's work on Indian commodity flows suggested a synthesis of formal and functional regions using a general field theory of spatial behavior comprising places, the attributes of those places and the interactions among them (B. J. L. Berry, 1966, and Brian J. L. Berry, 1968, pp. 419–428). Gauthier's research in the São Paulo region employed canonical correlation analysis to investigate the nature of the interrelationships between nodal accessibility and urban growth and identified the "lead" effects of transportation (H. L. Gauthier, 1968, pp. 77–94). Ray's analysis of Canadian census data revealed a hierarchy of heartland-hinterland relationships between economic and cultural characteristics that are related to centripetal and centrifugal spatial forces (D. M. Ray, 1971).

It is the purpose of this paper to describe and illustrate a number of interpretive devices which have recently emerged and which have not before been utilized in the geographic literature. These devices are the computation of the *canonical factor structure* matrix, the *variances extracted* from each measurement domain by the canonical factors, the *redundancy* of the *canonical factors* of one set given those of the other, and the *canonical factor scores*. The canonical factor structure matrix provides the correlations of the variables (or measurement domain) with the canonical factors and takes the place of the raw canonical vectors in which the variances are uncontrolled. The variances extracted from a measurement domain by a canonical factor may shrink to insignificance if its canonical correlation with the corresponding factor for the other measurement domain is low. A better measure of the interrelationships between the two measurement domains being analyzed is the redundancy measure, which is the product of the variance extracted and the variance shared for each pair of canonical factors. This paper also introduces the notion of canonical factor scores which correspond to the scores computed in principal components analysis and which provide a mapping of the observation units into the canonical factor space. The computation of these indexes is described in the mathematical section which follows. Two research examples are then provided to illustrate the application and interpretation of the technique.

THE MATHEMATICS

Canonical correlation analysis provides a method for exposing the structure of relationships between p_1 measurements in one vector z_1 and p_2 measurements in a second vector z_2 when both vectors of measurements have been taken on one population. The p_1 measurements in vector z_1 in the second research example presented in this paper are eleven labor force characteristics, and the p_2 measurements in vector z_2 are eight cultural characteristics, where the vectors z_1 and z_2 are any given census county in the two measurement domains. Hotelling introduced the technique to determine the substitutability of vector z_1 for vector z_2 in the measurement of an observation (H. Hotelling, 1935, pp. 139–142). By contrast, it may be assumed that the two measurement domains will be clearly distinguishable in geographic research and that the interest will focus on the interrelationships exposed. The example mentioned thus compares an acquired set of characteristics (labor force characteristics) with inherited characteristics (mother tongue). The canonical analysis is needed because neither the zero-order cross-correlations, of which there are $p_1 \times p_2$, nor the full set of multiple correlations of each measurement from each vector with all the measurements of the other vector, of which there are $p_1 + p_2$, provide a suitably parsimonious exposition of the structure of relationships, especially if p_1 and p_2 are sizeable numbers. Moreover, neither of these alternatives resolves the implications of the correlation structures *within* each vector for the understanding of the cross-correlation structure *between* vectors. The most competitive modeling procedure is to perform orthogonal factor analyses on the two measurement domains separately, and then display the zero-order or multiple correlations relating factors of the two domains. This alternative emphasizes the internal structure within each vector and underemphasizes the cross-correlation structure between the vectors, whereas the canonical analysis emphasizes the cross-correlations while also considering the internal correlations.

Although modern computing algorithms permit us to arrive at all the canonical correlations simultaneously, it is helpful to think of the analysis as proceeding in steps, and indeed it may be so computed. From this viewpoint, the first task is to locate one linear component of each measurement vector in a fashion that maximizes the correlation of the two components. Call these first canonical components x_1 and y_1 where:

$$x_1 = c_1' z_1 \quad \text{and} \quad y_1 = d_1' z_2.$$

The problem is to choose the coefficients vectors c_1 and d_1 so that the first canonical correlation, R_1, is maximized.

$$R_1 = \frac{1}{N} \sum_{i=1}^{N} x_1 y_1 \,|\, \max.$$

Note that all variables and components are assumed to be standardized to zero means and unit standard deviations throughout this discussion. Thus, in x_1 and y_1 we have selected the maximally correlated factors of the two measurement vectors z_1 and z_2.

The next step is to locate a second linear component of each measurement vector such that the correlation of these components is maximized under the restriction that the correlations of these second canonical components with the first canonical components must be zero. The second canonical components are

$$x_2 = c_2' z_1 \quad \text{and} \quad y_2 = d_2' z_2.$$

The second canonical correlation is

$$R_2 = \frac{1}{N} \sum_{i=1}^{N} x_i y_i \Big| \max.$$

The restrictions are

$$r_{x_1 x_2} = 0, \quad r_{x_1 v_2} = 0, \quad r_{v_1 v_2} = 0, \quad r_{x_2 v_1} = 0.$$

In further steps it is possible to locate additional pairs of canonical components such that at each step the new components are maximally correlated subject to the restriction that all their correlations with preceding components are zero. Assuming $p_2 \leqslant p_1$, there are p_2 canonical correlations available, but of course they descend in size as the steps progress and some of them may be trivial. It usually becomes a matter of scientific judgment as to how many pairs of canonical factors to include in a model for data. Let n be this number, which may be called the *rank* of the model for the data, where $n \leqslant p_2 \leqslant p_1$. We array the n column vectors of coefficients of z_1 in the matrix C, which then has p_1 rows and n columns. We array the n column vectors of coefficients for the n canonical components (or factors) of z_2 in the matrix D, which then has p_2 rows and n columns. If we let R_{11} stand for the square, symmetric matrix of intercorrelations among the p_1 measurements in vector z_1, the *factor structure* matrix containing the correlations of the measurements in z_1 with their canonical components is

$$S_1 = R_{11} C$$

in which the elements s_{jk} is the correlation of the jth measurement with the kth canonical component.

The structure coefficients giving the correlations of the measurements in z_2 with their canonical components are given by

$$S_2 = R_{22} D$$

where R_{22} is the matrix of intercorrelations among the p_2 measurements in z_2. These structure coefficients are most useful in understanding and interpreting the canonical components in terms of their relations with the known measurements on which they are based. Meredith first suggested the use of structure coefficients in the interpretation of canonical components in 1964 (M. Meredith, 1964, pp. 55–64). A convenient property of orthogonal factors, such as canonical factors, is that the structure coefficients are equal to the factor loadings.

Several indices may be computed from the factor structure matrices. The sum of squares across a row of S_1 or S_2 gives the communality of a measurement, which is the proportion of the variance of the measurement explained or extracted by the n canonical factors. The sum of squares for a column (a factor) of S_1 or S_2, divided by the number of rows (p_1 or p_2), gives the proportion of the generalized variance of the measurement domain extracted by that factor. The *redundancy* of each factor in one set of measurements when the corresponding factor of the other set is available is given by the product of the proportion of variance extracted by the factor times the squared canonical correlation. That is, the redundancy of factor k of z_1 when factor k of z_2 is available is

$$U_{1k} = \left| \frac{1}{p_1} \sum_{j=1}^{p_1} {s_{1_{jk}}}^2 \right| R_k^2.$$

Notice that the bracketed term is the proportion of variance extracted by the term. In general these redundancies of corresponding factors will not be

equal because the factors will not extract the same proportions of variance from their respective domains. Redundancy is an important index because a factor may have a strong canonical correlation but be a trivial explanatory construct for the measurement vector it is based on. The redundancy measure was proposed by Stewart and Love in 1968 (D. K. Stewart and W. A. Love, 1968, pp. 160–163). The total redundancy of one set of canonical factors given the availability of the other set is simply the sum of the redundancies for the factors. That is,

$$U_1 = \sum_{k=1}^{N} U_{1k} \quad \text{and} \quad U_2 = \sum_{k=1}^{N} U_{2k}.$$

T. W. Anderson (T. W. Anderson, 1958, pp. 288–306) proves that the complete canonical analysis is computed as the eigenstructure of the nonsymmetric matrix product

$$R_{22}^{-1} R_{21} R_{11}^{-1} R_{12}$$

where R_{11} and R_{22} are as previously defined, R_{12} is the cross-correlations between the measurements in z_1 and those in z_2, and

$$R_{21} = R_{12}'.$$

The eigenstructure is

$$(R_{22}^{-1} R_{21} R_{11}^{-1} R_{12}) V = VL$$

where L is a diagonal matrix of eigenvalues λ_j, and V is a matrix in which the jth column contains the eigenvector for λ_j. The eigenvalues of this matrix product are the squared canonical correlation coefficients R_k^2. The right eigenvectors are weights for the canonical variates of z_2. Since eigenvectors are defined only up to a constant of proportionality which may vary as a function of the numerical analysis procedure employed, it is necessary to scale the raw eigenvectors to guarantee unit variances for the canonical factors. Letting V be the raw eigenvectors, we get the desired coefficients for factors of z_2 as

$$D = V(V' R_{22} V)^{-\frac{1}{2}}$$

(the uncorrelatedness of the factors is verified by the observation that $V'\, R_{22}\, V$ is a diagonal matrix). When we have the coefficients d_k for the kth factor of z_2 (as the k-th column of D), the corresponding coefficients for the kth factor of z_1 are obtained as

$$c_k = (R_{11}^{-1} R_{12} d_k) \frac{1}{R_k}$$

This formula is derived by T. W. Anderson (T. W. Anderson, 1958, pp. 288–306). A complete account of the canonical analysis procedure may be found in Cooley and Lohnes who list a FORTRAN program for computing such analyses which was used to compute the two research examples which follow in this presentation, and who provide the following numerical example (W. W. Cooley and P. R. Lohnes, 1971).

Assume these relationships among two measures in z_1 and two measures in z_2:

$$R_{11} = \begin{vmatrix} 1.00 & .40 \\ .40 & 1.00 \end{vmatrix} \qquad R_{22} = \begin{vmatrix} 1.00 & .20 \\ .20 & 1.00 \end{vmatrix}$$

$$R_{12} = \begin{vmatrix} .50 & .60 \\ .30 & .40 \end{vmatrix}$$

Then the required matrix product $R_{22}^{-1} R_{21} R_{11}^{-1} R_{12}$ is:

$$\begin{vmatrix} 1.041 & -.208 \\ -.208 & 1.041 \end{vmatrix} \cdot \begin{vmatrix} .50 & .30 \\ .60 & .40 \end{vmatrix} \cdot \begin{vmatrix} 1.190 & -.476 \\ -.476 & 1.190 \end{vmatrix} \cdot \begin{vmatrix} .50 & .60 \\ .30 & .40 \end{vmatrix} = \begin{vmatrix} .206 & .251 \\ .278 & .341 \end{vmatrix}$$

The required eigenvalues may be obtained by expanding the following determinant and calculating the roots of the resulting quadratic equation:

$$\begin{vmatrix} .206-\lambda & .251 \\ .278 & .341-\lambda \end{vmatrix} = 0$$

These roots are $\lambda_1 = R_1^2 = .546$ and $\lambda_2 = R_2^2 = .001$, so that $R_1 = .74$ and $R_2 = .03$. We easily decide that the second canonical correlation is trivial in magnitude and should be ignored, so we choose the rank of our canonical model to be $n = 1$. The raw eigenvector that goes with the first eigenvalue may be obtained as the cofactors of the first row of the determinantal equation:

$$v_1 = \begin{vmatrix} -.205 \\ -.278 \end{vmatrix}$$

Solving the scaling equation $d_1 = v_1(v_1' R_{22} v_1)^{-\frac{1}{2}}$ yields

$$d_1 = \begin{vmatrix} .545 \\ .737 \end{vmatrix}$$

Solving the equation $c_1 = (R_{11}^{-1} R_{12} d_1) \frac{1}{R_1}$

$$c_1 = \begin{vmatrix} .856 \\ .278 \end{vmatrix}$$

Thus the first canonical factors are

$$x_1 = .856 z_{1_1} + .278 z_{1_2} \quad \text{and} \quad y_1 = .545\, z_{2_1} + .737\, z_{2_2}$$

To get the structure coefficients which represent the correlations between these canonical factors and the measurements on which they are based we form

$$s_1 = R_{11} c_1 = \begin{vmatrix} .967 \\ .620 \end{vmatrix} \quad \text{and} \quad s_2 = R_{22} d_2 = \begin{vmatrix} .692 \\ .846 \end{vmatrix}.$$

The proportion of the generalized variance in the first set of measures extracted by factor x_1 is $[(.967)^2 + (.620)^2]/2 = .660$, and this times the first canonical R^2 gives .36 as the redundancy of the first factor of z_1 given the availability of the first factor of z_2. Similarly, the proportion of the generalized variance in z_2 extracted by y_1 is .597, and the redundancy of y_1 given x_1 is .33. Since the canonical model is of rank one, it also appears that $U_1 = .36$ and $U_2 = .33$.

What has been revealed about the structure of relationships among these four variates by the canonical analysis? First of all, the three correlation matrices, R_{11}, R_{22}, and R_{12}, reveal that the cross-correlations between variates of z_1 and those of z_2 are stronger than the internal correlations within z_1 or within z_2. The strongest bivariate correlation in the system is .60 between the

first variate of z_1 and the second variate of z_2. A factor of z_1 has been located that correlates .74 with a located factor of z_2. The structure coefficients show that the canonical factor of z_1 has a very strong correlation (.97) with the first variate of z_1 and a moderate correlation (.62) with the second variate of z_1. This canonical factor is an important explanatory construct for z_1, inasmuch as it extracts 66% of the generalized variance. That the strongest correlation (.85) of the factor of z_2 is with the second variate of z_2 is consistent with the location of the strongest bivariate cross-correlation. Again this is a general factor of some importance in explaining z_2, since it extracts 60% of the generalized variance. The two redundancy indices are nearly the same (.36 and .33) and may be taken as indications that the canonical model shows approximately a third of the generalized variance in each measurement domain as redundant, given the other.

The most significant result is perhaps the overwhelming adequacy of a rank one model for the canonical correlation analysis of the data. Parsimony has been achieved without loss of precision.

MORBIDITY AND SOCIO-ECONOMIC TRAITS IN BUFFALO

The application and interpretation of canonical correlation analysis is now illustrated by a highly-simplified example based on a recent study of the interrelationships between and the spatial structure of morbidity and socio-economic traits in Buffalo.[1] Morbidity research has revealed significant relationships between disease and socio-economic characteristics in a number of United States urban centers. For example, the incidence of cancer of the respiratory system has been found to increase among the lower socio-economic groups in studies conducted in Buffalo and New Haven (E. M. Cohart, 1955, pp. 455–461). Nevertheless, no previous attempts had been made to explore the relationships between a wide range of morbidity and socio-economic characteristics.

Recent records for nine types of morbidity were gathered for the Buffalo Standard Metropolitan Statistical Area.[2] These data were coded and aggregated by census tract to make them compatible for analysis with a selection of seventy-four socio-economic variables from the 1960 census.

The morbidity and socio-economic measurement domains have been reduced to their underlying dimensions by separate factor analyses for the purposes of this example and a reworked canonical correlation analysis is presented below using a selection of four of the socio-economic and three of the morbidity factors. The first socio-economic dimension describes the young-married white population, including some Canadian and United Kingdom immigrants, predominantly in the age group 25 to 50 years, with children under 15. This group is associated with post-war constructed owner-occupied housing, and a diversified occupation structure including professionals, and machinery and transportation workers. A second white factor identifies young-single, and older im-

[1] The authors wish to thank Mr. Mindangus Matulionis for his assistance in providing the data, which were gathered for his M. A. thesis, State University of New York at Buffalo, Department of Geography, 1970.

[2] The data were made available by the Respiratory Disease Association of Western New York Inc., the Public Health Research Institute for Chronic Diseases of the State University of New York Medical School at Buffalo, the Erie County Health Department and the Roswell Park Memorial Institute, a Cancer research hospital. The morbidity data was classified where possible, by sex and race providing twenty-two morbidity measures.

migrant groups (predominantly Germans and Italians and with some Canadians and British) living in older downtown housing and largely in the 15 to 25 age group, but also with concentrations of population over 50 years. The black population forms a separate factor, reflecting their distinctive location, their relative concentration in unskilled occupations and their more-crowded housing conditions. Dilapidated housing, however, is not significantly associated with the proportion of census tract population that is black, and comprises a specific fourth factor.

Certain of these socio-economic characteristics are explicitly recognized in the morbidity data. The data on respiratory diseases are classified by the two dominant racial groups, with the exception of lung cancer and emphysema for which no data are available for the black population. The asthma data are restricted to children 16 years and under, the age group for which it is most prevalent.

The racial classification of the morbidity cases reflects itself in the factors. Asthma separates into two factors, the first identifying non-black ethnic groups, the second linking asthma and tuberculosis in both the male and female black population. The third morbidity factor identifies all male emphysema cases and mild female emphysema cases.

The scores of the 172 Buffalo census tracts on the four socio-economic and three morbidity factors are used as input variables for the canonical correlation analysis. The canonical factor structure matrix, shown in Table 1, reveals two significant and one residual pattern of association between the two measurement domains. Not surprisingly, the black population and incidence of respiratory diseases among negro population are strongly associated; indeed, every census tract with a high score (above 2.0) on the black socio-economic variable, has a correspondingly high score on the black respiratory morbidity incidence.

TABLE 1. Canonical correlation analysis of socio-economic and morbidity data: Buffalo, N. Y.

	Canonical Factors		
	I	II	III
Socio-economic variables			
(1) Young-married white	−.430	−.866	.214
(2) Single, downtown white	.241	.143	.541
(3) Black	.868	−.466	−.019
(4) Dilapidated housing	.030	−.120	−.818
Morbidity variables			
(1) Asthma (white)	−.259	−.719	.647
(2) Respiratory (black)	.939	−.343	−.029
(3) Emphysema (white)	.090	.586	.805
Canonical correlation	.915	.830	.093
Chi-square	498.53	193.03	1.44
Degrees of Freedom	12	6	2

Note: The variables are the scores on the dimensions of separate varimax-rotated factor analyses of socio-economic and morbidity data for the 172 census tracts of the Buffalo S.M.S.A.

Canonical variables I and II are highly significantly correlated.

The canonical correlation between the first pair of canonical factors, .915, is in fact a little higher than the simple correlation between the black socio-economic variable and the black respiratory-disease variable, which is 878 (see Table 1).

The second pair of canonical factors identifies the age differences between the predominant occurrence of asthma (children) and tuberculosis (under forty years) on the one hand, and emphysema (concentrated among males over forty) on the other, and underlines the high asthma incidence in census tracts with a predominant young-married white population. Once again the canonical correlation (.830) is higher than the simple correlation between the two variables with their highest loading on the canonical variates (.630 between young-married white and asthma). None of the socio-economic variables used identifies by a negative sign with emphysema in the second pair of canonical factors, and no importance can be attached to the high coefficients of emphysema, single downtown white population and dilapidated housing on the third, residual pair of canonical variates because of the low canonical correlation.

The variance extracted from the measurement domains by each of the three pairs of canonical correlates is much the same, but the redundancy measure for the third pair is severely reduced by the low canonical correlation (see Table 2).

TABLE 2. Canonical variance extracted and redundancy: Buffalo example

Canonical factors	Squared canonical correlation	Socio-economic variables		Morbidity variables	
		variance extracted	redundancy	variance extracted	redundancy
1	.837	.250	.209	.318	.266
2	.688	.251	.172	.326	.224
3	.009	.251	.002	.356	.003
Total		.752	.383	1.000	.494

This analysis of the Buffalo morbidity-socio-economic variables is highly simplified, and the complete data indicate other inter-relationships, such as the relative concentration of cancer of the digestive system among the Polish-ethnic population, related, presumably, to their diet. Nonetheless, the first two pairs of canonical variates in the above analysis also dominate the factor structure in the canonical correlation analysis of the complete data, and this analysis reveals more than can be learned from a simple correlation alone of the two sets of scores from the factor analysis, which, it will be recalled, is the most competitive modeling procedure.

THE OCCUPATION AND CULTURAL STRUCTURE IN CANADA

The second research example investigates the interrelationships and spatial structure of occupational and cultural characteristics in Canada. Since Confederation in 1867, Canada has achieved a rate of economic and population growth that is among the highest in the world. Migration has contributed a surprisingly small proportion of the population increase because of heavy em-

migration, but it has produced a cultural diversity that is retained as a national mosaic with associated occupational and social-class attributes. These attributes are spatially interwoven to produce distinctive regionalisms that threaten national unity and, as demonstrated by the recent tragic events in Quebec Province, "it is unlikely that regionalism is anywhere at the centre of more national problems" (G. Merrill, 1968, pp. 531–555).

An earlier attempt to investigate the interrelationships between cultural and economic characteristics in seven of Canada's ten provinces commented on the difficulty of identifying culture (D. M. Ray, 1971). Culture is defined by the Canada Royal Commission on Bilingualism and Biculturalism as "a way of being, thinking and feeling. It is a driving force animating a significant group of individuals united by a common tongue, and sharing the same customs, habits and experience" (Canada, 1967, p. XXXI). The earlier canonical correlation study employed census data on place of birth, ethnic origin and religion together with period of immigration to identify culture, although the analysis revealed serious errors in the census ethnic data. This example employs mother tongue as a surrogate measure of culture because it is less subject to ennumeration error than ethnic origin and because it may be a more sensitive indicator of cultural attachment.[3] This example also uses an expanded set of seven occupation groups supplemented by the labor-force participation rate, unemployment rate, male/female employment ratio and average family income for the whole of Canada.

Some high correlations occur both among the occupational and the mother tongue data, and between the two measurement domains (see Tables 3, 4 and 5). The occupational correlations underline the expected spatial association of the professional and managerial work-force (.643) and the contrast between areas with concentration of such workers and those with concentrations of farmers and a high male to female ratio of the labor force. The highest positive correlation in the complete R_{11} matrix is .770 between per cent of the county male labor force in professional occupations and county average family income. The highest negative correlation is — .724 between farmes and craftsmen.

The correlation table for mother tongue reveals the almost complete spatial separation of English and French cultures in Canada ($r = -.940$). This separation can be traced to the locational differences in initial colonization, the tendency of English settlers to migrate from any area becoming predominantly French-Canadian, and the tendency for minority groups to be assimilated. The complete R_{22} matrix indicates that the German, Italian, Polish and Ukrainian all have positive correlations with English and negative with French, revealing the tendency for minority ethnic groups to be located in regions that are predominantly English. The Jewish group is an important exception, being almost entirely located in the largest metropolitan centers, including Montreal.

The high within-group correlations make the patterns of between-group association more difficult to discern, though the R_{21} matrix indicates some of the important dimensions of association. The highest positive correlation with per cent of county labor force in managerial occupations is English (+.323) and the highest negative is French (−.262). The census data in fact reveals that 7.6% of the French ethnic labor force is employed in managerial work, compared with 12.1% of the British and 10.2% of the total labor force. Even in Quebec

[3] Ethnic origin is based upon the racial, linguistic or national origin of the individual, if he is an immigrant, or, in the case of native-born Canadians, of the paternal ancestor who first entered the North American continent. Mother tongue is more simply the language learned at birth and still spoken. Some assimilation of ethnic minority groups has occurred and, for example, 10% of Canada's population reporting French ethnic origin, have English mother tongue.

TABLE 3. Correlation matrix (R_{11}) for selected occupational data: Canada 1961

No.	Variable Name	1 Managers	2 Professional	3 Farmers	5 Miners
(2)	Professional	.643			
(3)	Farmers	−.396	−.534		
(5)	Miners	−.124	−.032	−.236	
(9)	Male/Female L.F. ratio	−.529	−.448	.197	.220

Note: The occupation data are for male labor force as a per cent of total male labor force. For definitions and a discussion of the ethnic composition of the labor force see Canada, Dominion Bureau of Statistics, **1961** *Census of Canada, General Review, Series 7.1 Bulletin 12 The Canadian Labor Force*, 1967.

The correlations indicate the degree and direction of spatial association at the census county level.

TABLE 4. Correlation matrix (R_{22}) for selected mother tongue data: Canada, 1961

Mother Tongue	English	French	Indian and Eskimo	Italian	Polish
French	−.940				
Indian and Eskimo	−.010	−.172			
Italian	.136	−.195	.024		
Polish	.190	−.410	.176	.283	
Yiddish (Jewish)	−.041	.004	−.053	.363	.147

Note: Mother tongue is language learned at birth and still spoken.

TABLE 5. Correlation matrix (R_{21}) for selected occupational and mother tongue variables by census county: Canada, 1961

Mother Tongue	Managerial	Professional	Farmers	Loggers etc.	Miners	Male/Female Labor Force Ratio	Average Family Income
English	.323	.089	−.060	.051	−.001	.123	.080
French	−.262	−.065	−.096	.021	−.043	−.145	−.098
Indian and Eskimo	−.155	−.068	.040	.134	.348	.190	−.072
Italian	.308	.379	−.334	−.106	.241	−.153	.507
Polish	−.198	−.152	.477	−.143	−.043	.064	−.156
Yiddish (Jewish)	.284	.347	−.166	−.135	−.031	−.229	.306

Note: "Loggers" is defined by the census of Canada to include fishermen, trappers and hunters. See Tables 3 and 4 for other data definitions.

Province, which is 81% French, "the firms which provide most employment and which most influence the course of economic development are owned and controlled by English-language interests," (Canada, 1967, pp. IX–IV) so that the proportion of the French-Canadian labor force in managerial occupations is exceeded, even in Quebec Province, by almost every other ethnic group (see Table 6). Thus A. H. Richmond (1969, p. 7) writes,

> "Already, the economic domination of American and English-Canadian companies means that, with the growth of industrialization, immigrants and native-born French-speaking Canadians alike have found themselves compelled to learn English in order to take advantages of opportunities for upward social mobility. The present crisis in Quebec is in large part an indication of the determination of French-Canadians that this will not be the sine qua non of individual or collective prosperity in the future. A 'quiet revolution' has been taking place ..."

The highest correlations between the county proportion of the male labor force in professional occupations and the mother tongue proportions of the county population are for Italian and Jewish. (See Table 5.) The Jewish labor force has the most extreme occupation profile in Canada with 13.7% in professional and technical occupations compared with 7.6% for the total labor force. By contrast only 2.8% of Italian labor force is in professional and technical occupations and 19.2% is employed as laborers, compared with 1.1% for the Jewish. But the Italian group is almost as concentrated in large urban centers as the Jewish (75% of the Italians compared with 94% of Jews living in the census metropolitan areas in 1961), so that their county intercorrelation (.362) is the highest either has with any mother tongue group. Thus the Italian group shares, vicariously, the high association of the Jewish population with professional occupations and high average family income. The Polish group identifies with farming, reflecting the relative concentration of the slavic population in the farming regions of western Canada. The Indian and Eskimo population has rather low correlations with all the occupational variables, but, as expected, has its highest correlations with loggers (which includes fishermen, trappers and hunters), miners and a high male to female labor force ratio.

As in the previous example, the interrelationships between the two measurement domains are more clearly focussed in the factor structure matrix than in the simple correlation matrix so that the correlations of the occupational and mother tongue variables are higher with their respective canonical factors than

TABLE 6. Per cent of male labor force, 15 years of age and over, in managerial occupations in Quebec Province, by ethnic group: 1961

Ethnic group	Per cent	Ethnic group	Per cent
French	7.9	Polish	13.3
British	15.4	Scandinavian	14.8
German	11.2	Ukrainian	8.1
Hungarian	9.9	Other European	15.2
Italian	6.1	Asiatic	25.1
Jewish	37.7	Native Indian	3.1
Netherlands	14.2	Total	9.6

Source: Canada, DBS, (1967), pp. 12–114 to 12–115.

with each other (see Table 7). The first canonical factor, which is examined here in some detail, loads on the Polish-Ukrainian-German population and farming. The structure coefficients are somewhat higher for the cultural than for the labor force variables so that the variance extracted (which equals the mean sum of squares) is .232 compared with .184 (see Table 8). Multiplying the variance extracted by the squared canonical correlation (.512) gives a redundancy in the labor force data given the cultural data of .094, and of .232 in the cultural data given the labor force data. In part, the difference between the two redundancy measures may reflect the three extra labor force variables which may add to the dimensionality, or rank, of the labor force matrix. In part, the difference in the redundancy measure reflects a truly asymmetric relationship between occupation and mother tongue on the first canonical factor in which three cultural groups are highly associated with a single occupation.

Two scores may be computed for each county on the first canonical factor score and the simple correlation computed between these scores and a set of

TABLE 7. Occupational and mother tongue canonical factors: Canada, 1961

	Canonical factors				
	I	II	III	IV	V
Labor force					
(1) Managerial	−209	476	508	−436	170
(2) Professional	−173	557	128	−248	509
(3) Farm Workers	803	−503	−052	−190	−222
(4) Loggers	−351	−122	117	528	−196
(5) Miners	003	453	−175	611	−241
(6) Craftsmen	−624	461	−312	−188	154
(7) Laborers	−625	109	−225	292	−004
(8) Labor Force Participation Rate	484	593	−012	−342	185
(9) Male/Female Labor Force Ratio	113	−172	176	486	−525
(10) Unemployment	−397	091	033	484	227
(11) Average Family Income	011	772	−038	−513	156
Cultural variables					
(1) English	041	222.	885	011	−313
(2) French	−358	−285	−806	−069	296
(3) German	665	034	137	−259	−220
(4) Indian & Eskimo	328	255	−032	764	−101
(5) Italian	−029	895	−109	−065	078
(6) Polish	746	340	028	055	109
(7) Ukrainian	788	−193	−022	208	248
(8) Yiddish	046	390	217	−217	836
Canonical Correlation	.715	.703	.599	.489	.372
Chi-Square	510.67	354.38	206.08	109.39	39.69
Degrees of freedom	88.	70	54	40	28

Note: Leading decimal points omitted in structure coefficients. Three additional factors are extracted but have low redundancies (see Table 8).

TABLE 8. Canonical variance extracted and redundancy: Canada example

Canonical factor	Squared canonical correlation	Labor force		Cultural variables	
		variance extracted	redundancy	variance extracted	Redundancy
1	.512	.184	.094	.232	.119
2	.494	.202	.100	.163	.080
3	.358	.046	.016	.187	.068
4	.240	.174	.042	.094	.023
5	.139	.077	.011	.128	.018
6	.044	.058	.003	.060	.003
7	.022	.066	.001	.040	.001
8	.012	.055	.001	.093	.001

marker variables to give the canonical factors spatial expression (see Table 9). Again the cultural characteristics have greater regularity and have higher correlations but both possess a significant east-west gradient that has been identified in previous studies. This gradient has been related to centripetal and centrifugal forces acting at an intercontinental scale during the early staple-export phase of Canada's economic development when it was closely tied to Northwest Europe (W. T. Easterbrook and M. H. Watkins, 1967, and D. M. Ray, 1971).

TABLE 9. Spatial structure of the canonical factors

Canonical factor	Simple correlation with: distance form					
	Population Potential	Vancouver	Winnipeg	Toronto	Montreal	Halifax
East-West Contrasts (I)						
Labor force scores	−.285	−.425	−.568	.222	.403	.615
Mother tongue scores	−.361	−.516	−.747	.339	.537	.650
Heartland-Hinterland (IV)						
Labor force scores	−.434	.153	.070	.320	.161	−.202
Mother tongue scores	−.250	.107	.034	.099	.062	−.079
Metropolitan (II and V)						
Labor force scores	.203	−.265	.027	−.205	.128	.308
Mother tongue scores	.182	−.265	.047	−.215	.170	.303
Labor force scores	.298	.165	.041	−.118	−.174	−.156
Mother tongue scores	.386	.226	−.148	−.137	−.325	−.118
English-French contrasts (III)						
Labor force scores	−.160	−.143	.004	.087	.320	−.020
Mother tongue scores	−.252	−.054	.090	.029	.428	−.050

The east-west progression of values on the first canonical factor are illustrated by the provincial averages for five of the variables with high loadings (see Table 10). The per cent of the county male labor force in farming increases east to west across Canada with a correlation of .429 and the provincial averages follow this trend with the severe exceptions of Prince Edward Island and British Columbia. The percentage of the male labor force amployed as craftsmen, who essentially comprise the factory workers, has a negative loading (−.624) on canonical factor one and a simple correlation with distance from Halifax of −.316, although the provincial range in values is much smaller than for farmers.

TABLE 10. Provincial values for characteristics with high coefficients on the east-west canonical factor

Province	Total population in 1000's	No. of Census Counties	% of male labor force employed as		% of population with mother tongue		
			farmers	craftsmen	German	Polish	Ukrainian
Newfoundland	458	10	1.8	27.7	0.1	—	—
Prince Edward Island	105	3	32.8	18.6	0.1	0.4	0.1
Nova Scotia	737	18	6.7	25.7	0.2	0.3	0.1
New Brunswick	598	15	9.2	25.7	0.2	0.2	0.1
Quebec	5,259	66	9.1	31.0	0.6	0.1	0.3
Ontario	6,236	54	8.8	31.5	2.9	1.4	1.4
Manitoba	922	20	21.3	23.9	9.1	1.4	9.2
Saskatchewan	925	18	43.2	16.6	9.7	1.9	7.3
Alberta	1,332	15	25.2	21.6	7.3	1.8	6.3
British Columbia	1,629	10	5.1	30.8	4.4	1.5	1.2
Canada	18,238	229	12.2	28.8	3.1	0.9	2.0

Note: Canada total includes 10 provinces and the Yukon and Northwest Territories
Source: Canada, DBS (1967), p. 12–9

The east-west increase in the per cent of the population with German, Polish and Ukrainian mother tongue is more regular than that of the employment characteristics; the simple correlations with distance from Halifax are .542, .531 and .431 respectively. These increases are indicative of the general increase in cultural heterogeneity east to west across Canada; no provinces show greater cultural heterogeneity than the Prairie Provinces (where more than a quarter of the population have a minority-group mother tongue) and none show less than the Atlantic Provinces (where less than 3% of the population have mother tongue other than English or French).

The east-west gradient of economic and cultural variation is a pervasive element of Canadian geography shared by many other characteristics not included in this analysis. Indeed, some characteristics such as population distinguished by Canadian province or country of birth and by period of immigration have much higher correlations with distance from Halifax than any of the characteristics analyzed here. This east-west gradient identifies the relation-

ships among accessibility, physical resource endowment and timing of regional development and the changing European origins of settlers as transportation improved and the scale of spatial interaction widened from trans-Atlantic to semi-global (Canada, 1965).

Similarly the remaining canonical factors identify other pervasive elements of Canadian geography. The fourth labor force canonical factor focuses on regional disparities in income found in areas with lower population potential and it displays a heartland-hinterland pattern (see Table 7 and 9). The highest labor force scores are for nothern Manitoba and northern Saskatchewan, and for the Newfoundland census counties excluding those along the Corner-Brook — St. John's development axis. The fourth mother tongue canonical factor identifies the areas with relative concentrations of Indian and Eskimo but the canonical correlation is rather low (.489) reducing the redundancy of this factor (see Table 8). Two pairs of metropolitan factors emerge which appropriately distinguish the Italian and Jewish mother tongues and the typical urban occupation profile and a concentration in professional occupations respectively. The scores and values for selected variables, given in Table 11 for the counties containing Canada's seventeen metropolitan areas, suggest the systematic heartland-hinterland and east-west gradients in the metropolitan characteristics of the second pair of canonical factors and the relative high concentration of Jewish population in just three cities, Montreal, Toronto and Winnipeg which helps to explain the low canonical correlation for the fifth pair of canonical correlates. English-French contrasts, which did not load on any canonical factor in a previous canonical correlation analysis, have the third highest canonical correlation in this analysis (.599) (D. M. Ray, 1971). Nevertheless, the variance extracted from the labor force data is only .046 and the corresponding redundancy, .016, compared with a redundancy of .068 for this factor on the cultural data. As far as can be judged from available census data, French-English differences are predominantly cultural, notwithstanding the low participation of French-speaking Canadians in managerial occupations.

CONCLUSIONS

Canonical correlation analysis is considerably enhanced by the four new interpretive devices described and applied in this paper and it becomes a much more powerful analytic technique than factor analysis, the most competitive modeling device, where two measurement domains are being analyzed. The factor structure matrix identifies the canonical factors by their correlations with the measurement domains, as in factor analysis. The variance extracted corresponds to the per cent eigenvalue in factor analysis. The product of the variance extracted by the squared canonical correlation indicates the variance accounted for in one measurement domain given the other, and it is the best single indication of where the factoring should stop, or the rank of the models. This redundancy measure has no equivalent in factor analysis. The canonical factor scores increase the flexibility of the technique and permit further analysis of the spatial expression of the canonical factors using marker variables, mapping and regionalization. The disparities between the scores for any factor on the two measurement domain may be mapped to reveal the spatial deviations from the factor association. Such disparities may also appear as secondary factors with low canonical correlations. The two research examples indeed only probe the application of this technique but should convince most readers of the power of canonical correlation in geographical analysis.

TABLE 11. The metropolitan scores and marker variables on factors II and V

Metropolitan area	Pop. in 1000's M.A.	Pop. in 1000's C.C.	Canonical factor II: labor force score	Canonical factor II: average family income	Canonical factor II: mother tongue score	Canonical factor II: % Italian	Canonical factor V: labor force score	Canonical factor V: % professional	Canonical factor V: mother tongue score	Canonical factor V: % Jewish
Atlantic Prov. Mean			−.14	$4773	−.42	.07	1.15	6.88	.01	.08
St. John's	91	189	−1.26	4043	−.54	.04	.45	6.36	−.20	.02
Halifax	184	226	.46	5331	−.33	.19	.160	7.42	−.01	.08
St. John	96	89	.36	4946	−.40	.09	1.40	6.87	.25	.13
Quebec Prov. Mean			1.21	5961	1.49	2.32	3.08	10.73	3.90	.95
Montreal	2110	1872	2.37	6098	3.43	4.49	4.24	10.39	7.26	1.77
Quebec	358	331	.04	5823	−.45	.19	1.92	11.07	.53	.10
Prov. of Ontario Mean			1.72	6026	2.06	3.32	.70	8.78	.85	.40
Hamilton	395	359	1.40	5914	3.30	5.07	.55	7.69	.04	.21
Kitchener	155	177	2.04	5822	.10	.50	.19	6.65	−.26	.12
London	181	221	1.16	5824	.68	.96	1.15	8.43	−.41	.10
Ottawa	430	353	1.51	6879	.95	1.87	2.79	15.19	1.00	.34
Sudbury	111	166	3.00	5973	2.34	4.05	−2.30	5.39	−.33	.10
Toronto	1824	1733	2.17	6459	4.48	6.55	1.49	10.79	5.35	1.57
Windsor	193	258	.79	5311	2.59	4.20	1.07	7.30	.57	.33
Prairie Prov. Mean			1.51	6010	.81	.94	1.23	9.41	2.76	.75
Winnipeg	476	476	1.30	5874	1.50	.83	1.33	8.71	7.98	1.85
Calgary	279	318	1.72	6255	.67	1.10	.95	10.21	.07	.24
Edmonton	338	411	1.52	9502	.25	.89	1.40	9.51	.25	.17
B. Columbia Mean			.53	5629	.61	1.02	.26	8.30	−.72	.09
Vancouver	790	908	.89	5816	.87	1.38	.65	8.97	−.61	.14
Victoria	154	291	.17	5442	.34	.65	−.14	7.62	−.82	.03
Can. Metro. Mean			1.16	5748	1.16	1.94	1.10	8.73	1.21	1.02
Can. Census Div. Mean			0.00	4463	0.00	.54	0.00	5.12	0.00	.08

Note: Population is given for the metropolitan area (M.A.) and census county (C.C.), all other data are for the census county in which the metropolitan area is located.

State University of New York, Buffalo

ACKNOWLEDGEMENTS

The authors wish to thank the State University of New York Education Department and the Faculty of Social Sciences and Administration of the State University of New York at Buffalo for travel grants to present this paper at Poznań.

BIBLIOGRAPHY

Government

Canada Dominion Bureau of Statistics, 1965, *1961 Census of Canada, General Review Series 7.1 Bulletin 6: Origins of the Canadian Population*, Ottawa, Queen's Printer.

Canada Royal Commission on Bilingualism and Biculturalism, 1967, *General introduction: Book I: The official languages*, Ottawa, Queen's Printer, p. XXXI.

Other

Anderson, T. W., 1958, *An introduction to multivariate statistical analysis*, New York, Wiley, 288–306.

Berry, B. J. L., 1966, *Essays on commodity flows and the spatial structure of the Indian economy*, University of Chicago, Dept. of Geography, Research Paper 111.

Berry, B. J. L., 1968, A synthesis of formal and functional regions using a general field theory of spatial behavior, in Brian J. L. Berry and Duane F. Marble (Eds.), *Spatial analysis: A reader in statistical geography*, Englewood Cliffs, N. J., Prentice-Hall Inc., 419–428.

Cohart, E. M., 1955, Socio-economic distribution of stomach cancer in New Haven, *Cancer*, 7, 455–461.

Cooley, W. W., and Lohnes, P. R., 1971, *Multivariate data analysis*, New York, Wiley.

Easterbrook, W. T. and Watkins, M. H. (Eds.), 1967, *Approaches to Canadian economic history*, Toronto, McClelland and Stewart, Ltd.

Gauthier, H. L., 1968, Transportation and the growth of the São Paulo economy, *J. Reg. Sci.*, 8, Summer, 77–94.

Gould, P. R., 1969, Methodological developments since the fifties, in Christopher Board *et al.* (Eds.), *Progress in geography*, London, Edward Arnold, 13–14.

Hotelling, H., 1935, The most predictable criterion, *J. Educ. Psychol.*, 26, 139–142.

King, L. J., 1969, *Statistical analysis in geography*, Englewood Cliffs, N. J., Prentice-Hall, Inc., 217–222.

Matulionis, M., 1971, *The spatial interrelationships of morbility and socio-economic characteristics: Buffalo, N.Y.*, M.A. Thesis, Dept. of Geography, State Univ. of New York at Buffalo.

Meredith, W., 1964, Canonical correlation with fallible data, *Psychometrika*, 29, 55–65.

Merrill, G., 1968, Regionalism and nationalism, in John Warkentin (Ed.), *Canada: A geographic interpretation*, Toronto, Methuen Press, 531–555.

Ray, D. M., 1971, The spatial inter-relationships of economic and cultural differences: A canonical ecology of Canada, *Econ. Geogr.*, 47, 344–355.

Richmond, A. H., 1969, Immigration and pluralism in Canada, *The International Migration Rev.*, 4, Fall, p. 7.

Stewart, D. K. and Love, W. A., 1968, A general canonical correlation index, *Psychol. Bull.*, 70, 160–163.

GEOGRAPHIA POLONICA 25, 1973

THE PRACTICAL APPLICATION OF ONE DIMENSIONAL SPECTRAL ANALYSIS

JOHN N. RAYNER

(1) INTRODUCTION

In recent years the geographer has been exposed to a bewildering variety of quantitative techniques the usefulness of many of which have yet to be demonstrated. Furthermore, discussions of the lesser known techniques tend to be very technical or, alternatively, are limited in scope. In consequence the relative values of these procedures are difficult to assess by the majority of geographers and are quickly dismissed. One such technique is spectral analysis which is often wrongly classified as being too complicated or being applicable only to periodic data sets of which there are few. This paper attempts to review briefly at a relatively non technical level the scope of the technique, or better, group of techniques, which may be labelled "spectral" or "Fourier", and to describe in detail the simple though long calculations involved. Central to these is the Fourier transformation which is nothing more than a particular form of curve fitting by least squares. At the outset it should be noted that these techniques usually apply to data which have equally spaced coordinates in space and/or time. Other arrangements of data are possible but they will not be included in the present discussion.

(2) WHAT IS SPECTRAL ANALYSIS?

(2.1) GENERAL

As used by the author this procedure involves the analysis of data through a particular transformation which arranges the results in the form of a spectrum. The term spectrum has the same meaning as is used for the decomposition of light, or more generally, of electromagnetic waves. It refers to the ordering of characteristics of a data set according to scale. Thus, for example, light may be classified by wavelengths varying between 4/100,000ths and 7/100,000ths of a centimeter. Alternatively, because these waves are travelling at a constant speed of 3×10^{10} cm s^{-1} they may be specified by the frequency at which they pass i.e. 7.5×10^{14} to 4.3×10^{14} cycles per second. These are extremely small wavelengths and extremely high frequencies but the same ideas may be used for other phenomena and other scales.

Many social and physical data are recorded continuously or at successive discrete time intervals. For instance, a large number of temporal series are available for the number of vehicles passing specific points in a road network per unit time or for atmospheric pressure systems passing a weather station.

Similarly maps and photographs record two dimensional series with space rather than time as the independent variable. The spectrum rearranges the fluctuations in the temporal (or spatial) series according to rank in size or scale of time (or space). Therefore, the spectrum of traffic presents the data according to the average intervals of time between single vehicles and groups of vehicles. It might be anticipated that a strong diurnal fluctuation would show up: more vehicles are on the streets during daylight than during darkness. This particular class or rank would be labelled one cycle per day, or seven cycles per week, or period of one day, etc. Other scales in time will appear. They may or may not be obvious in the data (see for example Fig. 1). Often, general consistent changes are apparent in the data. These are called trends and will appear at the very largest scales in a given study. For example, if interest were centered upon fluctuations of temperature of the scale of seconds and the total length of observations were an hour or so, the diurnal fluctuation would appear as a trend. At the other extreme, if temperature were considered over thousands of years back from the present, the general increase in temperature since the last glaciation would be called a trend.

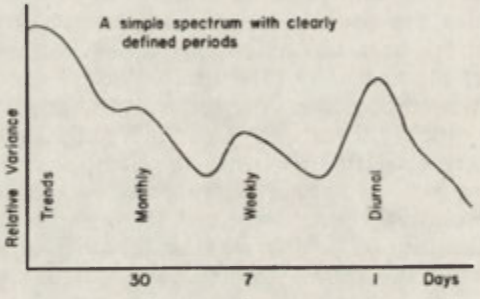


Fig. 1

The spectrum, then, is a particular set of spectacles through which a set of data may be viewed. The spectrum will contain the same information as was present in the original data but in its new form it may be 1) more easily described and interpreted or 2) more easily manipulated.

(2.2) THE SPECTRUM FOR DESCRIPTION AND INTERPRETATION

From the point of view of description and interpretation of data the spectrum must be placed alongside the mean and total variance as a statistical characteristic. This is because coordinate data are usually intercorrelated: adjacent observations tend to be similar. Therefore, even if the data are normally distributed, the mean and variance are not sufficient descriptors. If they were, then the next observation along a traverse would be within one standard deviation of the mean 68% of the time. For most naturally occurring series the last observation on a traverse must be taken into account in predicting the next. For example, if the last observation were in the tail of the distribution then it is highly likely that the next one would be there also and not close to the mean. To quantify this interdependence the autocovariances or autocorrelations are usually calculated. These indicate the degree of relationship between adjacent, alternate, every third, fourth etc. observations. Usually the correlation drops off from $+1$ at zero lag (distance between observations) and oscillates around zero at higher lags, (Fig. 2). The more slowly the autocorrelation function descends the more persistent will be the particular phenomenon being observed.

Unfortunately, it is not easy to put statistical confidence bands around the autocorrelation function and, because the function is itself autocorrelated, it is difficult to interpret beyond the first few lags. The same is not true of the spectrum of the function. This, it turns out, is the scale decomposition of the variance. In other words, it separates the total variance into components which characterize a given size of disturbance. For example, in a time series of temperature the diurnal and annual fluctuations are clearly two separate scales which contribute in an additive sense to the total variance. These components viewed as percentages of the total variance indicate their relative importance and probability of influencing the sequence of temperature. It should be standard practice therefore in describing coordinate data to present the spectrum of the variance as well as the mean and total variance.

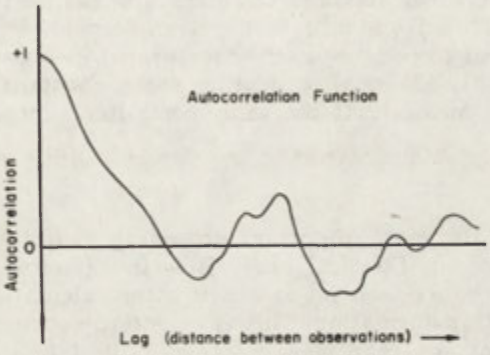


Fig. 2

But the student searches for more than mere description. He also seeks explanation in the form of associations and relationships. The spectrum can be a tool toward this end by indicating different generating processes. Frequently different processes have different spectra: they produce a peak or set of peaks at different scales and the spectra of the phenomena they influence will often contain these peaks. Consequently an analysis of the spectrum may suggest different lines of investigation. As a further step spectral correlation and regression can be carried out between the hypothesized process and the observed phenomenon (Rayner 1967).

Two types of spectra are produced depending upon the assumptions made about the original observations. If the data are assumed to be periodic and repeat themselves indefinitely beyond the limits of observation, the spectrum

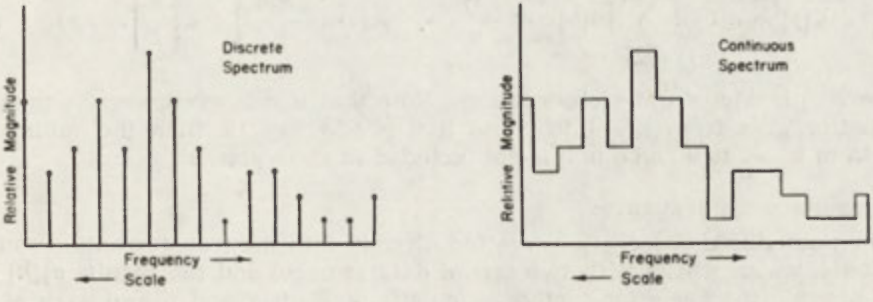


Fig. 3

will be made up of discrete lines corresponding to specific frequencies. On the other hand, if the data are assumed to be non-periodic then the spectrum will be continuous and the characteristic such as variance must be assigned to bands containing an infinite number of frequencies (Fig. 3). The sequence of calculations necessary for each type is different but the transformation procedure, which is relatively simple although somewhat tedious (see section 4), is the same.

(2.3) THE SPECTRUM FOR MANIPULATION

Use of the spectrum for the manipulation of data is analogous to the use of logarithms. Several types of calculation such as the filtering of data, the calculation of differentials, and the recognition of patterns are more easily carried out in the spectral domain. The data are assumed to be periodic, are transformed, manipulated and are then retransformed. Unlike logarithms all the data must be used to produce each transformed element and the procedure for retransformation is identical, excepting for a constant, to that for initial transformation. This means that the same computer subroutine or calculation sheets may be used.

(2.4) SUMMARY

A restatement of the more important preceding points may help clarify the technique for the reader. Spectral analysis is the general name applied to a large number of procedures central to which is the calculation of a spectrum by means of a Fourier transformation. This is essentially a curve fitting procedure where the coefficients of regression are explicitly labelled according to the scales of fluctuations present in the original data which may or may not be periodic. The uses to which a spectrum is put may be grouped into two classes. First, it may be used as a descriptive statistical characteristic to indicate the presence or absence of particular scales and to assess their relative importance in the data. It may confirm a priori hypotheses or suggest new ones about the generating processes giving rise to the fluctuations. Secondly, it may be used like logarithms as a means of simplifying calculations.

(3) OUTLINE OF THE ARITHMETIC

Production of the discrete spectrum is the basic step in the calculations and is dealt with separately. It involves the calculation of the coefficients, a and b for a Fourier series

$$X[j] = a[0] + \sum_{k=1}^{\frac{n-1}{2}} \left(a[k] \cos \frac{2\pi jk}{n} + b[k] \sin \frac{2\pi jk}{n} \right) + a\left[\frac{n}{2}\right] \cos \pi j, \qquad (3.0)$$

where $X[j]$ is one of the n observations. Note that if n is even, say 20, then the summation goes from $k = 1$ to 9 and if n is odd, say 19, then the summation goes from $k = 1$ to 9 and $a[n/2]$ is not included in the equation.

(3.1) THE DISCRETE SPECTRUM

The unmodified observations are fed directly into the transformation routine (section 4, which works with two sets of data at once) and the results $a_x[k]$ and $b_x[k]$ provide the basis for further calculations. Both a and b and each of the following excepting for (3.1.1) produce spectra.

(3.1.1) Frequency

The frequency k, an integer varying between 0 and $n/2$, specifies the scales calculated as number of cycles (number of complete oscillations of a cosine curve) over the data interval n. Therefore, if $n = 12$ and $k = 3$ the frequency would be 3 cycles in the interval of 12 observations. If the original observations were 2 hours apart ($\Delta t = 2$ hrs.) then this frequency could be expressed in cycles per hour by using the relationship,

$$\text{frequency}, f = k/n\Delta t, \qquad (3.1.1.1)$$

and in the above example f would equal $\frac{3}{12\times 2}=\frac{1}{8}$ cycle per hour. Alternatively, the coordinates of the spectral output may be expressed in terms of period or wavelength, the inverse of frequency. Thus with $k = 3$, $n = 12$ and $\Delta t = 2$ hrs. the period becomes 8 hrs. per cycle.

(3.1.2) Amplitude

One characteristic of a sinusoidal curve is the amplitude, half the height between the maximum and minimum. For a given frequency, k, this is related to the cosine and sine coefficients by

$$A_x[k] = (a_x^2[k]+b_x^2[k])^{1/2}. \qquad (3.1.2.1)$$

(3.1.3) Phase and Phase Shift

Another useful characteristic of the sinusoidal curve is the position of the maximum in the curve as expressed by the phase,

$$\Phi_x[k] = \arctan(b_x[k]/a_x[k]) \text{ (degrees)}. \qquad (3.1.3.1)$$

In terms of distance along the original coordinate space this may be converted to phase shift.

$$\text{Distance of maximum from origin in units of data spacing} = \frac{\Phi_x[k]\times n}{k\times 360} \qquad (3.1.3.2)$$

(3.1.4) Variance

The variance of a sinusoidal curve is related to the amplitude

$$\hat{\sigma}_x^2[k] = A_x^2[k]/2 = (a_x^2[k]+b_x^2[k])/2, \qquad (3.1.4.1)$$

excepting for $k = n/2$ (n even) when

$$\hat{\sigma}_x^2[n/2] = A_x^2[n/2] = a_x^2[n/2]. \qquad (3.1.4.2)$$

(3.1.5) Graphical Representation

Because of the trigonometric relationship

$$(a_x^2[k]+b_x^2[k])^{1/2}\cos\left(\frac{2\pi jk}{n} - \Phi_x[k]\right) = a_x\cos\frac{2\pi jk}{n} + b_x\sin\frac{2\pi jk}{n}. \qquad (3.1.5.1)$$

Equation (3.0) may be rewritten

$$X[j] = A_x[0]+\sum_{k=1}^{\frac{x-1}{2}} A_x[k]\cos\left(\frac{2\pi jk}{n} - \Phi_x[k]\right) + A_x[n/2]\cos\pi j, \qquad (3.1.5.2)$$

which says that the series $X[j]$ is the simple arithmetic sum at each $[j]$ of a finite number of cosine waves each having its own scale (frequency), amplitude and phase. As an example of one such wave, if $a_x[2] = 3$ and $b_x[2] = 4$, then, from equations (3.1.2.1) and (3.1.3.1), $A_x[2] = 5$ and $\Phi_x[2] = \arctan 4/3 = 53°8'$ (see Figure 4). It should be noted that $A[0] = a[0]$ is a constant and equal to the mean of $X[j]$.

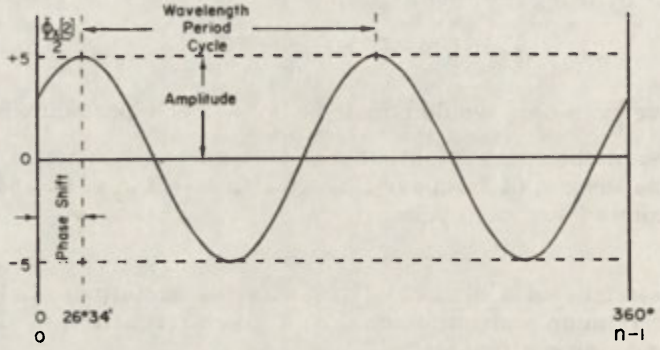


Fig. 4

(3.1.6) Aliasing

Because observations are made at discrete intervals of time or space, the calculable frequencies are limited ($k_{max} = n/2$). However, the frequencies beyond this limit may still be represented in the discrete data, and since every observation must be accounted for exactly by the calculable frequencies, the magnitude of the calculable frequencies may be significantly different from the true magnitudes. This effect is known as aliasing. Care is therefore necessary in selecting a datum spacing such that the magnitude of frequencies greater than k_{max} are relatively small.

(3.2) THE CONTINUOUS SPECTRUM

There are two different approaches to the calculation of the spectrum of non periodic data. For those with little knowledge of the subject it is suggested that they select initially the method which will most quickly give them results. The results will be comparable but the direct method has advantages of overall efficiency on large computers.

(3.2.1) The auto-covariance method

This is the older of the two methods and follows from the discussion of auto-covariances in section 2. The steps are

(a) Remove mean from the original data. X[i]

(b) Calculate the auto-covariances for approximately $n/10$ lags, i.e., maximum $j = m+1$

$$X[j] = \frac{1}{n-j+1} \sum_{i=1}^{n-j+1} X[i]X[i+j-1]. \qquad (3.2.1.1)$$

This produces $m+1$ estimates of $X[j]$

(c) Either

(i) Fit zero phase cosine curve to $X[j]$ $(1 \leqslant j \leqslant n_1+1)$

$$a[0] = \frac{1}{2m}(X[1]+X[m+1]) + \frac{1}{m}\sum_{j=2}^{m} X[j],$$

$$a[k] = \frac{1}{m}X[1] + \frac{2}{m}\sum_{j=2}^{m} X[j]\cos\frac{\pi k(j-1)}{m} + \frac{1}{m}X[m+1]\cos k\pi,$$

$$0 < k < m$$

$$a[m] = \frac{1}{2m}X[1] + (-1)^m X[m+1] + \frac{1}{m}\sum_{j=2}^{m}(-1)^{j-1}X[j], \qquad (3.2.1.2)$$

or

(ii) Increase $X[m+1]$ to $X[2m]$ by making

$$X[2m+2-j] = X[j] \qquad 1 < j \leqslant m. \qquad (3.2.1.3)$$

and apply the method in section 4 for $X[j]$. Since the function is symmetrical only a coefficients corresponding to those in (3.2.1.2) will be returned.

(d) Smooth final results to adjust for having analysed a sample of limited length from a continuum. For example, the number of terms summed in step b decreases with increasing lag. The results are $m+1$ estimates of the variance which must be assigned to bands. The frequencies of the centers of these bands are given by k. One simple smoothing procedure which may be used is $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$, known as hanning,

$$\begin{aligned}
\hat{\sigma}_x^2[0] &= 0.5\,(a_x[0] + a_x[1]),\\
\hat{\sigma}_x^2[r] &= 0.5\,a_x[k] + 0.25(a[k-1] + a[k+1]),\\
\hat{\sigma}_x^2[m] &= 0.5\,(a[m-1] + a[m]),
\end{aligned} \qquad (3.2.1.4)$$

where $r = k$.

There are no phases and the frequencies are given by $r/2m\Delta t$ in cycles per units of t. Each band is not completely independent of adjacent bands. Consequently, they share degrees of freedom of $2n/m$ each (the end bands have n/m degrees of freedom each).

(3.2.2) The direct method

(a) Remove mean from original data $x[j]$ of D observations.

(b) Smooth or taper the ends of data to account for the edge effects produced by removing a sample from the continuum. This is similar to smoothing in step d. of 3.2.1. Similarly various smoothing routines may be used. For example, a cosine bell of 6 observations would multiply the first x by 0 the second by .067, the third by .025, the fourth by .5, the fifth by .75, and the sixth by .933. The last six observations would be multiplied by the same factors in reverse order. This weighting function may be expressed as

$$h[j] = \frac{1}{2}\left(1 - \cos\frac{\pi j}{G}\right), \qquad (3.2.2.1)$$

where G is the number of observations to be smoothed. G is arbitrary but may be between $D/20$ and $D/4$.

(c) Add zeros to the end of the data to increase resolution of frequencies and to select an N which may be factored. This step produces a series $X[j]$ or N observations long.

(d) Apply transform of section 4.

(e) Calculate the variances of the elementary bands using; (3.1.3.2)

(f) Group the variances by summing the elementary bands to increase statistical confidence. For example, if $2z+1$ elementary bands are added in each group, this may be expressed as

$$\hat{\sigma}_x^2[0] = a_x^2[0] + \sum_{k=1}^{z} \frac{(a_x^2[k]+b_x^2[k])}{2},$$

$$\hat{\sigma}_x^2[r] = \sum_{k=r(2z+1)-z}^{r(2z+1)+z} \frac{(a_x^2[k]+b_x^2[k])}{2}, \quad 0 < r < m \qquad (3.2.2.2)$$

$$\hat{\sigma}_x^2[m] = \sum_{k=N/2-z}^{N/2-1} \frac{(a_x^2[k]+b_x^2[k])}{2} + a^2[N/2].$$

As in the auto-covariance method the frequencies of the centers of the bands are given by $r/2m\Delta t$. However, the bands are almost independent and the number of degrees of freedom is given by

$$\nu = \frac{(2z+1)(D-G)}{N/2} = \frac{D-G}{m}. \qquad (3.2.2.3)$$

(3.2.3) Confidence bands

The confidence bands of spectral estimates are given by the following:

$$\frac{\nu\hat{\sigma}_x^2[k]}{\chi^2} \leqslant \sigma_x[k] \leqslant \frac{\nu\sigma_x^2[k]}{\chi^2}. \qquad (3.2.3.1)$$

$$\left[\begin{matrix}\text{smaller} \\ \nu, \text{ probability level}\end{matrix}\right] \quad \left[\begin{matrix}\text{larger} \\ \nu, \text{ probability level}\end{matrix}\right]$$

Therefore, the true variance of a given band will be between degrees of freedom times the estimated variance divided by the chi square value at the 5% level and degrees of freedom times the estimated variance divided by the chi square value at the 95% level 90% of the time.

(3.2.4) Interpretation

Whereas simple periodic functions have been used in calculating the spectrum care should be taken to avoid the assumption that true periodicities necessarily exist. Any one band in the continuous spectrum represents an infinite number of frequencies centered upon the one used to identify that band. Any characteristic for that band, such as variance, is the sum of the continuous frequencies. It may well be that the actual variance fluctuates wildly (look, for example, at the elementary bands). The final estimate therefore is only an estimate of the central tendency for that band and consequently may be misleading. For instance, a band of periods of 9 to 5 days might have all its variance supplied by 9 days, yet it would be combined with and swamped by all the other periods and assigned to the central period of six days. On the other hand too much importance must not be attached to a given elementary band since it has only

two degrees of freedom. Obviously specific peaks or valleys must be taken as tentative and further analysis, by the taking of more samples or by the application of other procedures, will be necessary. The variance spectrum should be thought of as a probability function indicating the *probable* scales in the data. Also, remember that aliasing may be a source for serious error.

(3.3) SPECTRAL REGRESSION AND OTHER USES OF THE FOURIER TRANSFORM

Cross spectral analysis or spectral regression may be approached either through direct calculation of the Fourier coefficients or via a calculation of the lagged covariances. In either one the central procedure is the Fourier transform. The equations may be found in Rayner (1971).

The other uses referred to include filtering, differentiation and pattern recognition. Again the central procedure is the Fourier transform and the other details, which are outside the scope of this paper may be found in Rayner (1971).

(4) CALCULATION OF THE FOURIER COEFFICIENTS

As already indicated, regardless of whether the original data are assumed to be periodic or not the calculation of the spectrum involves the fitting of a Fourier Series (an ordered set of sinusoidal curves) by least squares. As such it may be considered similar to any other curve fitting procedure. However, because sinusoidal functions are orthogonal the arithmetic is considerably simplified and calculation of the covariance matrix is unnecessary. Also, a number of short-cut algorithms have been developed which involve the minimum in calculations. Of these, presently the most general and often the most efficient is the Fast Fourier Transform. (Gentleman and Sande, 1966). In essence this requires that the transform be computed in steps, each one involving a factor of the total number of observations. Procedures may be developed for any factor but that involving the factor 4 is the most efficient. In the present discussion the factors 4, 2, 3 and 5 will be used. Hence N, the total number of observations, is a restricted set (see table 4.1.1). In detail the Fast Fourier Transform may be set up in a number of ways. The one chosen here is one used by Sande. It allows the calculation of the transform of two series simultaneously. Also, in order to save computer storage space, the answers are stored in the locations of the original data and must be sorted at the end. The following 'hand' calculations are kept in this form so that the reader might easily convert them to a program for any size of computer.

TABLE 4.1.1. Numbers between 1 and 100 with factors 4, 2, 3 and/or 5

2	15	32	64
3	16	36	72
4	18	40	75
4	18	40	75
6	20	45	80
8	24	48	81
10	27	54	96
12	30	60	100

(4.1) THE TRANSFORMATION ALGORITHM

Input to the procedure are two equally spaced series X and Y, each with N observations which must have factors only 4, 2, 3 and/or 5. If only one series is available Y may be entered as zeros. Both X and Y have indices from 1 to N. It should be noted that this changes for the unravelling step.

The calculations are divided into steps, T, (see page 77 for the list and further explanation of symbols) each one referring to a different factor F_T where the factors are labelled 4's first followed by 2's, 3's and 5's. For example, if $N = 45$, $F_1 = 3$, $F_2 = 3$, $F_3 = 5$, and n, the total number of factors or steps, $= 3$.

For each factor a sheet has been drawn up indicating what copying, recopying and arithmetic must be performed. In any step N/F_T sheets of the same factor F_T must be used. Thus, if $N = 45$, there are required 15 sheets for factor $F_1 = 3$, 15 sheets for factor $F_2 = 3$ and 9 sheets for factor $F_3 = 5$, Each step is specified by its number T, a counter M and the value of the counter at the previous step MP, where $M = MP/F_T$ *and* MP in the first step equals N. Thus for $N = 45$ these variables will be:

Step no. T		1	2	3
Factor F_T		3	3	5
MP	$N = 45$	15	5	5
$M = MP/F_T$		15	5	1

For $N = 360$ this would read:

Step no. T		1	2	3	4	5
Factor F_T		4	2	3	3	5
MP	$N = 360$	90	45	15	5	5
$M = MP/F_T$		90	45	15	5	1

Within each step two indices J and K are used to specify which set of X's and Y's and which angles are to be used. Whereas T is the most slowly varying index, J is intermediate and K is the fastest. In each step J starts at 1 and increases by 1 to M whereas for each J, K starts at MP and increases by MP to N.

Therefore for $N = 45$ J and K take on the following values:

$T = 1,\ F_1 = 3,\ MP = N = 45,\ M = 15$

$J = 1\ \ 2\ \ 3\ \ 4\ \ 5\ \ 6\ \cdots\cdots\ 15$

$K = 45\ 45\ 45\ 45\ 45\ 45\ \cdots\cdots\ 45$

$T = 2,\ F_2 = 3,\ MP = 15,\ M = 5$

$J = 1 \qquad 2 \qquad 3 \qquad 4 \qquad 5$

$K = 15, 30, 45\ \ 15, 30, 45\ \ 15, 30, 45\ \ 15, 30, 45\ \ 15, 30, 45$

$T = 3,\ F_3 = 5,\ MP = 5,\ M = 1$

$J = 1$

$K = 5,\ 10,\ 15,\ 20,\ 25,\ 30,\ 35,\ 40,\ 45$

To prevent duplication and confusion the top right hand corner of each sheet contains sequencing information:

Step no. T of n,

Sheet no. $((J-1)\,\mathrm{N}+K)/MP$ of N/F_T.

In each step every observation in the X and Y series is processed and is replaced by an answer. These answers are used as input X and Y to the succeed-

ing step. Note that the trigonometric terms are functions of J and not of K. Therefore they remain the same in a given step for a given J and may be recopied. Also, in the final step the sine term is always zero and the cosine always one which simplifies the calculations on those sheets. In recopying for '5 AS A FACTOR' X(J0) and Y(J0) are needed twice from the previous step. Do not copy line 24 into 25.

Summary of procedure:

(1) Factor N and number the factors F_T.
(2) Calculate MP and M for each step.
(3) Find the values of J and K for each step.
(4) Label the necessary arithmetic sheets from the above information.
(5) Proceed by completing the calculations on each sheet whose order is indicated by the label.

The answers of the final step are the unsorted two sided complex coefficients.

KEY FOR SYMBOLS USED IN THE CALCULATIONS

N — the total number of observations in one series. N must be the same for two series which are to be transformed together

n — the number of factors in N

F_T — the particular factor of N for which a given sheet is designed e.g. where $F_T = 3$ the sheet will be needed "3 AS A FACTOR".

T — the step number. Is the most slowly varying index. T starts at 1 and increases by one for each separate factor

As an example of the above symbols

$$N = F_1 \cdot F_2 \cdot F_3 \cdot \ \dots \ \cdot F_T \cdot \ \dots \ \cdot F_n$$

the following magnitudes might be used:

$$N = 360,\ F_1 = 4,\ F_2 = 2,\ F_3 = 3,\ F_4 = 3,\ F_5 = 5,\ n = 5.$$

Note that the factors are ordered such that the 4's are computed first followed by 2's, 3's and 5's in that order. For sorting the 4's are ignored and only 2's, 3's and 5's are used as factors.

MP — a counter from the previous step. For the first step $MP = N$ and in succeeding steps $MP_T = M_{T-1}$ (M of previous step)

M — a counter $M = MP/F_T$

J — an index. Is intermediate between T and K. Starts at 1 and increases by 1 to M in a particular step T. Note that the cosines and sines for a fixed T and J remain constant so do not need to be recomputed as K changes

K — in index. It the most rapidly changing index. Starts at MP and increases by MP to N for a particular J.

(4.2) SORTING

The final answers are arranged in such a way that they be sorted by a set of nested indices which are controlled by the factors. It should be noted that 4 is no longer considered a factor and must be replaced by two 2's. Therefore for $N = 60$ the factors will be $F_{(i)} = 2$, $F_{(ii)} = 2$, $F_{(iii)} = 3$, $F_{(iv)} = 5$. Then the indices vary such that

I goes from H to N in steps of NF_1 $\qquad I = H, 60, 30$

H goes from G to $N/F_{(i)}$ in steps of $N/F_{(i)}F_{(ii)}$ $\qquad H = G, 30, 15$

Cont. on page 81

PROBLEM NAME _____

STEP NO _____ OF _____

SHEET NO _____ OF _____

4 AS A FACTOR

N = ____ T = ____ MP = ____ M = ____ J = ____ K = ____

TH = (J−1)360/MP = _____ COS(TH) = C1 = _____ SIN(TH) = D1 = _____

COS(2TH) = C2 = _____ SIN(2TH) = D2 = _____

COS(3TH) = C3 = _____ SIN(3TH) = D3 = _____

JO = J+K−MP = _____

J1 = JO+M = _____

J2 = J1+M = _____

J3 = J2+M = _____

	X(J0)		X(J1)		Y(J0)		Y(J1)	
	X(J2)		X(J3)		Y(J2)		Y(J3)	
ADD	A1		A2		A3		A4	
SUB	S1		S2		S3		S4	
	A1		S1		A3		S3	
	A2		S4		A4		S2	
ADD	X(J0)		A6		Y(J0)		A8	
SUB	S5		S6		S7		S8	
	S5.C2		A6.C1		S6.C3			
	S7.D2		S8.D1		A8.D3			
ADD	X(J1)		X(J2)		X(J3)			
	S7.C2		S8.C1		A8.C3			
	S5. D2		A6.D1		S6.D3			
SUB	Y(J1)		Y(J2)		Y(J3)			

PROBLEM NAME _____

STEP NO _____ OF _____

SHEET NO _____ OF _____

2 AS A FACTOR

N = _____ T = _____ M = _____ J = _____ K = _____

TH = (J−1)360/MP = _____ COS(TH) = C1 = _____ SIN(TH) = D1 = _____

J0 = J+K−MP = _____

J1 = J0+M = _____

	X(J0)		Y(J0)
	X(J1)		Y(J1)
ADD	X(J0)		Y(J0)
SUB	S1		S2
	S1.C1		S2.C1
	S2.D1		S1.D1
ADD	X(J1)	SUB	Y(J1)

PROBLEM NAME _____

STEP NO _____ OF _____

SHEET NO _____ OF _____

3 AS A FACTOR

N = _____ T = _____ MP = _____ M = _____ J = _____ K = _____

CA = COS(360/3)MP = −0.50000 DA = SIN(360/3) = 0.86603

TH = (J−1)360/MP = _____ COS(TH) = C1 = _____ SIN(TH) = D1 = _____

COS(2TH) = C2 = _____ SIN(2TH) = D2 = _____

JO = J+K−MP = _____

J1 = J0+M = _____

J2 = J1+M = _____

X(J0) _____ Y(J0) _____

X(J1) _____ Y(J1) _____

X(J2) _____ Y(J2) _____

ADD (ALL 3) X(J0) _____ Y(J0) _____

ADD (J1-J2) A1 _____ A2 _____

SUB (J1-J2) S1 _____ S2 _____

X(J0) _____ Y(J0) _____

A1.CA _____ A2.CA _____

ADD A3 _____ A4 _____

A3 _____ A4 _____

S4 _____ S3 _____

ADD A5 _____ A6 _____

SUB S6 _____ S5 _____

A5.C1 _____ S5.C2 _____

S6.D1 _____ A6.D2 _____

ADD X(J1) _____ X(J2) _____

S5.C1 _____ A6.C2 _____

A5.D1 _____ S6.D2 _____

SUB Y(J1) _____ Y(J2) _____

PROBLEM NAME _____

STEP NO _____ OF _____

SHEET NO _____ OF _____

5 AS A FACTOR

N = _____ T = _____ MP = _____ M = _____ J = _____ K = _____

CA = COS(360/5) = 0.30902 DA = SIN(360/5) = 0.95106

CA = COS(2.360/5) = −0.80902 DA = SIN(2.360/5) = 0.58779

TH = (J−1)360/MP = _____ COS(TH) = C1 = _____ SIN(TH) = D1 = _____

COS(2TH) = C2 = _____ SIN(2TH) = D2 = _____

COS(3TH) = C2 = _____ SIN(3TH) = D3 = _____

COS(4TH) = C4 = _____ SIN(4TH) = D4 = _____

J0 = J+K−MP = _____

J1 = J0+M = _____

J2 = J1+M = _____

J3 = J2+M = _____

J4 = J3+M = _____

	X(J1)		X(J2)		Y(J1)		Y(J2)	
	X(J4)		X(J3)		Y(J4)		Y(J3)	
ADD	A1		A2		A3		A4	
SUB	S1		S2		S3		S4	

	X(J0)		Y(J0)	
	A1		A3	
	A2		A4	
ADD	X(J0)		Y(J0)	

	XJ0)		X(J0)		Y(J0)		Y(J0)	
	A1.CA		A1.CB		A3.CA		A3.CB	
	A2.CB		A2.CA		A4.CB		A4.CA	
ADD	A5		A6		A7		A8	
	S1.DA		S3.DA		S1.DB		S3.DB	
	S2.DB		S4.DB		S2.DA		S4.DA	
ADD	A9		A10		S5		S6	
	A5		A6		A7		A8	
	A10		S6		A9		S5	
ADD	R1		R2		Q4		Q3	
SUB	R4		R3		Q1		Q2	
	R1.C1		R2.C2		R3.C3		R4.C4	
	Q1.D1		Q2.D2		Q3.D3		Q4.D4	
ADD	X(J1)		X(J2)		X(J3)		X(J4)	
	Q1.C.1		Q2.C2		Q3.C3		Q4.C4	
	R1.D1		R2.D2		R3.D3		R4.D4	
SUB	Y(J1)		Y(J2)		Y(J3)		Y(J4)	

G goes from F to $N/F_{(i)}F_{(ii)}$ in steps of $N/F_{(i)}F_{(ii)}F_{(iii)}$ $\quad G = F, 15, 5$
F goes from E to $N/F_{(i)}F_{(ii)}F_{(iii)}$ in steps of $N/F_{(i)}F_{(ii)}$
$F_{(iii)}F_{(iv)}$ $\quad F = E, 5, 1$
E is 1

and in full with L increasing by 1

$E = 01$
$F = 01$
$G = 01 \quad 06 \quad 11 \quad 02$
$H = 01 \quad 16 \quad 06 \quad 21 \quad 11 \quad 26 \quad 02 \quad 17$
$J = 01, 31, 16, 46, 06, 36, 21, 51, 11, 41, 26, 56, 02, 32, 17,$ etc.
$L = 01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12, 13, 14, 15,$

Then correctly labelled

$X(1)$ is in $X(1)$
$X(2)$ is in $X(31)$
$X(3)$ is in $X(16)$
etc.
$X(L)$ is in $X(I)$

For more factors the number of indices must be increased accordingly. The most slowly varying index must start at one.

(4.3) UNRAVELLING THE ODD AND EVEN PARTS

The sorted answers contain the cosine and sine coefficients of the X and Y series as even and odd functions. If $a_x[k]$, $a_y[k]$, $b_x[k]$, and $b_y[k]$ are the coefficients of the cosine and sine of functions at frequency k (an integer varying from 0 to $N-1$) then the computed series

$$X[k] = (a_x[k] + b_y[k])N/2,$$

and

$$Y[k] = (a_y[k] - b_x[k])N/2.$$

Now since the a's are even (symmetrical about $N/2$) and the b's are odd (asymmetrical about $N/2$) the four separate coefficients may be unravelled by the following:

$$\left.\begin{aligned} a_x[k] &= (X[k] + X[N-k-2])/N \\ b_x[k] &= (X[k] - X[N-k-2])/N \\ a_x[k] &= (Y[k] + Y[N-k-2])/N \\ b_x[k] &= -(Y[k] - Y[N-k-2])/N \end{aligned}\right\} \text{For } k = 1 \text{ to } \frac{N-1}{2}$$

Note that $a_x[0] = X[0]/N$, $a_y[0] = Y[0]/N$, $b_x[0] = b_y[0] = 0$ and if N is even $a_x[N/2] = X[N/2]/N$, $a_y[N/2] = Y[N/2]/N$, $b_x[N/2] = b_y[N/2] = 0$.

(5) AN EXAMPLE

For demonstration two sets of periodic data have been selected. These are observations of mean monthly temperatures at 18° intervals along latitude 60°N for the months of January and July. Observations, as listed in Table 5.1, were

interpolated from small scale atlas maps and are only approximate. With N as 20 the factors are $F_1 = 4$ and $F_2 = 5$. The necessary nine sheets of calculations are given in the following pages. Some rounding was performed so any recalculation may give small differences in the last decimal places.

TABLE 5.1. Monthly mean temperature at 60° N in °F

Lat. E	Jan.	Jul.	Lat. W	Jan.	Jul.
0	36	55	180	15	46
18	25	60	162	0	50
36	12	62	144	25	56
54	4	63	126	0	64
72	−6	64	108	−15	61
90	−9	65	90	−21	49
108	−18	63	72	−13	47
126	−32	67	54	10	47
144	−15	55	36	30	52
162	−5	55	18	38	55

The relationships between the unsorted I, sorted L and frequency indices K are given in Table 5.2. The sorted results are listed in Table 5.3 and the unravelled coefficients in Table 5.4. The equations in 3.2.1 were applied to produce the amplitudes, phase shifts, and variance components for January and July. They are given in Tables 5.5 and 5.6 but, for better appreciation of their relative magnitudes, the reader may wish to plot them as discrete spectra (Fig. 3). Positive and negative phase shifts refer to degrees longitude east and west respectively.

TABLE 5.2. Relationships between indices
I — unsorted; L — sorted; K — frequency

L	I	K	L	I	K	L	I	K	L	I	K
1	1	0	6	12	5	11	8	10	16	19	15
2	11	1	7	7	6	12	18	11	17	5	16
3	6	2	8	17	7	13	4	12	18	15	17
4	16	3	9	3	8	14	14	13	19	10	18
5	2	4	10	13	9	15	9	14	20	20	19

Most of the results are self explanatory. In winter the mean temperature for 60°N is low but the variance is nine times larger than in summer. This suggests that factors controlling the degree of variation in temperature along the latitude are much more effective in winter than in summer. Usually it is assumed that land and sea differences account for most of the variation and that their effect is reversed with the seasons. The magnitude of the effect is seldom discussed.

Further information on the often accepted notion of seasonal temperature reversal is provided by the harmonics. Clearly the relative temperature of land and sea do change over. However, it is not a symmetrical change or the phases for the two months should be different by 180°.

TABLE 5.3. Sorted results of the calculations performed on the 60° N temperature

K	$X(K)$	$Y(K)$	K	$X(K)$	$Y(K)$
0	61.	1136.	10	41.00004	14.00004
1	200.74634	59.16919	11	−51.59653	16.22517
2	223.26517	−76.90573	12	33.56717	−15.87638
3	−8.64100	140.25755	13	2.99461	11.95756
4	−20.09388	−1.92878	14	−9.78194	−21.73210
5	10.00006	−38.99998	15	−3.99993	24.99999
6	−9.46401	30.89101	16	10.78896	−40.86906
7	4.13436	−26.46730	17	0.65302	−51.37708
8	19.73781	−2.32454	18	159.98101	16.74675
9	−29.39373	4.26040	19	85.10308	−50.01520

TABLE 5.4. The Fourier coefficients

k	$a_x[k]$	$b_x[k]$	$a_y[k]$	$b_y[k]$
0	3.05000		56.80000	
1	14.29247	−5.45921	0.45769	5.78216
2	19.16230	4.68262	−3.00794	3.16420
3	−0.39939	−9.58173	4.44402	−0.46470
4	−0.46524	−1.94701	−2.13989	−1.54414
5	0.30000	3.19996	0.69999	0.69999
6	−0.96229	−2.63115	0.45794	0.01589
7	0.35644	1.92124	−0.72548	0.05698
8	2.66524	−0.67759	−0.91004	−0.69146
9	−4.04951	−0.59823	1.02427	1.11014
10	2.05000		0.70000	

TABLE 5.5. Spectral estimates for January temperatures

k	$A_x[k]$	$\Phi_x[k]/k$	$\hat{\sigma}_x^2[k]$	% σ^2
0	3.05			
1	15.30	−21	117.0	30
2	19.73	7	194.6	50
3	9.59	−31	45.9	12
4	2.00	−26	2.0	1
5	3.21	17	5.1	1
6	2.80	−19	3.9	1
7	1.95	11	1.9	0
8	2.75	−2	3.7	1
9	4.09	19	8.3	2
10	2.05	0	4.2	1
			total 386.9	

TABLE 5.6. Spectral estimates for July temperatures

k	$A_v[k]$	$\Phi_v[k]/k$	$\hat{\sigma}_v^2[k]$	% σ^2
0	56.80			
1	5.80	85	16.8	39
2	4.36	67	9.5	22
3	4.46	−2	9.9	23
4	2.63	−36	3.4	8
5	0.99	27	.4	1
6	0.45	0	.1	0
7	0.72	25	.2	1
8	1.14	−17	.6	2
9	1.51	5	1.1	3
10	−0.70	18	.4	1
			total 42.9	

In each season the first three harmonics are the most important. In January the second harmonic stands out, accounting for 50% of the variance. With maxima over Europe (7°E) and the N. Pacific (173°W) and minima over E.N. America (83°W) and W. Central Asia (97°E) this appears to describe nicely the land and sea (currents) relationship. The first harmonic (30%) just emphasizes the warmth of the Atlantic compared to the cold of eastern Asia.

In summer the first harmonic becomes more prominent (39%) centered over Asia but some 74° west of the cold pole. This may be accounted for in part by the relative warmth of the North Atlantic. The coldest part is in fact over the Canadian archipelago. The second harmonic is no more important than the third which effects it cancels out over Canada. Despite the relative increase in importance of the higher harmonics in summer their actual magnitudes are small.

Of course, the above discussion is very superficial. However, the calculations do show clear differences between summer and winter which cannot be completely accounted for by a simple seasonal reversal in the effects of land and sea. It is suggestive of new lines of enquiry which may be pursued further using spectral or other techniques. The example serves to underline the fact that spectral techniques like other quantitative techniques do not solve problems themselves. They are vehicles for testing ideas and creating new ones.

(6) COMMENT

This paper has attempted to give an introduction to Fourier techniques for the geographer who has little knowledge of the subject and little mathematical background. It should serve as a starting point for reading further and understanding some of the substantive papers based upon these techniques. A fuller exposition is given in Rayner (1971) and more advanced treatments are to be found in Blackman and Tukey (1959), Granger and Hatanaka (1964), and Jenkins and Watts (1968). Some examples of application are climatology (Horn and Bryson, 1960), economics (Hamermesh 1969), geography (Tobler 1969), geomorphology (Speight 1967), meteorology (Fiedler and Panofsky, 1970), oceanography (N.A.S., 1963), pattern recognition (Bauer *et al.*, 1967), and water pollution (Wastler 1969).

Cont. on page 92

PROBLEM NAME TEMP. 60°N JAN+JUL.
STEP NO 1 OF 2
SHEET NO 1 OF 5

4 AS A FACTOR

N = 20 T = 1 MP = 20 M = 5 J = 1 K = 20

TH = (J−1)360/MP = 0.0 COS(TH) = C1 = 1.0 SIN(TH) = D1 = 0.0
COS(2TH) = C2 = 1.0 SIN(2TH) = D2 = 0.0
COS(3TH) = C3 = 1.0 SIN(3TH) = D3 = 0.0

J0 = J+K−MP = 1
J1 = J0+M = 6
J2 = J1+M = 11
J3 = J2+M = 16

	X(J0) 36.	X(J1) −9.	Y(J0) 55.	Y(J1) 65.
	X(J2) 15.	X(J3) −21.	Y(J2) 46.	Y(J3) 49.
ADD	A1 51.	A2 −30.	A3 101.	A4 114.
SUB	S1 21.	S2 12.	S3 9.	S4 16.
	A1 51.	S1 21.	A3 101.	S3 9.
	A2 −30.	S4 16.	A4 114.	S2 12.
ADD	X(J0) 21.	A6 37.	Y(J0) 215.	A8 21.
SUB	S5 81.	S6 5.	S7 −13.	S8 −3.
	S5.C2	A6.C1	S6.C3
	S7.D2	S8.D1	A8.D3
ADD	X(J1) 81.	X(J2) 37.	X(J3) 5.
	S7.C2	S8.C1	A8.C3
	S5.D2	A6.D1	S6.D3
SUB	Y(J1) −13.	Y(J2) −3.	Y(J3) 21.

PROBLEM NAME TEMP. 60°N JAN+JUL
STEP NO 1 OF 2
SHEET NO 2 OF 5

4 AS A FACTOR

N = 20 T = 1 MP = 20 M = 5 J = 2 K = 20

TH = (J−1)360/MP = 18. COS(TH) = C1 = 0.95106 SIN(TH) = D1 = 0.30902
COS(2TH) = C2 = 0.80902 SIN(2TH) = D2 = 0.58779
COS(3TH) = C3 = 0.58779 SIN(3TH) = D3 = 0.80902

J0 = J+K−MP = 2
J1 = J0+M = 7
J2 = J1+M = 12
J3 = J2+M = 17

	X(J0) 25.	X(J1) −18.	Y(J0) 60.	Y(J1) 63.
	X(J2) 0.	X(J3) −13.	Y(J2) 50.	Y(J3) 47.
ADD	A1 25.	A2 −31.	A3 110.	A4 110.
SUB	S1 25.	S2 −5.	S3 10.	S4 16.
	A1 25.	S1 25.	A3 110.	S3 10.
	A2 −31.	S4 16.	A4 110.	S2 −5.
ADD	X(J0) −6.	A6 41.	Y(J0) 220.	A8 5.
SUB	S5 56.	S6 9.	S7 0.	S8 15.
	S5.C2 45.30495	A6.C1 38.99332	S6.C3 5.29007
	S7.D2 0.	S8.D1 4.63525	A8.D3 4.04508
ADD	X(J1) 45.30495	X(J2) 43.62857	X(J3) 9.33515
	S7.C2 0.	S8.C1 14.26585	A8.C3 2.93893
	S5.D2 32.91595	A6.D1 12.66969	S6.D3 7.28115
SUB	Y(J1) −32.91595	Y(J2) 1.59616	Y(J3) −4.34222

PROBLEM NAME TEMP 60°N JAN+JUL
STEP NO 1 OF 2
SHEET NO 3 OF 5

4 AS A FACTOR

N = 20 T = 1 MP = 20 M = 5 J = 3 K = 20

TH = (J−1)360/MP = 36.

COS(TH) = C1 =	0.80902	SIN(TH) = D1 =	0.58779
COS(2TH) = C2 =	0.30902	SIN(2TH) = D2 =	0.95106
COS(3TH) = C3 =	0.30902	SIN(3TH) = D3 =	0.95106

J0 = J+K−MP = 3
J1 = J0+M = 8
J2 = J1+M = 13
J3 = J2+M = 18

	X(J0)	12.	X(J1)	−32.	Y(J0)	62.0	Y(J1)	67.
	X(J2)	25.	X(J3)	10.	Y(J2)	56.0	Y(J3)	47.
ADD	A1	37.	A2	−22.	A3	118.0	A4	114.
SUB	S1	−13.	S2	−42.	S3	6.0	S4	20.

	A1	37.	S1	−13.	A3	118.	S3	6.
	A2	−22.	S4	20.	A4	114.	S2	−42.
ADD	X(J0)	15.	A6	7.	Y(J0)	232.	A8	−36.
SUB	S5	59.	S6	−33.	S7	4.	S8	48.

	S5.C2	18.23201	A6.C1	5.66312	S6.C3	10.19755	
	S7.D2	3.80423	S8.D1	28.21368	A8.D3	−34.23802	
ADD	X(J1)	22.03624	X(J2)	33.87680	X(J3)	−24.04047	

	S7.C2	1.23607	S8.C1	38.83281	A8.C3	11.12460	
	S5.D2	56.11232	A6.D1	4.11450	S6.D3	−31.38486	
SUB	Y(J1)	−54.87625	Y(J2)	34.71831	Y(J3)	42.50946	

PROBLEM NAME TEMP 60°N JAN+JUL
STEP NO 1 OF 2
SHEET NO 4 OF 5

4 AS A FACTOR

N = 20 T = 1 MP = 20 M = 5 J = 4 K = 20

TH = (J−1)360/MP = 54.

COS(TH) = C1 =	0.58779	SIN(TH) = D1 =	0.80902
COS(2TH) = C2 =	−0.30902	SIN(2TH) = D2 =	0.95106
COS(3TH) = C3 =	−0.95106	SIN(3TH) = D3 =	0.30902

J0 = J+K−MP = 4
J1 = J0+M = 9
J2 = J1+M = 14
J3 = J2+M = 19

	X(J0)	4.	X(J1)	−15.	Y(J0)	63.	Y(J1)	55.
	X(J2)	0.	X(J3)	30.	Y(J2)	64.	Y(J3)	52.
ADD	A1	4.	A2	15.	A3	127.	A4	107.
SUB	S1	4.	S2	−45.	S3	−1.	S4	3.

	A1	4.	S1	4.	A3	127.	S3	−1.
	A2	15.	S4	3.	A4	107.	S2	−45.
ADD	X(J0)	19.	A6	7.	Y(J0)	234.	A8	−46.
SUB	S5	−11.	S6	1.	S7	20.	S8	44.

	S5.C2	3.39918	A6.C1	4.11450	S6.C3	−0.95106	
	S7.D2	19.02112	S8.D1	35.59673	A8.D3	−14.21483	
ADD	X(J1)	22.42030	X(J2)	39.71123	X(J3)	−15.16589	

	S7.C2	−6.18033	S8.C1	25.86255	A8.C3	43.74857	
	S5.D2	−10.46162	A6.D1	5.66312	S6.D3	0.30902	
SUB	Y(J1)	4.28129	Y(J2)	20.19943	Y(J3)	43.43955	

PROBLEM NAME TEMP 60°N JAN+JUL

STEP NO 1 OF 2

SHEET NO 5 OF 5

4 AS A FACTOR

N = 20 T = 1 MP = 20 M = 5 J = 5 K = 20

TH = (J−1)360/MP = 72.

COS(TH) = C1 =	0.30902	SIN(TH) = D1 =	0.95106
COS(2TH) = C2 =	−.80902	SIN(2TH) = D2 =	0.58779
COS(3TH) = C3 =	−.80902	SIN(3TH) = D3 =	−0.58778

J0 = J+K−MP = 5
J1 = J0+M = 10
J2 = J1+M = 15
J3 = J2+M = 20

	X(J0)	−6.	X(J1)	−5.	Y(J0)	64.	Y(J1)	55.
	X(J2)	−15.	X(J3)	38.	Y(J2)	61.	Y(J3)	55.
ADD	A1	−21.	A2	33.	A3	125.	A4	110.
SUB	S1	9.	S2	−43.	S3	3.	S4	0.

	A1	−21.	S1	9.	A3	125.	S3	
	A2	33.	S4	0.	A4	110.	S2	−43.
ADD	X(J0)	12.	A6	9.	Y(J0)	235.	A8	−40.
SUB	S5	−54.	S6	9.	S7	15.	S8	46.

S5.C2	43.68689	A6.C1	2.78116	S6.C3	−7.28116
S7.D2	8.81679	S8.D1	43.74858	A8.D3	23.51137
ADD X(J1)	52.50368	X(J2)	46.52974	X(J3)	16.23021

S7.C2	−12.13525	S8.C1	14.21480	A8.C3	32.36070
S5.D2	−31.74043	A6.D1	8.55951	S6.D3	−5.29006
SUB Y(J1)	19.60518	Y(J2)	5.65529	Y(J3)	37.65076

PROBLEM NAME TEMP 60°N JAN+JUL
STEP NO 2 OF 2
SHEET NO 1 OF 4

5 AS A FACTOR

N = 20 T = 2 MP = 5 M = 1 J = 1 K = 5

CA = COS(360/5) = 0.30902 DA = SIN(360/5) = 0.95106
CB = COS(2.360/5) = −0.80902 DB = SIN(2.360/5) = 0.58779

TH = (J−1)360/MP = 0. COS(TH) = C1 = 1.0 SIN(TH) = D1 = 0.
COS(2TH) = C2 = 1.0 SIN(2TH) = D2 = 0.
COS(3TH) = C2 = 1.0 SIN(3TH) = D3 = 0.
COS(4TH) = C4 = 1.0 SIN(4TH) = D4 = 0.

J0 = J+K−MP = 1
J1 = J0+M = 2
J2 = J1+M = 3
J3 = J2+M = 4
J4 = J3+M = 5

	X(J1)	−6.	X(J2)	51.	Y(J1)	220.	Y(J2)	232.
	X(J4)	12.	X(J3)	19.	Y(J4)	235.	Y(J3)	234.
ADD	A1	6.	A2	34.	A3	455.	A4	466.
SUB	S1	−18.	S2	−4.	S3	−15.	S4	−2.

	X(J0)	21.	Y(J0)	215.
	A1	6.	A3	455.
	A2	34.	A4	466.
ADD	X(J0)	61.	Y(J0)	1136.

	X(J0)	21.	X(J0)	21.	Y(J0)	215.	Y(J0)	215.
	A1.CA	1.85410	A1.CB	−4.85410	A3.CA	140.60289	A3.CB	−368.10254
	A2.CB	−27.50636	A2.CA	10.50659	A4.CB	−337.00171	A4.CA	144.00208
ADD	A5	−4.65246	A6	26.65249	A7	−21.39892	A8	−9.10046

	S1.DA	−17.11900	S3.DA	−14.26585	S1.DB	−10.58014	S3.DB	−8.81679
	S2.DB	−2.35114	S4.DB	−1.17557	S2.DA	−3.80422	S4.DA	−1.90211
ADD	A9	−19.47014	A10	−15.44142	SUB	S5 −6.77592	S6	−6.91468

	A5	−4.65246	A6	26.65249	A7	−21.39892	A8	−9.10046
	A10	−15.44142	S6	−6.91468	A9	−19.47014	S5	−6.77592
ADD	R1	−20.09388	R2	19.73781	Q4	−40.86906	Q3	−15.87638
SUB	R4	10.78896	R3	33.56717	Q1	−1.92878	Q2	−2.32454

	R1.C1		R2.C2		R3.C3		R4.C4	
	Q1.D1		Q2.D2		Q3.D3		Q4.D4	
ADD	X(J1)	−20.09388	X(J2)	19.73781	X(J3)	33.56717	X(J4)	10.78896

	Q1.C1		Q2.C2		Q3.C3		Q4.C4	
	R1.D1		R2.D2		R3.D3		R4.D4	
SUB	Y(J1)	−1.92878	Y(J2)	−2.32454	Y(J3)	−15.87638	Y(J4)	−40.86906

PROBLEM NAME TEMP 60°N JAN+JUL

STEP NO 2 OF 2

SHEET NO 2 OF 4

5 AS A FACTOR

N = 20 T = 2 MP = 5 M = 1 J = 1 K = 10

CA = COS(360/5) = 0.30902 DA = SIN(360/5) = 0.95106

CB = COS(2.360/5) = −0.80902 DB = SIN(2.360/5) = 0.58779

TH = (J−1)360/MP = 0.

COS(TH) = C1 =	1.0	SIN(TH) = D1 =	0.
COS(2TH) = C2 =	1.0	SIN(2TH) = D2 =	0.
COS(3TH) = C2 =	1.0	SIN(3TH) = D3 =	0.
COS(4TH) = C4 =	1.0	SIN(4TH) = D4 =	0.

J0 = J+K−MP = 6

J1 = J0+M = 7

J2 = J1+M = 8

J3 = J2+M = 9

J4 = J3+M = 10

	X(J1)	45.30495	X(J2)	22.03624	Y(J1)	−32.91595	Y(J2)	−54.87625
	X(J4)	52.50368	X(J3)	22.42030	Y(J4)	19.60518	Y(J3)	4.28129
ADD	A1	97.80863	A2	44.45654	A3	−13.31077	A4	−50.59496
SUB	S1	−7.19873	S2	−0.38406	S3	−52.52113	S4	−59.15754

X(J0)	81.	Y(J0)	−13.
A1	97.80863	A3	−13.31077
A2	44.45654	A4	−50.59496
ADD X(J0)	223.26517	Y(J0)	−76.90573

	X(J0)	81.	X(J0)	81.	Y(J0)	−13.	Y(J0)	−23.
	A1.CA	30.22455	A1.CB	−79.12878	A3.CA	−4.11326	A3.CB	10.76864
	A2.CB	−35.96605	A2.CA	13.73783	A4.CB	40.93214	A4.CA	−15.63471
ADD	A5	75.25850	A6	15.60905	A7	23.81888	A8	−17.86607

	S1.DA	−6.84638	S3.DA	−49.95055	S1.DB	−4.23130	S3.DB	−30.87117
	S2.DB	−0.22575	S4.DB	−34.77196	S2.DA	−0.36527	S4.DA	−56.26216
ADD	A9	−7.07213	A10	−84.72251	SUB S5	−3.86603	S6	25.39099

	A5	75.25850	A6	15.60905	A7	23.81888	A8	−17.86607
	A10	−84.72251	S6	25.39099	A9	−7.07213	S5	−3.86603
ADD	R1	−9.46401	R2	41.00004	Q4	16.74675	Q3	−21.73210
SUB	R4	159.98101	R3	−9.78194	Q1	30.89101	Q2	14.00004

	R1.C1		R2.C2		R3.C3		R4.C3	
	Q1.D1		Q2.D2		Q3.D3		Q4.D4	
ADD	X(J1)	−9.46401	X(J2)	41.00004	X(J3)	−9.78194	X(J4)	159.98101

	Q1.C1		Q2.C2		Q3.C3		Q4.C4	
	R1.D1		R2.D2		R3.D3		R4.D4	
SUB	Y(J1)	30.89101	Y(J2)	14.00004	Y(J3)	−21.73210	Y(J4)	16.74675

PROBLEM NAME TEMP 60°N JAN+JUL

STEP NO 2 OF 2

SHEET NO 3 OF 4

5 AS A FACTOR

N = 20 T = 2 MP = 5 M = 1 J = 1 K = 15

CA = COS(360/5) = 0.30902 DA = SIN(360/5) = 0.95106

CB = COS(2.360/5) = −0.80902 DB = SIN(2.360/5) = 0.58779

TH = (J−1)360/MP = 0.0

COS(TH) = C1 =	1.0	SIN(TH) = D1 =	0.
COS(2TH) = C2 =	1.0	SIN(2TH) = D2 =	0.
COS(3TH) = C2 =	1.0	SIN(3TH) = D3 =	0.
COS(4TH) = C4 =	1.0	SIN(4TH) = D4 =	0.

J0 = J+K−MP = 11

J1 = J0+M = 12

J2 = J1+M = 13

J3 = J2+M = 14

J4 = J3+M = 15

	X(J1)	43.62857	X(J2)	33.87680	Y(J1)	1.59616	Y(J2)	34.71831
	X(J4)	46.52974	X(J3)	39.71123	Y(J4)	5.65529	Y(J3)	20.19943
ADD	A1	90.15831	A2	73.58803	A3	7.25145	A4	54.91774
SUB	S1	−2.90117	S2	−5.83443	S3	−4.05914	S4	14.51888

X(J0)	37.	Y(J0)	−3.
A1	90.15831	A3	7.25145
A2	73.58803	A4	54.91774
ADD X(J0)	200.74634	Y(J0)	59.16919

	X(J0)	37.	X(J0)	37.	Y(J0)	−3.	Y(J0)	−3.
	A1.CA	27.86046	A1.CB	−72.93953	A3.CA	2.24082	A3.CB	−5.86654
	A2.CB	−59.53392	A2.CA	22.73996	A4.CB	−44.42935	A4.CA	16.97052
ADD	A5	5.32654	A6	−13.19956	A7	−45.18853	A8	8.10398

	S1.DA	−2.75917	S3.DA	−3.86047	S1.DB	−1.70527	S3.DB	−2.38590
	S2.DB	−3.42938	S4.DB	8.53399	S2.DA	−5.54885	S4.DA	13.80827
ADD	A9	−6.18855	A10	4.67352	SUB S5	3.84358	S6	−16.19417

	A5	5.32654	A6	−13.19956	A7	−45.18853	A8	8.10398
	A10	4.67352	S6	−16.19417	A9	−6.18855	S5	3.84358
ADD	R1	10.00006	R2	−29.39373	Q4	−51.37708	Q3	11.95756
SUB	R4	0.65302	R3	2.99461	Q1	−38.99998	Q2	4.26040

	R1.C1		R2.C2		R3.C3		R4.C4	
	Q1.D1		Q2.D2		Q3.D3		Q4.D4	
ADD	X(J1)	10.00006	X(J2)	−29.39373	X(J3)	2.99461	X(J4)	0.65302

	Q1.C1		Q2.C2		Q3.C3		Q4.C4	
	R1.D1		R2.D2		R3.D3		R4.D4	
SUB	Y(J1)	−38.99998	Y(J2)	4.26040	Y(J3)	11.95756	Y(J4)	−51.37708

PROBLEM NAME TEMP 60°N JAN+JUL

STEP NO 2 OF 2

SHEET NO 4 OF 4

5 AS A FACTOR

N = 20	T = 2	MP = 5	M = 1	J = 1	K = 20

CA = COS(360/5) = 0.30902 DA = SIN(360/5) = 0.95106

CB = COS(2.360/5) = −0.80902 DB = SIN(2.360/5) = 0.58779

TH = (J−1)360/MP = 0.

COS(TH) = C1 =	1.0	SIN(TH) = D1 =	0.
COS(2TH) = C2 =	1.0	SIN(2TH) = D2 =	0.
COS(3TH) = C3 =	1.0	SIN(3TH) = D3 =	0.
COS(4TH) = C4 =	1.0	SIN(4TH) = D4 =	0.

J0 = J+K−MP =	16
J1 = J0+M =	17
J2 = J1+M =	18
J3 = J2+M =	19
J4 = J3+M =	20

	X(J1)	9.33515	X(J2)	−24.04047	Y(J1)	−4.34222	Y(J2)	42.50946
	X(J4)	16.23021	X(J3)	−15.16589	Y(J4)	37.65076	Y(J3)	43.43955
ADD	A1	25.56536	A2	−39.20636	A3	33.30854	A4	85.94901
SUB	S1	−6.89506	S2	−8.87458	S3	−41.99298	S4	−0.93009

X(J0)	5.	Y(J0)	21.
A1	25.56536	A3	33.30854
A2	−39.20636	A4	85.94901
ADD X(J0)	−8.64100	Y(J0)	140.25755

	X(J0)	5.	X(J0)	5.	Y(J0)	21.	Y(J0)	21.0
	A1.CA	7.90014	A1.CB	−20.68279	A3.CA	10.29291	A3.CB	−26.94714
	A2.CB	31.71858	A2.CA	−12.11544	A4.CB	−69.53416	A4.CA	26.55972
ADD	A5	44.61872	A6	−27.79823	A7	−38.24125	A8	20.61258

	S1.DA	−6.55759	S3.DA	−39.93767	S1.DB	−4.05282	S3.DB	−24.68286
	S2.DB	−5.21636	S4.DB	−0.54669	S2.DA	−8.44023	S4.DA	−0.88456
ADD	A9	−11.77395	A10	−40.48436	SUB S5	4.38741	S6	−23.79830

	A5	44.61872	A6	−27.79823	A7	−38.24125	A8	20.61258
	A10	−40.48436	S6	−23.79830	A9	−11.77395	S5	4.38741
ADD	R1	4.13436	R2	−51.59653	Q4	−50.01520	Q3	24.99999
SUB	R4	85.10308	R3	−3.99993	Q1	−26.46730	Q2	16.22517

	R1.C1		R2.C2		R3.C3		R4.C4	
	Q1.D1		Q2.D2		Q3.D3		Q4.D4	
ADD	X(J1)	4.13436	X(J2)	−51.59653	X(J3)	−3.99993	X(J4)	85.10308

	Q1.C1		Q2.C2		Q3.C3		Q4.C4	
	R1.D1		R2.D2		R3.D3		R4.D4	
SUB	Y(J1)	−26.46730	Y(J2)	16.22517	Y(J3)	24.99999	Y(J4)	−50.01520

A very simple example with periodic data has been used to illustrate the fast Fourier transform, an economical way of obtaining regression coefficients. The procedure would have been identical for non-periodic data had steps *a* through *c* of 3.2.2 been applied. Finally, steps *e* and *f* would have produced the variance spectrum which would have been continuous rather than discrete. On the other hand, the simple periodic coefficients are the ones required, if purely manipulative procedures are to follow, regardless of whether the data are assumed periodic or not.

A word of caution is necessary about spectral analysis and quantitative techniques in general. Research in geogprahy will pose many types of question. The researcher must select that technique or set of techniques which are best suited to the problem at hand. Variance is an important basic statistical characteristic of data but it is not always useful. Similarly for sequenced data the variance spectrum is a basic statistical function yet it is not always the best for a specific problem (see Box *et al.*, 1967).

The Ohio State University, Columbus

BIBLIOGRAPHY

Bauer, A., Fontanel, A., Grau, G., 1967, The application of optical filtering in coherent light to the study of aerial photographs of Greenland glaciers, *J. Glaciol.*, 6, 781–793.

Blackman, R. B., Tukey, J. W., 1959, *The measurement of power spectra, New York*, Dover.

Box, G. E. P., Jenkins, G. M., Bacon, D. W., 1967, Models for forecasting seasonal and non-seasonal time series, in B. Harris(ed.) *Advanced Seminar on Spectral Analysis of Time Series*, New York, Wiley, 271–311.

Fiedler, F., Panofsky, H. A., 1970, Atmospheric scales and spectral gaps, *Bull. Amer. Meteorol. Soc.*, 51, 1114–1119.

Gentleman, W. M., Sande, G., 1966, Fast Fourier transforms for fun and profit, *American Federation of Information Processing Societies, Proceedings of the 1966 Fall Computer Conference*, 563–578.

Granger, C. W., Hatanaka, M., 1964, *An analysis of economic time series*, Princeton, Princeton Univ. Press.

Hamermesh, D. S., 1969, Spectral analysis of the relation between gross employment changes and output changes, 1958–1966, *Rev. Economic Statist.*, 51, 62–69.

Jenkins, G. M., Watts, D. G., 1968, *Spectral analysis*, San Francisco, Holden Day.

N.A.S. (National Academy of Science), 1963, *Ocean wave spectra*, Englewood Cliffs, Prentice-Hall.

Rayner, J. N., 1967, Correlation between surfaces by spectral methods, *Computer Contributions*, 12, 31–37 (Kansas Geological Survey).

Rayner J. N., 1971, *An introduction to spectral analysis*, London, Pion.

Speight, J. G., 1967, Spectral analysis of meanders of some Australian rivers, in J. N. Jennings and J. A. Marbbutt (eds.) *Landform studies from Australia and New Guinea*, Canberra, Australian National Univ. Press, 48–63.

Tobler, W. R., 1969, The spectrum of U.S. 40, *Papers, Reg. Sci. Ass.*, 23, 45–52.

Wastler, T. A., 1969, *Spectral analysis*, Washington, U.S. Government Printing Office.

GEOGRAPHIA POLONICA 25, 1973

HARMONIC ANALYSIS OF URBAN SPATIAL GROWTH

PIOTR KORCELLI AND BENIAMIN KOSTRUBIEC

INTRODUCTION

Several authors have recently suggested that urban growth can be represented as a wave-like diffusion process (Blumenfeld 1954, Boyce 1966, Morrill 1968 and 1970, Korcelli 1969, 1970 and 1972). It may be assumed that this line of research will also expand in the future. Its relation to other approaches, as well as some insights it gives into the nature of the spread of urbanization, are discussed elswhere (Korcelli 1972). This form of analysis, however, beside certain advantages it offers, brings also some dangers, which can not be overemphasized. Two particular problems may be noted:

The improvement of the concepts has not been supported by an extensive body of sound, empirical evidence. This, in part, is a consequence of scarcity of adequate data, especially when large spatial and temporal series should be employed. If it persists, such a gap may eventually prevent further development of the theory.

The second problem relates to the methodology itself. While it is usually tempting to classify a phenomenon under investigation as a part of a broader system, one may loose, by doing so, some of its rather essential properties. The following citation from Beckmann (1970, p. 116) well illustrates the point:

"Although it is interesting that the same mathematical equation appears to apply to a particle, heat diffusion, and to human migration, this conclusion should not be accepted uncritically. After all, we do not seek to reproduce the well-known equations of mathematical physics but to develop models that best reflect economic conditions".

The objections of this paper are, therefore, twofold. First, we attempt to find some statistical evidence, however limited, for the aforementioned concepts of urban growth. The method applied is believed to be consistent with the theory tested. Second, we want to trace, on the basis of the data employed, some of the specific features of the urban growth process, as opposed to other spatial diffusion processes.

TRANSFORMATION OF THE BASIC DATA

This is a crucial step of the analysis, since it specifies what is meant by the urban growth wave. Both its course, speed and dimensions are dependent on the convention chosen since it is an abstract construct and there are no methods

of measuring it directly. If we now stipulate that population increase should be the phenomenon under consideration, we may define our dependent variable as:

$$Y_{ij} = \frac{p_{ij}}{P_i} \quad \begin{matrix} (i = 1,2,3, \ldots, n), \\ (j = 0,1,2, \ldots, k), \end{matrix} \tag{1}$$

where p_{ij} denotes the rate of population increase within the j-th areal unit (zone) and in the i-th time period; while P_i is the rate of population increase for the whole region. (See Fig. 1 and 2).

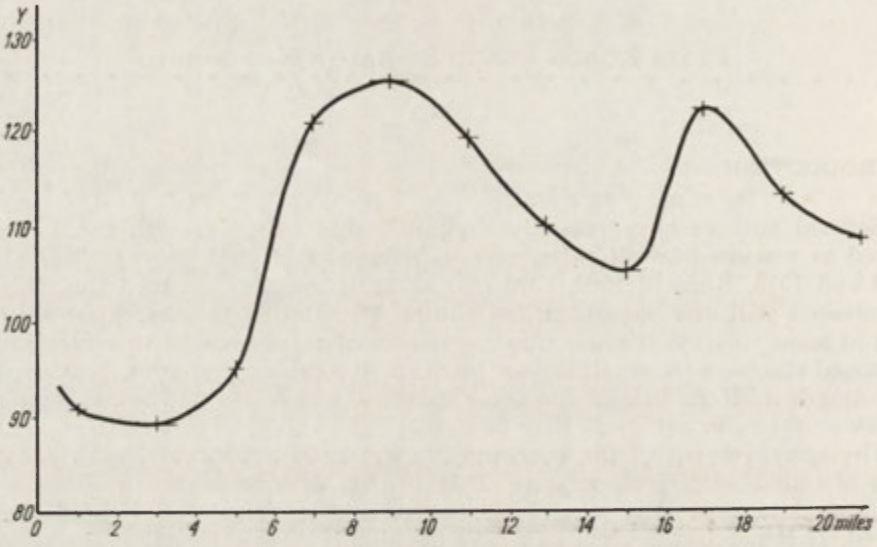


Fig. 1. Rate of population growth 1940–1950. Philadelphia region
Data source: Blumenfeld (1954)

The advantages in selecting this particular method are as follows:

(1) The dynamics of urban growth can be effectively represented for both the inner and the outer segments of the metropolitan area;

(2) By taking into account variations in the regional growth rates, the indices for various time periods are made more comparable with each other;

(3) There is a greater likelihood that more than one crest can be identified;

(4) Periodicity in the growth process, if such exists, is more easily detected.

The limitations, however, are as numerous:

(1) Because of wide variations in population density within a metropolitan area the same value of Y tends to indicate different growth cycles at different locations. At very low densities the amplitudes may assume high values even at a quite moderate absolute increase level.

(2) Negative displacements are, by definition, smaller than the positive ones.

(3) Mistreatment of the time factor. This is probably the most serious limitation of the method, since a continuous wave is being simulated by a discrete data series. Only a partial justification of this procedure is possible. We may suppose that intensity of change in the time interval for which the data are available (usually a decade) is also representative for a single point in time in the middle of this interval. In fact, a growth rate can not be identified on a cross-sectional

basis, though an approximation may be obtained if the length of time intervals is substantially contracted.

(4) The problem of areal units, which may be interpreted in a way analogous to that of temporal units. It is also commonly encountered in alternative approaches, e.g., the analysis of transformations of population density profiles.

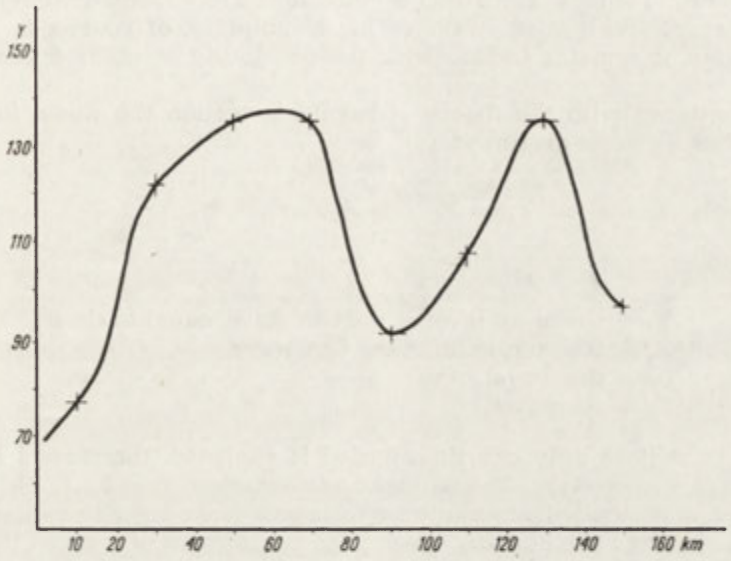


Fig. 2. Rate of population growth 1950–1960. San Francisco region
Data source: Korcelli (1969)

THE ANALYSIS

Harmonic analysis belongs to a family of statistical techniques concerned with the analysis of temporal and spatial series of data. Application of these methods to the geographical problems has been so far limited (King 1969), though quite recently a number of studies dealing with the subject have been appearing (Curry 1967, Rayner 1970, Tobler 1969 and 1970, and the articles by Rayner and Tobler in this issue). It is not difficult to foresee an expansion of this literature in the future, as application of time series analysis in other social sciences, particularly economics (Granger 1969) has proved extremely fruitful.

For the present study harmonic analysis has been performed on two spatial data series, relating to population change within the Philadelphia and San Francisco metropolitan areas. In the former case the data pertain to the time period 1940–1950, and in the latter to 1950–1960. A radially symmetric field of urban growth is assumed, thus the basic census areal units (tracts or townships) are aggregated by concentric zones. These for the Philadelphia region are 2 miles, and for the San Francisco region 20 kilometers (about 12.5 miles) in diameter.

The traditional applications of the technique have concerned cases, such as investigation of vibrations or sound waves, where the fundamental period is rather evident. It may be noted that any curve of finite length can also be described mathematically with the help of harmonic analysis, though the for-

mula employed may often involve a large number of elements. As the wave-like concept of urban growth does assume periodicity of change in the value of Y, its testing by way of harmonic analysis seems justified. According to this concept, an urban area experiences in its development the passage of several growth cycles, each cycle being composed of a definite sequence of phases. This hypothesis may be transferred to the analysis of the spatial series of data as shown in Fig. 1 and 2. Here the culminations are supposed to represent the crests of successive growth waves. This assumption, of course, is not a very safe one, but it remains in line with the previously mentioned general hypotheses.

In accordance with the theory of harmonic motion the wave, in an elastic medium, has a profile given by:

$$Y_{ij} = \overline{Y} + \sum_{j=0}^{k} A_i \sin(b_i j + c_i), \tag{2}$$

where:

$\overline{Y}_i$ = the mean level of the i-th wave, equal to sin 0°,
A_i = the amplitude of the i-th wave,
b_i = the wavelength,
c_i = the phase.

For both regions only one time period is analysed, therefore i is equal to either 1 or 2, respectively. The points of measurement ($j = 0,1,2,.., k$) are regularly spaced and situated at increasing distances from the city center, which is assumed to be the source of the waves. The coordinates of a given point on the sinusoid are:

$$Y_{ij} = A_i \sin(b_i j + c_i). \tag{3}$$

If we now denote by d_i the maximum difference between the pairs of the following numbers:

$$Y_{i0}, Y_{i1}, Y_{i2}, \ldots, Y_{i10},$$

then the amplitude of the i-th wave is equal to:

$$A_i = \frac{1}{1.90} d_i,$$

where 1.90 is the mean maximum difference between the numbers:

$$\sin(b_i j + c_i). \qquad (j = 0, 1, 2, \ldots, 10)$$

For the Philadelphia region (Fig. 1) A_i equals 18, while d_i is 35. In order to estimate the period of the sinusoid, b_i, and the phase, c_i, with the short series of data available, the least squares technique has been used. The values of b_i and c_i are selected so as to minimize the differences between:

$$c_i \text{ and } \varphi_0, \quad b_i + c_i \text{ and } \varphi_1, \quad 2b_i + c_i \text{ and } \varphi_2, \quad \ldots, \quad 10b_i + c_i \text{ and } \varphi_{10}.$$

To estimate these values we minimize the following function:

$$F(c, b) = \sum_{j=0}^{10} (bj + c - \varphi_j)^2 \tag{4}$$

and then compare its partial derivatives with respect to the parameters to zero:

$$\frac{\partial F}{\partial c} = 11c + 55b - S_1 + 0,$$

where

$$S_1 = \varphi_0, \varphi_1, \varphi_2, \ldots, \varphi_{10}$$

and

$$\frac{\partial F}{\partial b} = 385b + 55c - S_2 = 0,$$

where

$$S_2 = \varphi_1 + 2\varphi_2 + 3\varphi_3 + \ldots + 10\varphi_{10}.$$

From the above equations we obtain estimators of the wavelength b_i, and the phase, c_i

$$b = -\tfrac{1}{110}(5S_1 - S_2),$$

$$c = \tfrac{1}{22}(7S_1 - S_2).$$

The angles are found from the equation as follows:

$$\sin(b_i j + c_i) = \frac{Y_{ij} - \overline{Y}_i}{A_i}. \tag{5}$$

The arc sins of the right sides give the angles φ_j as

$$j \cdot 180° + 90° + \varphi_j$$

from which we select the series most closely approximating an arithmetic progression with a rate of increase equal to b, where $b_1 = 149°$.

The basic parameters of the population growth wave for the Philadelphia region are therefore as follows: the mean equilibrium level $\overline{Y}_1 = 107.1$; the amplitude, $A_1 = 18$, the wavelength, $b_1 = 149°$, and the phase, $c_1 = 267°$.

There is usually little likelihood that an empirical curve can be closely approximated by the fundamental term only. This is also true in case of this data set. The next step is to complicate the function by adding a second term, with different amplitude and period. The latter must be a submultiple of the fundamental period.

The basic parameters of the second harmonic term are:

$$\overline{Y}_1' = -1.9;\ A_1' = 30;\ b_1' = 25°;\ c_1' = 118°.$$

One could pursue this procedure to a point at which the theoretical curve would almost meet the empirical one. It is not, however, the purpose of the present paper. Inadequacies related to the basic data and their transformation, as well as numerous simplifying assumptions, which are built into the model, preclude a meaningful interpretation of subsequent harmonic terms. Therefore, we have to be satisfied with the approximation afforded by the fundamental and second harmonics (see Table 1).

This yields the equation:

$$Y_{ij} = \overline{Y}_i + A_i \sin(b_i j + c_i) + \overline{Y}_i' + A_i'(\sin b_i' j + c_i') + \varepsilon_j, \tag{6}$$

where ε_j is a random element

TABLE 1. Predicted and observed rates of population growth for the Philadelphia region (1949–1950)

j	Fundamental	Second harmonic	Value of Y_{1j}	
			predicted	observed
0	91.0	−0.5	90.5	90.7
1	123.9	−14.8	109.1	89.3
2	104.3	−8.0	96.3	94.9
3	107.2	10.8	118.0	121.6
4	119.8	11.0	130.8	124.1
5	92.6	27.0	119.6	118.4
6	106.5	14.2	120.7	110.1
7	95.2	22.0	117.2	105.1
8	103.3	13.0	116.3	122.7
9	125.8	−0.5	125.3	113.0
10	96.8	11.6	108.4	108.5

By substituting the numerical values we obtain:

$$Y_{1j} = 107.1 + 18\sin(149°j + 267°) + 30\sin(25°j + 118°) + \varepsilon_j \qquad (7)$$

A similar analysis has been carried out for the San Francisco region (see Fig. 2 and Table 2). As in the former case, the estimators of wavelength and phase have been found by using least squares techniques. The set of equations has the form:

$$28b + 8c - S_1 = 0,$$
$$140b + 28c - S_2 = 0.$$

TABLE 2. Predicted and observed rates of population growth for the San Francisco region (1950–1960)

j	Fundamental	Second harmonic	Value of Y_{1j}	
			predicted	observed
0	83.7	−8.9	74.8	75.0
1	131.0	−24.3	106.7	121.1
2	121.0	6.1	127.1	135.3
3	111.2	30.5	141.7	136.3
4	116.0	−21.8	94.2	91.1
5	91.2	7.9	99.1	98.9
6	139.7	−10.8	128.9	136.5
7	79.6	22.7	102.3	96.9

These yield the values of the parameters:

$$b = \frac{1}{84}(2S_2 - 7S_1),$$

$$c = \frac{1}{12}(5S_1 - S_2),$$

where

$$S_1 = \sum_{j=0}^{7} \varphi_j,$$

$$S_2 = \sum_{j=0}^{7} j\varphi_j.$$

The analysis was terminated after the second harmonic term was identified. The period of this latter term was found to be half of that of the fundamental component ($b_2 = 162°30'$ while $b_2 = 81°10'$). The values of the other parameters are:

$$A_2 = 34.52, \qquad A_2' = 28,$$
$$c_2 = 232°30', \qquad c_2' = 24°40',$$
$$\bar{Y}_2 = 111.20, \qquad \bar{Y}_2' = 2.72.$$

The solution of the equation (2) for the San Francisco region is the following (see also Table 2):

$$Y_{2j} = 113.92 + 34.52\sin(162°20'j + 232°30') - 28\sin(24°40'j + 81°10') + \varepsilon_j.$$

CONCLUSIONS

The scope of this study has been rather narrow. In fact, we have made an attempt to extend the analysis onto other time periods (back to 1900), but the variations in the parameters that have been revealed do not allow far-reaching generalizations about the behavior of urban growth waves. The one parameter which is characterized by a certain level of stability is the wavelength. Its value for the San Francisco region was about 45–50 miles in four out of seven periods, for Philadelphia it was about 10 miles in three out of five cases. It should be noted that these figures are not strictly comparable with each other, as the areal units for which the data had been aggregated were different for both regions.

The virtue of harmonic analysis lies in its ability to separate various overlapping components of a given phenomenon. It has been earlier hypothesized (Korcelli 1972) that spatial variations in the intensity of urban growth are products of several processes superimposed one upon another. These growth components are related to regular oscillations, wave asymmetry, changes in the rate of growth of the system (metropolitan area), "maturity" of an area, and, finally, to its various levels of attractiveness. We can identify our fundamental term with the basic (cyclical) growth component, but we may only suspect the second harmonic term to be associated with the wave asymmetry. Hypotheses concerning the correspondence of subsequent harmonic terms to the postulated growth components could be tested on more extensive sets of data. Theoretically, we would assume the random element in our formula to represent the spatial variations in the level of attractiveness (local resistance), but because of data limitations it is treated here as a more inclusive category.

An alternative, and possibly even more challenging approach to the analysis of urban growth waves would be to study the change in the temporal rather than in the spatial dimension. As this problem will be the subject of subsequent work, we shall not treat it here in any detail. One point should be noted. Figure 3 demonstrates that in order to be able to observe a full cycle of growth for

a given area one needs a series of data extending over at least several decades. The length of the cycle may be different for various places; one can also anticipate that at a finer areal scale faster rates of change could be recorded.

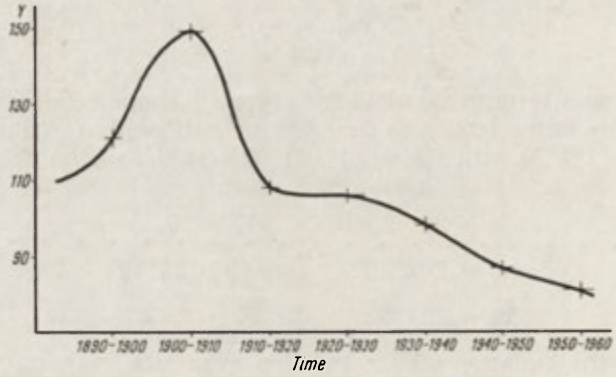


Fig. 3. Rates of population growth, 1890–1960, within a concentric zone extending 10 to 20 km from the center of San Francisco

We have attempted to demonstrate a way in which harmonic analysis can be used for predicting the intensity of population change within a metropolitan area and in a broader zone that surrounds it. Although the results are not as satisfactory as might have been anticipated, we would attribute this fact largely to inadequancies of data rather than to the approach selected. Arbitrary elements, however, pose certain questions. Some of the problems, like those relating to the units of time and space, are rather difficult to overcome, while others, like a more precise definition of the dependent variable, should be solved in future research.

Institute of Geography PAN, Warsaw
Wrocław University

BIBLIOGRAPHY

Beckmann, M., 1970, The analysis of spatial diffusion processes, *Papers, Reg. Sci. Ass.*, 25, 109–118.

Blumenfeld, H., 1954, The tidal wave of metropolitan expansion, *J. Amer. Inst. Planners*, 20, 3–14.

Bojarski, A., 1965, *Matematyka dla ekonomistów* (Mathematics for economists), PWE, Warszawa.

Boyce, R. R., 1966, The edge of the metropolis: the wave theory analog approach, *British Columbia Geogr. Ser.*, 7, 31–40.

Curry, L., 1967, Central places in the random spatial economy, *J. Reg. Sci.*, 7 (Suppl), 217–238.

Granger, C. W. J., 1969, Spatial data and time series analysis, in A. J. Scott (ed), *Studies in Regional Science*, London, 1–24.

Granger, C. W. J., 1970, *Random variables in time and space*, Paper read at the ARPUD-70 Conference, Dortmund, Oct. 5–9.

King, L. J., 1969, *Statistical analysis in geography*, Prentice-Hall, Englewood Cliffs, N. J.

Korcelli, P., 1969, *Rozwój struktury przestrzennej obszarów metropolitalnych Kalifornii* (Sum.: Evolution of spatial structure of California metropolitan areas), *Prace Geogr. IG PAN*, 78, Warszawa.

Korcelli, P., 1970, *A wave-like model of metropolitan spatial growth*, *Papers, Reg. Sci. Ass.*, 24, 127–138.

Korcelli, P., 1972, Urban spatial growth: a wave-like approach, *Geogr. Pol.*, 24, Warszawa.

Morrill, R. L., 1968, Waves of spatial diffusion, *J. Reg. Sci.*, 8, 1–18.

Morrill, R. L., 1970, The shape of diffusion in time and space, *Econ. Geogr.*, 46, 2 (Suppl.), 259–268.

Perkal, J., 1961, O pewnym schemacie cybernetycznym (On certain cybernetic scheme), *Prace Wrocł. Tow. Nauk.*, Ser. B., 103, 39–51.

Rayner, J. N., 1970, *The practical application of one dimentional spectral analysis*. Paper read at the IGU Commission on Quantitative Methods Conference, Poznań, Sept. 20–24 (revised version in this volume).

Tobler, W. R., 1969, The spectrum of. U.S. 40, *Papers, Reg. Sci. Ass.*, 23, 45–52.

Tobler, W. R., 1970, A computer movie simulating urban growth in the Detroit region, *Econ. Geogr.*, 46, 2 (Suppl.), 234–240.

GEOGRAPHIA POLONICA 25, 1973

REGIONAL ANALISIS
TIME SERIES EXTENDED TO TWO DIMENSIONS

Waldo R. Tobler

For the last several years an effort has been made to explore the relevance of an extension of time series analysis to the study of geographical problems. To date this has resulted in several published papers, with several more in preparation, and the introduction of a course entitled "Regional Analysis" at the University of Michigan during the fall semester of 1969. It would be a fairly accurate description of this course to take the twelve chapter outlines from H. T. Davis' *The Analysis of Economic Time Series* and to recast them all into a two-dimensional framework, modifying the content to emphasize the geographical distribution of population and of poverty. From a mathematical point of view the extension to two dimensions is a fairly natural generalization with only a few really interesting aspects. The formal treatment requires only that the students be familiar with probability theory, complex variables, some linear algebra, integral calculus, and so on. Our concern has been more with the substantive geographical interpretations, and social applications, of the concepts rather than the mathematical details. The suggestion that it might be useful to extend to the spatial case those methods used for the study of time series is of course not at all new.

Historically the study of time series has two sources. One is from the sciences, both social and natural, sources related to those which gave rise to the field of statistics. The other source is from telegraphy and electrical transmission studies. Norbert Wiener was able to combine these two sources and the hybrid has shown the expected vigor. The reason that time series studies differ from other studies is of course that the observations are not independent of each other. The central hypothesis of the historian is that the present is related to the past. A similar situation holds in the case of spatial series. The central dogma in geography asserts that what happens at one place is not independent of what happens at another. The most extensive development of the two-dimensional variants of time series analysis which are of interest to geographers are in geophysics, including seismology, meteorology and oceanography, and in picture processing with the related fields of visual perception, pattern recognition, optics and holography. The challenge is to demonstrate the validity, if any, of concepts from these fields to studies of society. Only a small number of people seem to be following these lines.

In the first instance consider a geographical distribution to be completely static, an atemporal or purely spatial point of view. Probably the simplest approach is to consider only one spatial dimension, e.g., the distribution of population along highway 40 from Baltimore to San Francisco. This distribution can be, and in fact has been, studied and analyzed by time series methods. The motivation comes from central place theory which suggests a repetitiveness

in the distribution of populated places throughout the landscape. Extending the analysis to examine the two-dimensional distribution of population is relatively simple. One of the interesting results, an empirical observation, is the near rotational symmetry of the bivariate autocorrelation function, which suggests that an isotropic assumption may be invoked without severe violence to the actual geographical processes.

This example is used to illustrate an important transfer of concepts from the temporal to the spatial domain. The spatial sampling along US 40 in the study cited resulted in one observation every mile of the 2900 mile long route. In time series it is usually required that one have a long record of equally spaced observations. What is a long spatial record? If the earth's circumference is assumed to be 40,000 km, one-fourth of which is land, then one might obtain a record of 10,000 observations, sampling every kilometer. This would be considered a long record in economics, and a reasonable record in electronics. The two-dimensional case is more complicated. Here it is not only the *size* of the sampling interval, but also its *shape* and *orientation* which affect the resolution of the data. This suggests the use of interval independent measures. Instead, most demographic data are made available on a rather absurd basis, by countries, counties, or census tracts, etc. We now have had experience in converting these data into "square" units, e.g., 1.5-mile squares for a 90-by-90-mile region including Detroit, or five-degree quadrilaterals for world population data. Our experience is that it is easier than one might expect. We do not do this simply to satisfy a Teutonic sense of orderliness, but because it allows a greater analysis capability. For example, if one has demographic data on a square lattice, then a rather obvious interpretation of the notion of "population pressure" is to compute the finite difference approximation to the spatial gradient. The vector field calculated in this manner does in fact define the edges of cities rather well. One could object that the objects of interest are coordinate invariants, which is true, but the practical advantages of uniform spatial data intervals seem great. The sampling theorem would also appear to have some relevance to the study of spatial distributions. Statistics books which deal with methods of demographic sampling do not mention this theorem, a rather curious omission. Presumably this is due to the fact that most demographers are interested in the frequency distributions, and not the geographical distributions, of their data.

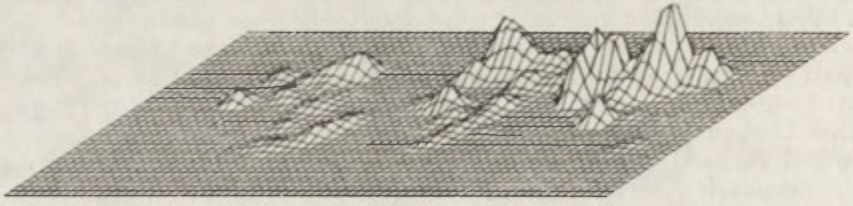


Fig. 1. World population by 5 degree quadrilaterals
Department of Geography, University of Michigan 15:55.20; Oct. 28, 1969

Spatial sampling is thus one area in which the comparison of time series analysis with regional analysis provides some insight. Another useful idea, or set of ideas, is contained in the notion of a trend. The data in these cases is decomposed into a set of components, and these may be given a spatial interpretation: Regional trends, National trends, local trends. A national trend might be from the Northwest to the Southeast, for example. With the advent of

digital computers these methods have enjoyed great popularity and many formal methods have evolved. These include the fitting of bivariate functions, smoothing, and so on. These methods can be extended to treat the entire earth using spherical harmonics.

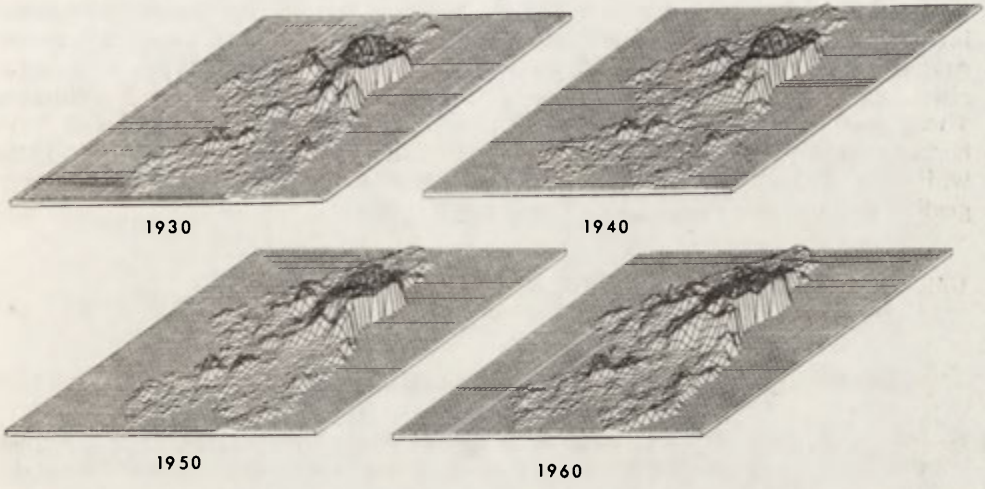


Fig. 2. Actual population growth, Detroit, Region
Non-linear vertical scale

Since the spherical shape of the earth has now been mentioned it should be pointed out that all geographical distributions are truly periodic, doubly periodic in fact, but in a rather trivial sense. The topology of the earth has other consequences; for example, any vector valued function on such a surface must have at least one singularity, there are no squares, the distinction between interpolation and extrapolation is less clear and so on. These topological facts do not hold for all two-dimensional spaces of course.

At a somewhat different level, historians divide history into periods, and so do geologists. This process may be referred to "epochodization." Geographers do the two-dimensional equivalent when they partition space into disjoint sets. This is known as "regionalization." One interpretation of these procedures is that they are attempts to find domains within which the phenomena of concern are stationary. But the subject does not appear to have been approached from this point of view, and it might be fruitful to do so.

Now consider both space and time, a spatio-temporal series. The notion of temporal lags is quite familiar to model builders. It is also obvious that most systems involve spatial lags, which are not one-sided but three or more sided. Typical forecasting procedures are weighted such that the recent past has more influence than the far past. The central dogma of geography asserts that a similar decay in space is involved. A space-time cone of recent and nearby events is relevant. A technical question which then arises concerns the compatibility between the spatial and temporal observation scales for data collection purposes.

The use of positionally invariant linear operators has been of great utility in the study of electrical signals over time. These same procedures may be used to compare geographical distributions at two different time periods; the mathematics are similar to those used in optics. Not all people behave the

same so that one would not expect a linear space invariant model to be very effective. It is however quite easy to formulate a migration model in these terms — essentially a finite difference form of the diffusion equation — and this in fact fits the empirical data quite well. Scientifically it is an oversimplification but it provides a useful null model. In the long run it will be necessary to develop multivariate spatio-temporal analyses for the transportation system influences the population distribution, and vice versa. Cross-correlation between spatial series provides a formal method for comparing some types of geographical maps, and these ideas have been extended to cross-spectral estimates. These methods appear quite complicated when examined in detail but have become more practicable with the advent of the Fast Fourier transform. Thus, while regional analysis is not a panacea, some of the approaches may be suggestive and provide insight for demographic studies.

University of Michigan, Ann Arbor

BIBLIOGRAPHY

Tobler, W. R., 1966, Spectral analysis of spatial series, *Proceedings*, Fourth Annual Conference on Urban Planning Information Systems and Programs, University of California, Berkeley, 179–186.

Tobler, W. R., 1969, The spectrum of US 40, *Papers, Reg. Sci. Ass.*, 23, 45–52.

Tobler, W. R., 1969, Geographical filters and their inverses, *Geogr. Analysis*, 1,3, 234–253.

Tobler, W. R., 1970, A computer movie simulating urban growth in the Detroit Region, *Econ. Geogr.*, 46,2 (Suppl.), 234–240.

Tobler, W. R. and Barton, B., 1971, A spectral analysis of innovation diffusion, *Geogr. Analysis* 3, 182—186).

Tobler, W. R., and Moellering, H., 1972. The analysis of scale-variance *Geogr. Analysis*, 4, 34—50.

GEOGRAPHIA POLONICA 25, 1973

SOME ASPECTS OF NETWORK THEORY

Nurudeen Alao

In so far as one may adequately characterize network theory as a body of facts about the structure of points and their interconnecting links, one may justly argue that every aspect of network theory is geographically relevant. Indeed, the virtual universal applicability of results from network theory is attested to by the importance accorded the theory in vastly divergent subjects such as neurology, psychology, sociology, economics, geography, electrical engineering and even history (see Pitts (1965)). Unfortunately the less trivial aspects of this theory (many of which are important in geography) are accessible mainly to the mathematically sophisticated. There are at least two conditions which render the less trivial aspects relatively inaccessible. The first has to do with the vast combinatorial problems involved. Thus several transport network problems are deceptively simple to enunciate but would require the examination of a vast number of cases to solve completely. The second condition relates to the difficulty of obtaining analytic solutions and consequently the necessity for providing efficient algorithms. Consequently, to assert that every aspect of network theory has geographic relevance is to suggest that geographers do have lots of theoretical problems to contend with for a long time to come.

The account presented in the following pages is not intended to cover all applications of network theory in geography. Neither is it designed to survey all network techniques at present available. Rather, it identifies some significant aspects of network theory and techniques which have recently received attention in geography or which are potentially useful in tackling some geographic problems. The approach is expository although wherever extra gains are derivable from rigorous treatment, we do not hesitate to be rigorous. Experts in this area may find that sometimes many of the steps included could be omitted. We have however insisted on including many elementary steps in order to make the exposition as self-contained as possible. Our aim, we reiterate is to select major problems and techniques in this area and treat them as simply and as efficiently as possible. In this regard, a word of caution is in order. In some of the problems dealt with here the geometric approach is employed. Because of its pedagogical advantages, the geometric approach is usually appealing. However, geometric intuition can quite often be misleading as is evident in the geometric "demonstration" of the nonsense that every plane triangle is isosceles. Consequently, whenever we appeal to geometric intuition in any construction, we shall usually take pains to prove that the construction makes sense.

Our account is divided into four sections. Section I is concerned with notation, general definitions and general descriptive techniques used on networks.

It sets forth essential graph theoretic notions which are used in comparing networks. Section I is very brief in its discussion of the use of major indices; it places great emphasis on their interpretation. Section II deals with the efficient geometric structure of networks in the two dimensional Euclidean space E^2; it unifies many problems in this area via the Reflection Principle and its generalization on the one hand, and recent mathematical formulations of the Steiner problem by Melzak (1961), DeMar (1968), Gilbert and Pollak (1968) and Cockayne (1967), on the other. Section III is devoted to the analysis of network flow problems of which two broad types are considered. First are the direct link "many sources-many destinations" flow problems and next, the transshipment types. This section reviews the analyses of flow problems and provides interpretation of the key steps in such analyses. Section IV combines Sections I–III through the consideration of networks in which construction and flow costs are incorporated into the minimand. In geography, the major contribution towards solutions of the problems of Section IV are due largely to Garrison and Marble (1958, 1965) and Werner (1968).

Throughout this paper examples will be drawn mainly from human geography. The more important applications in physical geography (most hydrology and geomorphology) have been reviewed in a recent book by Haggett and Chorley (1969).

SECTION I

In this introductory section we shall deal with the basic descriptive techniques for networks and set out general notations and definitions which apply to other sections as well. Most of the descriptive notions that have been applied to networks belong in graph theory. A major reason for setting out definitions clearly is that in network theory different authors use various terms differently with the result that only a few terms can be called standard.

NOTATION AND DEFINITIONS

D I. i A *network* N consists of a set of points (that will variously be referred to as nodes, vertices, sources) together with lines joining them. The points of the network are denoted $\{A_i\}_{i \in I}$ for an appropriate index set I.

D I.ii A *link* in a network is a line joining any two points in the network. A link is also called an *arc* or an *edge*. A link between A_i and A_j is denoted variously as $\overline{A_iA_j}$, (A_i, A_j) while the Euclidean length of such a link is denoted $|\overline{A_iA_j}|$. If the link is directed we write $A_i \rightarrow A_j$ to indicate that flow is feasible in the direction from A_i to A_j and if flow is possible in both directions we write the link as $\overline{A_iA_j}$ or more emphatically $A_i \leftrightarrows A_j$. We call N a directed network in case its links are directed, and we denote it $\vec{N}$.

D. I.iii Let A_i, A_j be any two points of N. By a chain $C_{A_i}^{A_j}$ between the points A_i and A_j, we mean a sequence of consecutive links, $\overline{A_iA_{i_1}}$, $\overline{A_{i_1}A_2}$, ..., $\overline{A_{i_m}A_j}$, the first member of which has A_i as the first of its defining points and the last member of which has A_j as the second of its defining points.

D I.iv A path $P_{A_i}^{A_j}$ between A_i and A_j in N is a chain through which a flow from A_i to A_j is realizable. The order of the path is the number of links in the corresponding chain. $P_{A_j}^{A_i}[k]$ is a path of order k linking A_i to A_j. A_i would be referred to as the initial point and A_j as the terminal point of the path.

D I.v A *circuit* is a path whose initial and terminal points are identical. A circuit defined by a path of order one is called a *loop*.

There are at least three important levels at which network descriptive techniques may be applied. The first level concerns the use of graph theoretic measures of a network to summarize its salient properties. At another level, numerical indices may be constructed from the topological properties of a network to facilitate comparison with other networks. At a third level one may use some of the descriptive techniques for the evaluation of the performance characteristics of the network. In order fully to delineate the advantages and problems associated with the applications of these measures and indices we need a few more definitions and concepts.

D I.vi Henceforth we shall regard networks and graphs as synonymous.

D I.vii Let $A = \{A_i | i \in I\}$ denote the set of given vertices of N, indexed by the finite set I. Then N is said to be *planar* if it can be represented on the two dimensional plane. More precisely N is *planar* if for i, j, k, l in I, $A_iA_j \cap \overline{A_kAl}$ belongs in A whenever this intersection is non-empty.

N is *non-planar* if for any i, j, k, l in I the following are realizable:

$$(a)\ A_iA_j \cap A_kA_l = \emptyset,$$

$$(b)\ A_iA_j \cap A_kA_l \in A,$$

$$(c)\ \emptyset \neq A_iA_j \cap A_kA_q \in A.$$

The simple examples shown in Figs. 1 and 2 are quite instructive. In the first example we have a network consisting of five vertices represented by Fig. 1a and its realization on the two dimensional plane is shown in Fig. 1b.

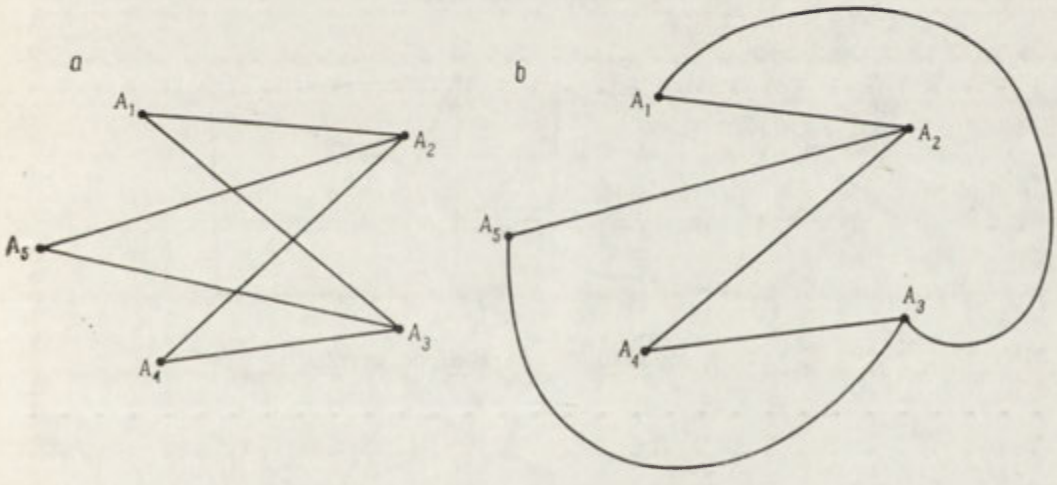


Fig. 1. Examples of planar graphs

Note that in Fig. 1b the intersection of every pair of links (if non-void) occurs in one of the given vertices and that Figs. 1a and 1b are isomorphic (D I.viii). On the other hand, Fig. 2a and Fig. 2b illustrate the case of a *non-planar* graph, since at least one new vertex A^* results from the intersection of pairs of links in the network.

D I.viii Two graphs N_1 and N_2 are said to be isomorphic if and only if

(a) there exists a one-one correspondence (i) between the vertices of N_1 and the vertices of N_2 and (ii) between the links of N_1 and the links of N_2.

(b) such one-one correspondence preserves neighborhood (incidence) properties. (See Fig. 3a and Fig. 3b).

D I.ix Consider a planar graph isomorphic with a circuit of order $n \geqslant 4$, in which all vertices are directly linked. For many purposes, it is convenient to operate on a convex polygon equivalent of such a circuit (see Fig. 4). We shall refer to such a polygon as the c.p. isomorph of the graph N. A link such as A_1A_3

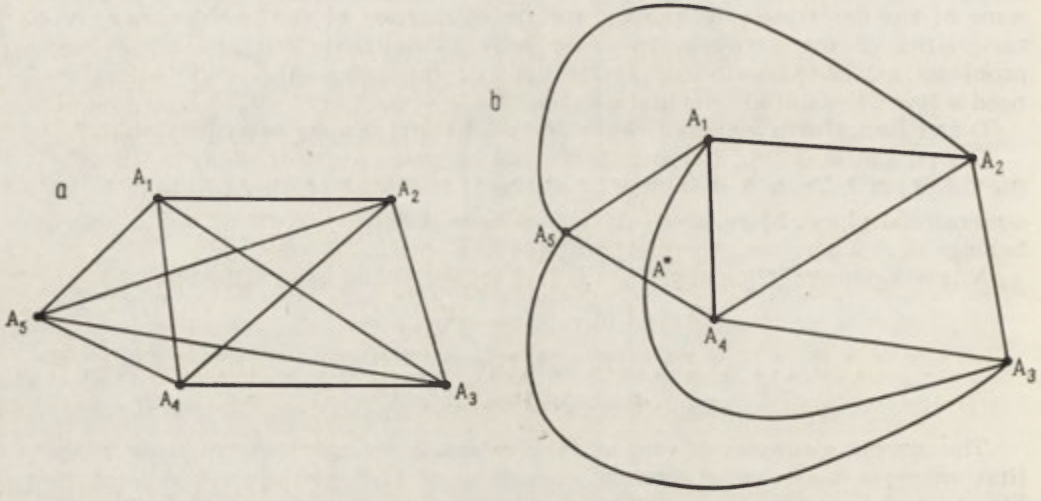


Fig. 2. Non-planar graphs

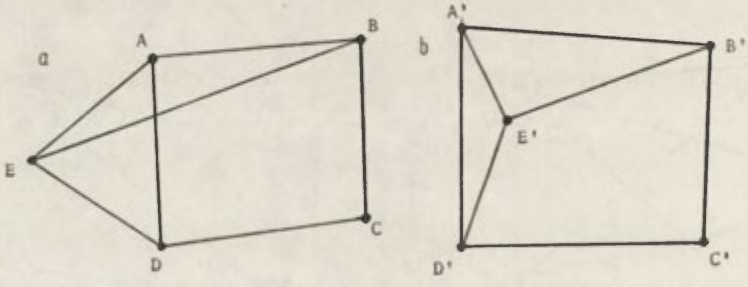


Fig. 3. An example of isomorphic graphs

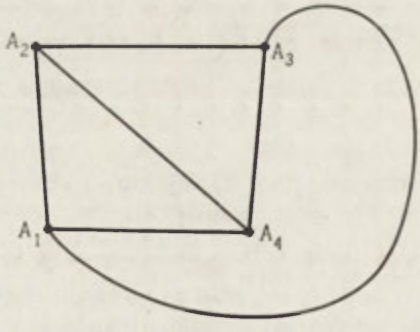


Fig. 4.

will subsequently be referred to as an *exterior link*, while a link such as $\overline{A_2A_4}$ will be referred to as an interior link. Any two links in a graph will be called non-intersecting if either they have no point in common or the only point they have in common belongs in A.

SOME ELEMENTARY IDENTITIES

ID 1. If N is *non-planar*, there are $\binom{n}{2}$ pair-wise combinations of its given points. Hence N can have no more than $\frac{1}{2}n(n-1)$ links.

ID 2. If N is *planar* and contains n given points, then N can have no more than $3(n-2)$ links.

DESCRIPTIVE INDICES AND MEASURES

It will become apparent after we have defined the various indices that have hitherto been used in geography that for any given network these are easy to compute. Consequently, rather than give detailed examples of how to compute these indices, we shall merely refer to the various literature in which they have been used and summarize their results. However, we shall discuss in some detail what seems to be more important, and what has up to now not been adequately dealt with in the literature: the problem of interpreting these indices.

For any network, there exists a variety of interesting indices which relate the number of vertices of the network to the number of links. Of these indices only four (Cyclomatic number, Alpha index, Beta index and Gamma index) have been accorded the widest application in geography.

Cyclomatic number: Suppose we have a network N determined by n given nodes and e actual number of edges. Then the cyclomatic number μ of N is defined as

$$\mu = e-n+1, \qquad \text{if the network is connected,}$$

$$\mu = e-n+p, \qquad \text{if the network can be broken into } p \text{ subsets each of which is connected.}$$

Since the minimum number of links required to connect all n points is $n-1$, $e-n+1$ gives the number of links which are in excess of the minimum required. The same interpretation (mutatis mutandis) holds for $e-n+p$. Consequently, μ may be interpreted as the degree of circuity in the network.

Alpha index: Let N be a network with n vertices. The maximum circuity in N is equivalent to the difference between the maximum realizable number of links and the number of links in the smallest tree connecting all n vertices of N. That is, the maximum circuity is equal to

$$\text{(i)}\ 3(n-2)-(n-1) = 2n-5, \qquad \text{if } N \text{ is planar,}$$

or

$$\text{(ii)}\ \frac{n}{2}(n-1)-(n-1) = \frac{(n-1)(n-2)}{2}, \qquad \text{if } N \text{ is non-planar.}$$

The Alpha index is then defined as:

$$\text{(i)}\ \alpha = \mu/2n-5, \qquad \text{if } N \text{ is planar,}$$

or

$$\text{(ii)}\ \alpha = 2\mu/(n-1)(n-2), \qquad \text{if } N \text{ is non-planar,}$$

where μ is the cyclomatic number. Clearly α has its minimum when the circuity μ is smallest, i.e., when $\mu=0$ and has its maximum when μ is largest, i.e., when $\mu=2n-5$ or $\mu=\frac{1}{2}(n-1)(n-2)$. Thus $0 \leqslant \alpha \leqslant 1$.

Beta index: The Beta index relates the actual number of links to the total number of vertices and is written as

$$\beta = e/n.$$

Clearly $\beta=0$ for a completely disconnected network and $\beta=1$ if the network is a circuit of order n.

Gamma index: The Gamma index measures the ratio of the actual number of links to the maximum possible. It is defined as

$$\gamma = e/3(n-2), \qquad \text{if } N \text{ is planar,}$$

$$\gamma = 2e/n(n-1), \qquad \text{if } N \text{ is non-planar.}$$

For examples see Figs. 5a and 5b.

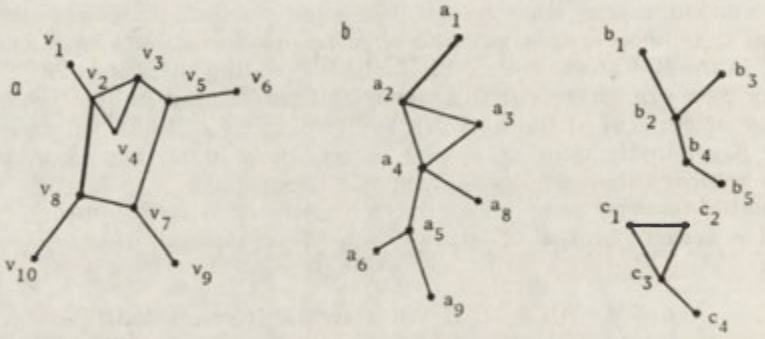


Fig. 5. Examples illustrating computation of (planar) network indices

(a) $\mu = 11-9+1 = 3$
$\alpha = 3/(2 \cdot 10-5) = 1/5$
$\beta = 11/10$
$\gamma = 11/3(10-2) = 11/24$

(b) $\mu = 8+4+4-(8-5-4)+3 = 20$
$\alpha = 2/2n-5 = 2/34-5 = 2/29$
$\beta = 16/17$
$\gamma = 16/48 = 1/3$

We shall next briefly consider the following questions:

(1) How have these graph theoretic indices been used empirically?
(2) How can they be correctly interpreted?
(3) To what other geographical problems can they be applied and how?

Probably the most thorough applications of these indices are represented in the works by Garrison and Marble (1965) and Kansky (1963), who have applied these indices on a global scale. Kansky postulated that a consistent relationship exists between several of the graph theoretic indices and levels of economic development of various countries. Kansky (1963) related selected graph theoretic indices (for railroads and highways) by a linear regression technique to selected components of the regional economy of twenty-five countries of the world. He reported a "strong relation" between the two variables at the country level, but a weaker relation at the lower regional level.

In evaluating and interpreting the measures themselves, the first thing to note is that each of them is in general a many-to-one function. This fact has both advantages and disadvantages. Assuming that each of the measures has

well-established interpretations (i.e., each of the measures has an unambiguous meaning) then their many-to-one property yields few exhaustive groups of network types for detailed theoretical study. In the absence of such interpretations, the many-to-oneness can hardly be meaningfully (i.e., scientifically) exploited. Furthermore, on a comparative basis, Werner *et al.* (1968), using eight regular types of networks labelled *A*, *B*, ..., *H*, have found that "all three measures rank the networks in the same sequence, i.e., the relationship $\sigma(I) \lesseqgtr \sigma(J)$ is the same for α, β, γ and $I, J \in A, \{B, ..., H\}$."

At an empirical level, Garrison and Marble (1965) posed the following questions: "Can the structure of transportation systems be related to the features of the areas within which they are located?" They answered this question by establishing a series of regression equations between level of socioeconomic development and the physical features of an area (as independent variables) on the one hand and the various indices as dependent variables on the other. Whilst they concluded that the fit they obtained was generally good, they warned that "the ultimate answer to the question requires generating the actual transportation network, given the characteristics of the area that contain the network." Thus we necessarily reach the conclusion that this goal is unattainable through the indices defined above principally because of their many-to-oneness.

The interdependence among the various measures has led Garrison and Marble to employ a principal component analysis (with data on 22 nations) and they obtained three principal dimensions: the first dimension is a combination of number of nodes, number of edges and the cyclomatic number; the second is a combination of the α — and γ — indices, combines a number of nodes, a number of edges and the network diameter.

In concluding this section, it is important to stress two points. First, the empirical applications of the various graph theoretic measures are largely still at the initial stages. Powerful inductive techniques still have to be brought to bear on these indices to elevate their interpretations to more general levels which enable interregional comparisons to be made. Themes that may profitably be pursued have been enunciated by Garrison and Marble (1965), Kansky (1963) and Werner *et al.* (1968). However, there is at present no body of comprehensive general statements that could serve as indicators of expectations against which the empirical observations derived from the applications of these indices could be tested.

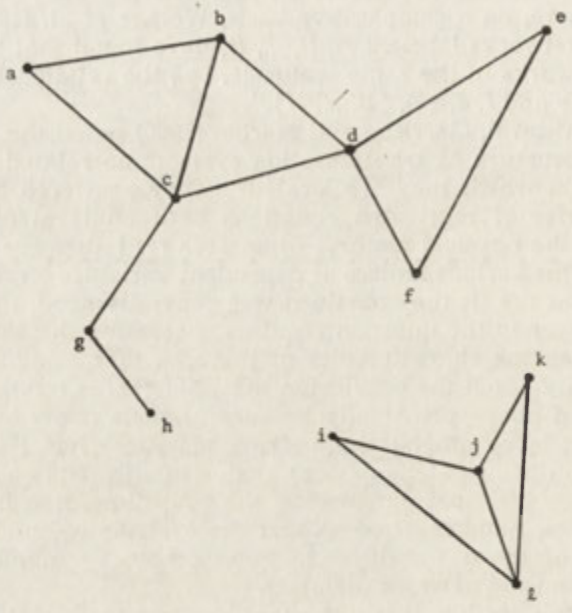
Secondly, there are several questions which the geographic literature has put to many of these indices, but which these indices can never answer meaningfully. In other instances, several mathematical operations have been performed on these indices in fashions which should make one raise questions about the validity of the operations, in the circumstances in which they are used. For example, what does it mean (for empirical purposes) to add these indices, or to take linear combinations of them? All of these point to our earlier remark that we are far from a full understanding of these indices vis-a-vis network structure and network performance.

NETWORKS AS MATRICES

To every oriented network *N* there corresponds a square matrix (linear transformation) called the *adjacency matrix* or *connectivity matrix* whose elements α_{ij} are defined as follows:

$$\alpha_{ij} = \begin{cases} 1 & \text{if the point } A_i \in N \text{ is linked to the point } A_j, \\ 0 & \text{if } A_i \text{ is not connected to } A_j. \end{cases}$$

We remark that the connectivity matrix Q defines the topology of the network *N*. Thus any statement about the topology of a network is a statement about its connectivity matrix. (See Fig. 6).



	a	b	c	d	e	f	g	h	i	j	k	ℓ
a	0	1	1	0	0	0	0	0	0	0	0	0
b	1	0	1	1	0	0	0	0	0	0	0	0
c	1	1	0	1	0	0	1	0	0	0	0	0
d	0	1	1	0	1	1	0	0	0	0	0	0
e	0	0	0	1	0	1	0	0	0	0	0	0
f	0	0	0	1	1	0	0	0	0	0	0	0
g	0	0	1	0	0	0	0	1	0	0	0	0
h	0	0	0	0	0	0	1	0	0	0	0	0
i	0	0	0	0	0	0	0	0	0	1	0	1
j	0	0	0	0	0	0	0	0	0	0	1	1
k	0	0	0	0	0	0	0	0	0	1	0	1
ℓ	0	0	0	0	0	0	0	0	1	1	1	0

Fig. 6.
Connectivity matrix associated with Fig. 6

Although several useful properties can be extracted from connectivity matrices, the mathematical notions by which such extraction is accomplished are fairly deep and advanced (see Alao, 1970). Consequently, it is not surprising that most applications of the connectivity matrix in geography have been based on the result of direct matrix multiplication. We therefore proceed to exhibit the important properties associated with such multiplication and the application to geographic problems.

D I.x Pick any points A_i, A_j in N. The product

$$(*) \qquad (a_{iv_1}a_{v_1v_2}a_{v_2v_3}\dots a_{v_{k-1}v_k} = j) = a_{iv_1}\prod_{h-1}^{k-1} a_{v_hv_{h+1}}, \text{ with } v_k = j,$$

is a path of order k in N in case this product has the value 1. In case the above product is zero, we shall say that there exists no path between A_i and $A_j \in N$ through the $(k-1)$ points (vertices or nodes) $A_{v_1}, A_{v_2}, \dots, A_{v_{k-1}}$.

Consider the matrix multiplication rule applied to the adjacency matrix Q. Let a_{ij} denote the (i, j)-th element of $Q\cdot Q$ or Q^2. Then

$$(**) \qquad \xi_{ij} = \sum_{k=1}^{n} a_{ik}a_{kj}.$$

Let $Z^1(i, j) = \{t \mid a_{it}a_{tj} = 1\}$. If $v_h \in Z^1(i, j)$ then there is a path from the point $A_i \in N$, through the point $A_{v_h} \in N$ to the terminal $A_j \in N$. Thus ξ_{ij} is the number of elements in (i.e., the cardinality of) the finite set $Z^1(i, j)$. Consequently, we arrive at the interpretation of the elements of Q^2 as the number of ways of connecting pairs of points in N through two links. Here lies the foundation of the combinatorial application (and implication) of the adjacency matrix.

Again let η_{ij} denote the (ij) element of $Q^3 = Q\cdot Q^2$. Then

$$\eta_{ij} = \sum_{k=1}^{n} a_{ik}\xi_{kj}, \text{ where } \xi_{kj} \text{ is the cardinality of } Z^1(k, j)$$

$$= \sum_{k=1}^{n} a_{ik}\sum_{h=1}^{n} a_{kh}a_{hj}$$

$$= \sum_{k=1}^{n}\sum_{h=1}^{n} a_{ik}a_{kh}a_{hj}$$

which is a sum of products of the form (*) above. Let

$$Z^2(i, j) = \left\{\{v_h\}_{h=1}^{2} \middle| (a_{iv_1})\prod_{h=1}^{2} a_{v_hv_{h+1}} = 1, \quad \text{where } v_3 = j\right\}.$$

Note that $Z^2(i, j)$ is a set of sequences. Now it becomes clear that η_{ij} is the cardinality of $Z^2(i, j)$ and we can thus logically interpret η_{ij} as the total number of ways of linking the point $A_i \in N$ through the points $\{A_{v_h}\}_{h+1} \subset N$ to the terminal point $A_j \in N$. Consequently each element of Q^3 is exactly the number of ways of connecting pairs of points in N through three links.

Hence by the principle of recursive definition, we arrive at the conclusion that for any positive integer N, the elements of Q^N are integers which represent the number of ways of making pairwise connections of points on N through N links, where, in general,

$$Z^N(i, j) = \left\{\{v_h\}_{h=1}^{N} \middle| (a_{iv_1})\prod_{k=1}^{N} a_{v_hv_{h+1}} = 1, \ v_{h+1} = j\right\}.$$

We now proceed to define two more indices:

Let $\psi(i, j)$ denote the minimum of the set $\{k \mid Z^k(ij) \neq \Phi\}$ for the pair A_i, A_j of points in N. Assume that N contains n given points. Then the functions

$$D(A_i) = \max_{1 \leq j \leq n} \psi(i, j),$$

$$S(A_i) = \sum_{j=1}^{n} \psi(i, j)$$

are respectively called the *associated number* and the *Shimbel-index* of the point A_i of N.

Evidently these two indices contain statements of the "degree" of accessibility to individual vertices of the network. In respect of the associated number, if we adopt the convention that whenever the set $\{k \mid Z^k(i, j) \neq \Phi\}$ is empty, $D(A_i) = \infty$ (it can be shown that this convention makes sense mathematically), then we can interpret the associated number in the following manner: the smaller the associated number of a node the higher is its accessibility, and in particular, a completely isolated node has associated number ∞. The Shimbel-index on the other hand is an unweighted sum of certain critical linkages to a given node. From purely a priori considerations we may say that if the interest in accessibility is to isolate critical link sequences (and Kissling's, 1969, applications to highways in Canada would lead one to conclude that there are very many instances in which this is the case) the associated number is a more powerful index than the Shimbel index.

The two indices have been tested in a number of other ways in order to evaluate their use in discriminating various types of regular networks. Probably the most comprehensive is that reported in Werner *et al.* (1968), where frequency arrays of the scores of eight different types of network on the two indices have been carefully studied. On the basis of comparison of the means and variances of these scores, Werner *et al.* concluded that the Shimbel-index has by far the greater discriminating power in so far as the eight networks are concerned. Neither of the indices can however indicate the presence of directional bias in the structure of accessibility of the network.

Finally Kissling (1969) used Shimbel's shortest distance matrix for indentifying the importance of highway linkages and claimed that the index was useful in revealing probable growth points in the system. A shorter method for computing shortest distance between points in a network is identified at the end of the third section.

SECTION II

This section is devoted to a class of problems which (following the majority of scholars in this area) we shall call Steiner problems. Although the solution strategy adopted here partially integrates the abstract versions given by Gilbert and Pollak (1968), DeMar (1968), Cockayne (1967) and Melzak (1961), our motivation is different and simpler, deriving largely from elementary geometric principles. The generalized Steiner problem has practical as well as theoretical interest. For example, it has applications to optimal layout of pipelines, to optimal location of central facilities and allied geographical problems. All the examples which follow are assumed to be set in the Euclidean plane E^2 (i.e., the space of ordered pairs in which length is defined by the well-known Pythagorean theorem and in which the notion of the inner product helps us define angles). For us a consequence of assuming E^2 is that if construction cost is a monotone function of length and if construction cost is the only cost incurred

in creating networks, then the least cost network is the minimum length network.

The simplest form of the Steiner network problem may be stated thus:

(a) Given three points in E^2, find the shortest network connecting them.

(b) Given three points $\{A_i\}_{i=1}^3$ in E^2, find a fourth point $\bar{A}_4$ in E^2, so that the sum $\sum_{i=1}^{3}{}' |\bar{A}_4 A_i|$ is as small as possible. $\bar{A}_4$ may coincide with A_i, $i = 1,2,3$.

It will soon become clear that (a) and (b) are equivalent, in the sense that a solution to the one is automatically a solution to the other and vice versa. For expository simplicity, we shall concentrate on (b) (which is in fact the mathematical dual of the well-known Weberian location problem) and work through Figs. 7–9 to formulate general solution principles.

Consider Fig. 7 in which A_1, A_2 and A_3 have given fixed locations. Choose a fourth point A_4 whose location is arbitrary except that it belongs in the convex hull of A_1, A_2 and A_3 (i.e., the smallest convex set containing A_1, A_2, A_3 which is obviously the triangle $\Delta A_1 A_2 A_3$).

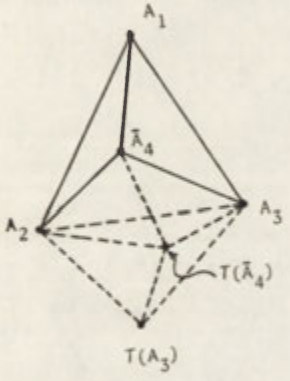


Fig. 7.

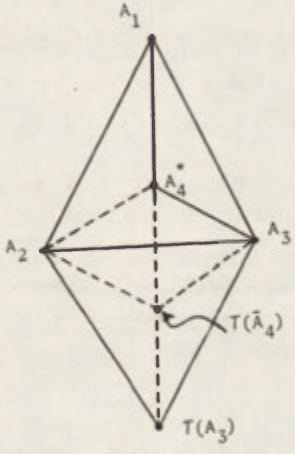


Fig. 8.

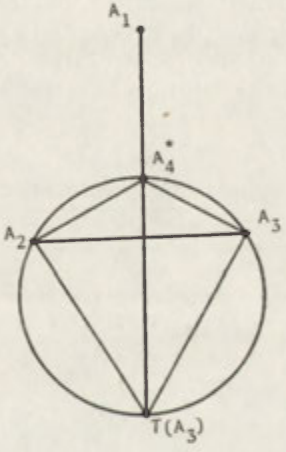


Fig. 9.

Consider the set of lines $\overline{A_4 T(\bar{A}_4)}$, $\overline{T(\bar{A}_4)T(A_3)}$ with the property that $|\overline{A_2 \bar{A}_4}| = |\overline{A_4 T(\bar{A}_4)}|$ and $|\overline{A_2 \bar{A}_4}| = |\overline{\bar{A}_4 T(A_3)}|$. (Figs. 7–9).

Clearly for any arbitrary location of A_4 inside $\triangle A_1 A_2 A_3$ it is possible to generate lines $\overline{\bar{A}_4 T(\bar{A}_4)}$ and $\overline{T(\bar{A}_4)T(A_3)}$ with the properties stated in the last paragraph. We then pose the question, what rigid motion (s) of the plane would enable us to achieve such properties? In this case, it is simpler to employ rotations. Thus for any arbitrary location of A_4 in $\triangle ABC$, a rotation of $\triangle A_2 A_3 A_4$ onto $\triangle A_2 T(A_3) T(\bar{A}_4)$ gives all the above properties. With such a rotation for any location of $\bar{A}_4$ in $\triangle ABC$, we have

$$\sum_{i=1}^{3}{}' |\overline{\bar{A}_4 A_i}| = |\overline{A_1 \bar{A}_4}| + |\overline{\bar{A}_4 T(\bar{A}_4)}| + |\overline{T(\bar{A}_4)T(A_3)}|.$$

The right hand side of this identity is smallest when and only when A_1, $\bar{A}_4$, $T(\bar{A}_4)$, $T(A_3)$ are collinear. Consequently, problem (b) reduces to finding the location for $\bar{A}_4$ such that A_1, $\bar{A}_4$, $T(\bar{A}_4)$, $T(A_3)$ are collinear.

Thus we shall completely answer question (b) if we can exhibit the exact location of $\bar{A}_4$ which satisfies the collinearity condition of the last paragraph. We turn to Fig. 8 for the basis of such a solution. Since we rotated $\wedge A_2A_3\bar{A}_4$ onto $\triangle A_2T(A_3)T(\bar{A}_4)$, we necessarily have that

(i) $\triangle A_2\bar{A}_4T(\bar{A}_4)$ is equilateral when each of its angles is 60°,

(ii) $\sphericalangle A_3A_2T(A_3) = 60°$,

(iii) $|\overline{A_2T(A_3)}| = |\overline{A_2\bar{A}_3}|$.

In consequence of facts (ii) and (iii), $\triangle A_2A_3T(A_3)$ is equilateral, which in turn implies that

(iv) $\sphericalangle A_2\bar{A}_4T(A_3) = \sphericalangle A_2A_3T(A_3)$.

In consequence of (iv) the points A_2, $\bar{A}_4$, A_3, $T(A_3)$ must lie on a circle, i.e., these points together with the line segments joining them constitute a cyclic quadrilateral. In consequence of this fact $\sphericalangle A_2A_4^*A_3 + \sphericalangle A_2T(A_3)A_3 = 180°$, which implies that $\sphericalangle A_2A_4^*A_3 = 180 - 60 = 120°$. Consequently we have arrived at a complete constructive proof of the solution of problem (b). The essential steps of the solution are now isolated (Fig. 9).

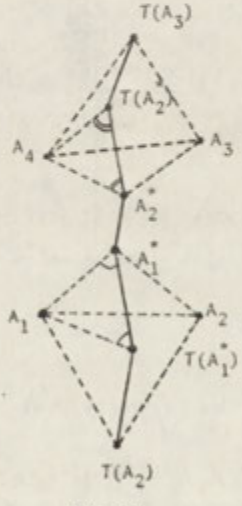


Fig. 10.

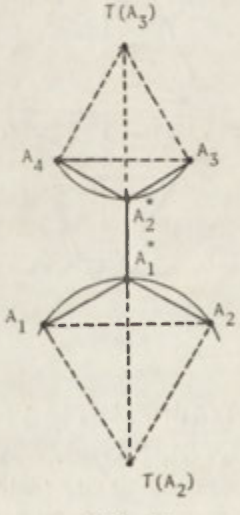


Fig. 11.

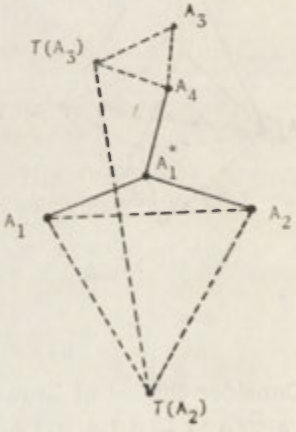


Fig. 12.

Step I: Using $\overline{A_2A_3}$ as base, construct an equilateral $\triangle A_2A_3T(A_3)$ with (A_3) as apex located on the side of BC opposite to that in which A_1 is located. It is clear from the preceding paragraphs that an implicit assumption is that $\sphericalangle A_2A_1A_3 < 120°$.

Step II: Join A_1 to $T(A_3)$ and $\overline{A_1T(A_3)}$ is the required minimal line.

Step III: Circumscribe $\triangle A_2A_3T(A_3)$ in a circle and call the intersection of this circle with $\overline{A_1T(A_3)}$ the point A_4^*. Thus the network consisting of the line segments $\overline{A_1A_4^*}$, $\overline{A_2A_4^*}$ and $\overline{A_3A_4^*}$ is the required optimal network. In case any of the angles of $\triangle A_1A_2A_3$ is greater than 120°, it is clear that A_4^* must coincide with that point.

To solve problem (a), i.e., the Steiner problem for N (>3) points, we must formulate an appropriate (b)-equivalent. To do this, we need to extract the more general properties from the $N-3$ case. We first give a couple of definitions.

D II.i Every point (such as A_4^* in Fig. 9) added to the given N points and located in their convex hull will be called an A^*-point. Furthermore, if I^* is the index set for A^*-points,

$$A^* = \{A_i^* \mid i \in I^*\}.$$

Let $\mu(A^*)$ = number of elements in A^*, and $M = N + \mu(A^*)$.

D II.ii We shall refer to the case where each of the angles of $\triangle A_1A_2A_3$ is less than 120° as the normal case. Consider Fig. 8 and observe the three angles on A_4^* which are associated with the given points A_1, A_2, A_3. We shall say that V_4^* is a carrier of (or carries) each of these angles.

D. II.iii $S\langle L, \{n_i\}_{i=1}^M, l(n_i, n_j), A^*\rangle$ is the solution to the Steiner problem (a) where L is the total length of the associated network, the set $\{n_i\}_{i=1}^M$ is the totality of nodes in the networks, and $l(n_i, n_j)$ is the link joining node n_i to n_j. For brevity, we shall occasionally denote the set simply by S.

PROPERTIES OF $S\langle L, \{n_i\}_{i=1}^M, l(n_i, n_j), A^*\rangle$.

Property 1. All $l(n_i, n_j)$, $i \neq j$, $i, j = 1, \ldots, M$, must be straight lines. Here M = number of A^*-points plus N.

Remark. Property 1 is necessarily true since the points are located in E^2.

Property 2. If S is normal, every A^*-point carries angles each of which is exactly 120°. In case A^*-points coincide with given points, then any angle carried must exceed 120°.

Proof. Suppose specifically that $\alpha < 120°$ is carried by A_i^* and is associated with given points A_k, A_j. Consider now $\triangle A_i^* A_k A_j$. It is easy to see that steps I, II and III can be used to create a new A^*-point, say $(A_i^*)_{kj}$, which carries an angle of 120° associated with A_k, A_j, and which creates a new S' of length L' with $L' < L$ contradicting the optimality of S.

Property 3. If $A_i^* \in A^*$, there must exist links $\{l(n_{i_k}, n_i)\}_{k=1}^3$ with $n_i = A_i^*$ such that $i_1 \neq i_2 \neq i_3 \neq i_1$ and

$$\bigcap_{k=1}^{3} l(n_{i_k} n_i) = \{A_i^*\}.$$

Proof. This property simply states that every A^*-point is the intersection of exactly three links. First suppose that $A_i^* \in A^*$ and that A_i^* is the intersection of two links. Then A_i^* must carry one angle greater than 120° which implies that A_i^* is exactly coincident with one of the given points hence $A_i^* \in A^*$. If, however, $A_i^* \in A^*$ and A_i^* is a terminal of one of the links in S, then we claim that S cannot be optimal, for we can generate $S' = S - \{A_i^*\}$ of smaller length and in which all given points are connected. Finally, if $A_i^* \in A^*$ is the intersection of more than three links, then one of the angles carried by A_i^* must be less than 120°, which contradicts Property 2. This completes the proof of Property 3.

Property 4. $S\langle L, \{n_i\}_{i=1}^M, l(n_i, n_j), A^*\rangle$ cannot contain a loop.

Property 5. $\mu(A^*) \leqslant N-2$.

Proof. Let $\mu(A^*) = r$. Since there are N given points, we have a total of $(N+r)$ points. In the light of Properties 1−4 together with the fact that we are operating in E^2, S is connected if S contains $(N+r-1)$ links. On the other hand, by Property 3, each member of A^* is the intersection of exactly three links. So

that by this intersection property (which forces us to include several links twice) we have $3r$ links. The links that are double-counted are those between the A^*-points of which (links) there are most $(r-1)$. Consequently S contains at least $3r-(r-1) = 2r+1$ links, i.e.,

$$N+r-1 \geqslant 2r+1,$$
$$\text{or } -r \geqslant 2-N,$$
$$\text{i.e., } r \leqslant N-2,$$

which proves Property 5.

Remark. If S is normal, then $r = N-2$.

We are now in a position to formulate the (b)-equivalent of the generalized Steiner problem.

(*b*)-*Equivalent Problem*: Let $\{A_i\}_{i=1}^N$ be a set of points located in E^2. By creating r A^*-points in the convex hull of $\{A_i\}_{i=1}^N$, $r \leqslant N-2$, construct the network of minimal length which connects all the given points via the created A^*-points. In other words, construct

$$S = S\langle L, \{n_i\}_{i=1}^M, l(n_i, n_j), \{A_i^*\}_{i=1}^r \rangle.$$

The tools for solving this problem are essentially given in Properties 1–5 and steps I–III. In the interest of concreteness, we shall illustrate the solution with $N = 4$ and $N = 5$ cases and conclude with general remarks.

Examine Fig. 10 in which the given points are A_1, A_2, A_3 and A_4. We want to construct the network of minimum length connecting these points. In this case $r \leqslant 2$. So we must really examine minimum length networks with no A^*-point, with one A^*-point and with two A^*-points to determine the global minimum. Clearly, once we specify the exact value of r, we have determined possible topologies of the network. Thus we define the minimum length network associated with a specified value of r as a relatively minimum length network (RELM), i.e., relative to a specified topology. We shall illustrate the principles with reference to the maximum value of r. We shall refer to networks associated with the maximum value of r as networks with complete topology (NECOM).

Thus, in Fig. 10, we attempt to expose the basis of a minimum length NECOM for the four given points A_1, A_2, A_3, A_4. In this case, $r = 2$ and we assume some arbitrary location for A_1^* and A_2^* initially. Using Property 3, we know that the RELM network must have the following connection pattern: $A_4 \rightleftarrows A_2^*$, $A_3 \rightleftarrows A_2^*$, $A_2^* \rightleftarrows A_1^*$, $A_1 \rightleftarrows A_1^*$, $A_2 \rightleftarrows A_1^*$. Consider the sequence of line segments (paths) which connect the points $T(A_2)$, $T(A_1^*)$, A_1^*, A_2^*, $T(A_2^*)$ and $T(A_3)$ with $|\overline{A_1A_1^*}| = |\overline{A_1^*T(A_1^*)}|$, $|\overline{A_1^*A_2}| = |\overline{T(A_1^*)T(A_2)}|$, $|\overline{A_4A_2^*}| = |\overline{A_2^*T(A_2^*)}|$ and $|\overline{A_2^*A_3}| = |\overline{T(A_2^*)T(A_3)}|$. Observe that this path has the same length as that of the network. So the network is of minimum length when this path is a straight line. Steps I–III are used to construct the associated RELM (Fig. 11) as follows:

(1) Partition the four points into two sets by joining A_1 to A_2 and A_4 to A_3.

(2) Construct equilateral $\triangle T(A_3)A_3A_4$ with apex $T(A_3)$. Construct equilateral $\triangle T(A_2)A_1A_2$ with apex $T(A_2)$. Join $\overline{T(A_2)T(A_3)}$. The circumcircle on the first equilateral $\triangle$ cuts $\overline{T(A_2)T(A_3)}$ at A_2^* and the circumcircle on the second equilateral $\triangle$ cuts $\overline{T(A_2)T(A_3)}$ at A_3^*.

(3) Join A_2^* to A_4, A_2^* to A_3, A_1^* to A_1, and A_1^* to A_2.

Given the value r, the alternative topologies are determined by the partitioning adopted in (1) so that for $r = 2$, an alternative to Fig. 11 is Fig. 14.

Consequently, the NECOM corresponding to $r = 2$ for four given points is of two types given by Fig. 11 and Fig. 14 and the appropriate RELM for $r = 2$ is Fig. 11 or Fig. 14 according as the former or the latter has the smaller total length. It could happen that the location of some of the given points, e.g., A_3, A_4 in Fig. 12, is such that the ($r = 2$) topology cannot be realized. Thus in Fig. 12, the RELM would consist of the following set of links: $\overline{A_1A_1^*}$, $\overline{A_2A_1^*}$, $\overline{A_1^*A_4}$, $\overline{A_4A_3}$.

Figure 13 illustrates a NECOM case in which there are five given points A_1, A_2, A_3, ..., A_5. In a NECOM with five given points $r = 3$. Again we partition the points into two sets of pairs and a third set consisting of the single point A_5. Next construct the equilateral $\triangle A_3A_4T(A_3)$, and $\triangle A_1A_2T(A_2)$. Join $T(A_2)T(A_3)$. The problem is now essentially reduced to that of finding the minimum length network for the three points A_5, $T(A_2)$, $T(A_3)$. Construct the equilateral $\triangle T(A_2)T(A_3)T(T(A_3))$. Join A_5 to $T(T(A_3))$ and this line intersects the circumcircle of the last equilateral triangle at A_3^*. Join $T(A_2)$ to A_3^*, $T(A_3)$ to A_3^* and these two lines intersect the circumcircle of $\triangle T(A_2)A_1A_2$ and $\triangle T(A_3)A_3A_4$ respectively at A_1^* and A_2^*. Join A_1 to A_1^*, A_2 to A_1^*, A_3 to A_2^* and A_4 to A_2^*.

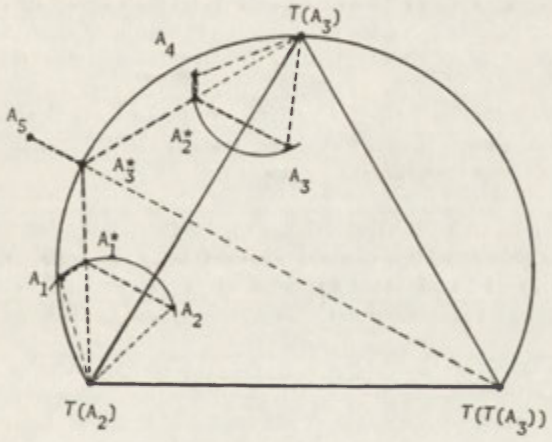


Fig. 13.

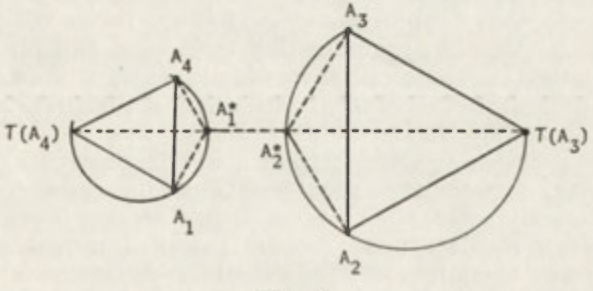


Fig. 14.

It should now be clear that for any given finite number of points, the same solution procedure applies. The major problem, however, is that, to identify the RELM, in this case we have to examine all possible partitions (or pairing of the set of given points) and construct the associated NECOM. This is certainly

an unsatisfactory state of affairs since we do not possess any general systematic procedure for eliminating alternative topologies so that to obtain minimal length networks in the global sense would involve tremendous computational drudgery. Compounding this is the fact that global optima are by no means unique.

Examples of applications in geography of the network notions which we have been discussing in this section are scanty. In fact, to date probably the only geographer who has approached network problems in a vein similar to ours is Werner (1969). However, geographical problems to which these notions are applicable are many (at least on a theoretical level) so that the scarcity of published geographical examples may be argued to be due to the fact that the type of formulation expounded in this section is relatively unknown.

SECTION III. FLOWS ON NETWORKS

III-1 OPTIMAL MULTI-SOURCE MULTI-DESTINATION FLOWS WHERE ONLY DIRECT UNCAPACITATED LINKS ARE PERMITTED

This problem is of course the network interpretation of the fundamental Hitchcock problem. Basically one seeks a pattern of demand-supply linkages (over a network of direct links) which (pattern) ensures that the following conditions are met:

(a) no supply location is called upon to supply more than its capacity,
(b) the demand of every location for the commodity is met,
(c) the total flow cost is as small as possible.

Denote the set of supply locations by S, and the set of demand points by D. Further, for every $i \in S$ and $j \in D$, let (a) t_{ij}, x_{ij} respectively be the unit flow cost and the amount of commodity moved from i to j, and (b) σ_i, d_j respectively represent availability at i and total demand at j.

The problem verbalized above may be symbolized as follows:

$$\text{Min}\Big\{\sum_{i\in S}\sum_{j\in D} t_{ij}x_{ij}\Big|\ \sum_{i\in S} x_{ij} \geqslant d_j;\ -\sum_{j\in D} x_{ij} \geqslant -\sigma_i;\ \sum_{j\in D} d = \sum_{i\in S}\sigma_i;$$

$$x_{ij} \geqslant 0,\ i \in S,\ j \in D\Big\}.$$

Several algorithms have been constructed for solving this problem. The four best known are (a) the stepping-stone algorithm, (b) modified distribution method or MODI, (c) Vogel's approximation technique, and (d) simplex algorithm. Space limitation prevents our reviewing these algorithms but the interested reader will find a particularly lucid exposition of the first three in Metzger (1958) and of the fourth in Dantzig (1962).

The Hitchcock problem as characterized above has been the foundation for several more general network flow problems. Generalizations have involved the introduction of capacity constraints (reviewed in Garrison (1960)), the inclusion of flow through intermediate nodes or the so-called transshipment problem (see Orden (1956) and Quandt (1960) for two different methods of solving this problem) and the introduction of multiple commodities (see Werner *et al.*, 1968).

Related to this basic Hitchcock problem but characteristically involving many intermediate nodes is the maximum flow problem. Simply, it is concerned with the question: Given a source, a destination and a directed network of links from the source through a finite number of nodes to the destination, what is the maximum flow that can arrive at the destination from the source given

that there are capacity constraints on the links? Probably the simplest solution technique for this problem is that due to Ford and Fulkerson (1962), which is based on the max-flow-min-cut theorem. Complete exposition of the solution method cannot be undertaken in an essay of this length and so the reader is referred to Kaufman (1967) for a thorough treatment of this problem. An important application of this maximum flow problem is to the evaluation of the efficiency of networks with respect to flows in an economy. In this respect, the maximum flow problem surpasses in power any of the indices which we evaluated in the first section. Furthermore, it can be shown that its dual solves the minimum path problem of Shimbel which forms the basis of Kissling's (1969) investigations of the linkage importance of regional highways in Canada.

Two Examples of the Maximal Flow Problem: As with the generalized Hitchcock problem, we deal with flow from an origin through several intermediate nodes to a destination. In this part, however, we put capacity constraints on the various links and ask: What is the maximum flow that the network system can accommodate per unit of time? The initial data are summarized in the network shown in Fig. 15.

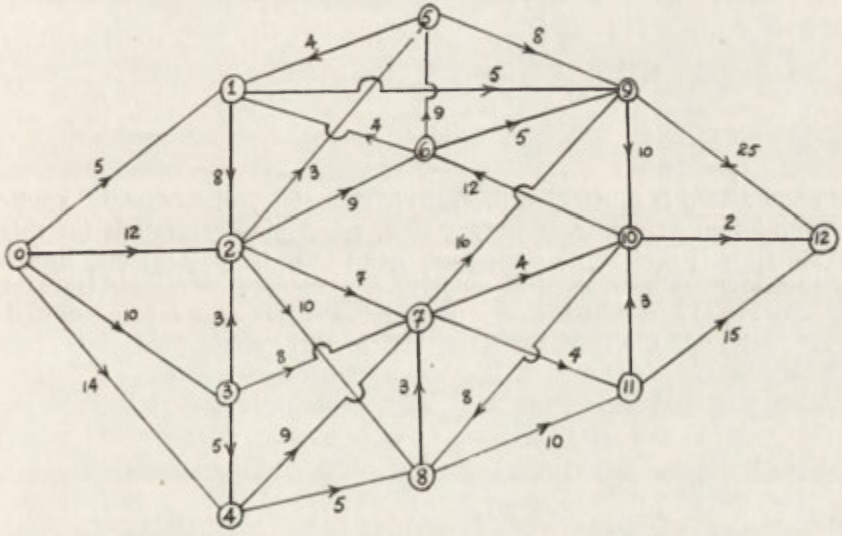


Fig. 15.

Consider Fig. 15. The numbers in circles are the nodes of which there are twelve. Node ⓪ refers to the origin and ⑫ to the terminus (or destination) of all flows. The arrows indicate the possible directions of flow and the numbers beside each arrow indicate the maximum capacity of the corresponding link. We solve the problem in a series of steps illustrated by Figs. 16–19. The first step in the solution consists in applying the law of conservation of flows to each node. Simply stated, one relabels the links in such a way that for each node the sum of incoming flows is equal to that of outgoing flows. Thus Fig. 16 is derived from Fig. 15 by application of the conservation principle. Consider node ⑦ in Fig. 16. Total incoming flow is $4+5+9+3 = 21$; total outgoing flow is $13+4+4 = 21$. The reader can check that every other node of Fig. 16 has the property that total inflow minus total outflow associated with it is zero. In

applying the conservation principle, it is essential that the capacities of links are not exceeded. Consequently, wherever the capacity of a link has been attained, we show the link by pecked lines.

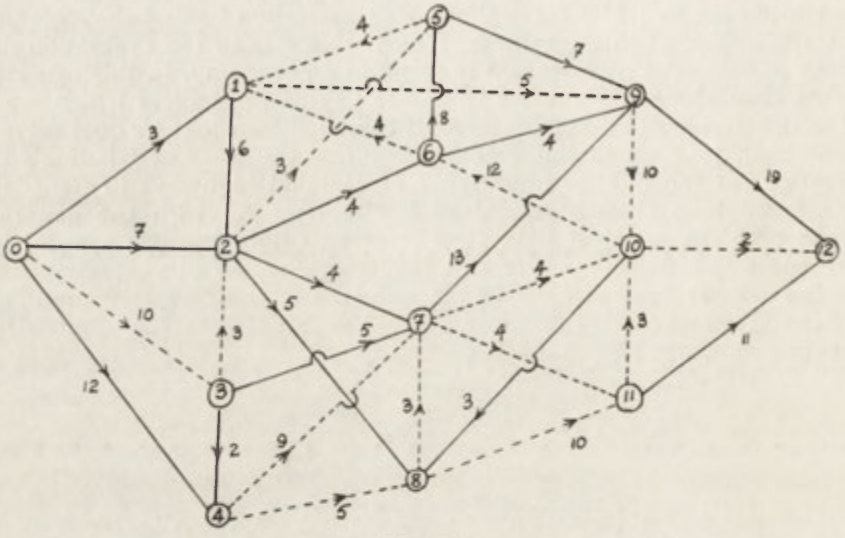


Fig. 16.

The second step is illustrated in Fig. 17. This step consists in the application of the *completion principle*. Simply stated, an origin-destination (in this case a (0) — (12)) flow is said to be complete if, and only if, at least one link in the path of the flow has been used to capacity. Thus step 2 consists in transforming Fig. 16 into Fig. 17 in which every possible path has at least one dotted link, whilst for each node the conservation principle still holds.

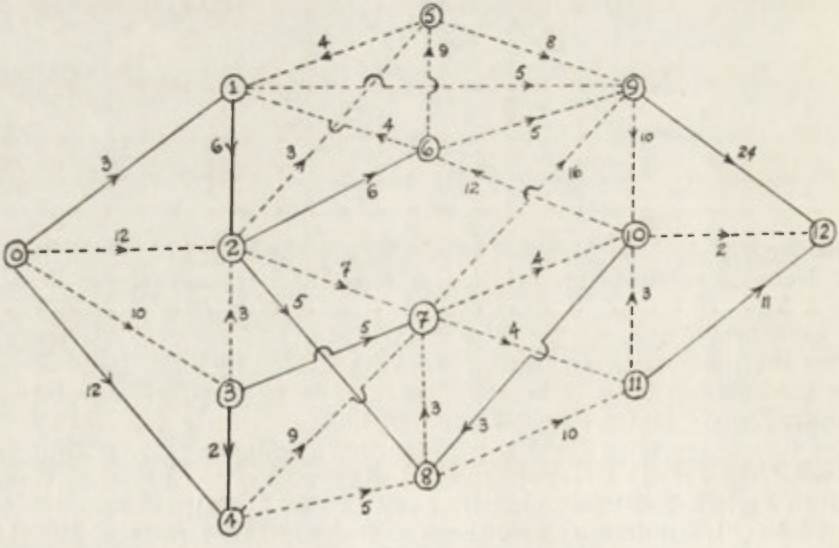


Fig. 17.

The third step is a little more tricky and somewhat more difficult to describe. This step enables us to answer the question: Given the "complete" flow structure of Fig. 17, is it possible to increase the number of "complete" links and maintain conservation of flows in each node? To answer the question we consider the set Z_0^{12}, each member of which is a chain (consisting of a sequence of successive adjacent links) with the property that the first link is incident on ⓪ and the last is incident on ⑫. Thus defining $C_m^n[k]$ as a chain consisting of k links with the first link incident on m and the last link incident on n, we note that $C_m^n[k] \in Z_0^{12}$ if $m = 0$ and $n = 12$. Now consider $C_0^{12}[k] \in Z_0^{12}$, and for the sake of simplicity label the links of $C_0^{12}[k]$ successively as $(0, \sigma(1))$, $(\sigma(1), \sigma(2))$, ..., $(\sigma(k-12, 12)$. Put a $[+]$ at the origin. There are essentially three cases to consider for any such $C_0^{12}[k]$.

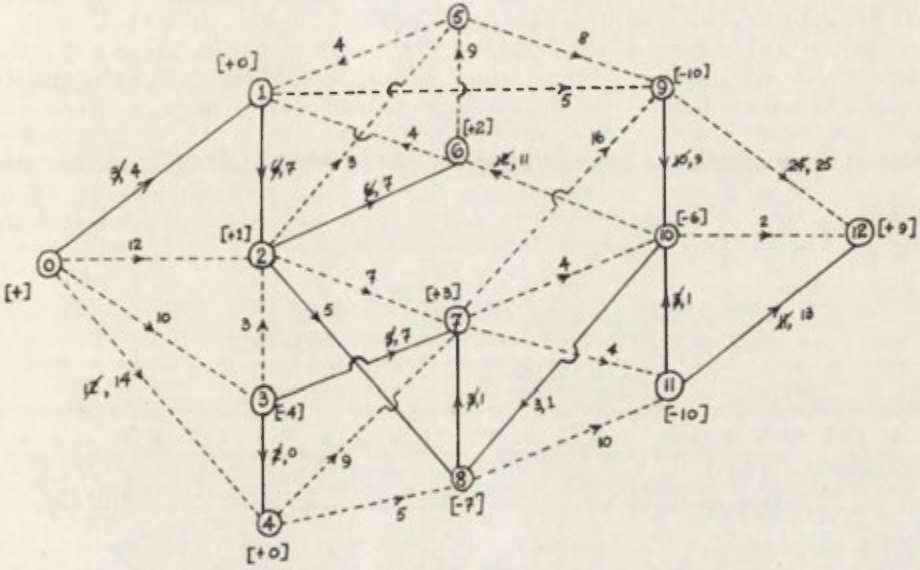


Fig. 18. First adjustment

(1) None of the links is saturated.

(2) One or more of the links are saturated and the feasible flow direction is exactly that indicated by the labelling order.

(3) One or more of the links are saturated but the feasible flow direction of such link(s) is opposite to the direction indicated by our labelling.

The reader might ask, how is case (1) possible when the flow structure of Fig. 17 is complete? The only answer is that case (1) is possible since Z_0^{12} is direction-free, i.e., elements of Z_0^{12} need not be "paths" linking ⓪ to ⑫. whereas "completeness" is path-dependent. We deal with the cases as follows:

Case. (1). Mark the origin with a $[+]$. Next consider the link $(0, \sigma(1))$. If the flow is feasible in that direction put $[+0]$ on node $\sigma(1)$; if flow is feasible only in the opposite direction, put $[-0]$ on node $\sigma(1)$. Perform the same operation on every link and every node of $C_0^{12}[k]$, so that the operation ends with $[+\sigma(k-1)]$ on node ⑫. In case there are links whose terminal nodes bear negative signs, among them identify that whose flow value v is smallest. Next consider the set of links whose terminal nodes bear positive signs; for the i-th member of this set let P_i be the difference between the flow value and the capacity, and let

$p = \min_i \{P_i\}$. Let $q = \min\{p, v\}$. If $q = v$, add v to the flow value on each positive link and subtract v from each negative link. As a result of this subtraction, the chain $C_0^{12}[k]$ will contain a link with flow value zero. In this case (unless $p = v$) no link can be completed in the chain. On the other hand, if $q = p$, add p to all positive links and subtract v from all negative links. This last operation results in at least one link in $C_0^{12}[k]$ being completed and case (1) becomes case (2).

Case (2). For case (2), we may without loss of generality suppose that only one link in $C_0^{12}[k]$ is saturated and that link has a flow direction which exactly follows the direction of the labelling. Clearly in this case there is no way of increasing the number of saturated links without exceeding the capacity of at least one link (the initially saturated link).

Case (3). In this case, use the steps of case (1) until either one of the negative links has zero flow or case (3) is reduced to case (2).

Thus by taking care of case (1)–(3) for all chains, we arrive at a situation in which we can find no $C_0^{12}[k]$ in which any additional link can be completed. How is this situation reflected in the flow structure? This question leads us to one of the most important theorems in network flow theory: max-flow-min-cut theorem. Examine Fig. 18 again. Place [+] beside the origin. Consider the chain $C_0^{12}[6]$ defined by the ordered set of links (⓪ ①), (① ②), (② ⑥), (⑥ ⑩), (⑩ ⑪), (⑪ ⑫). In view of the feasible direction of flow we have the associated sign

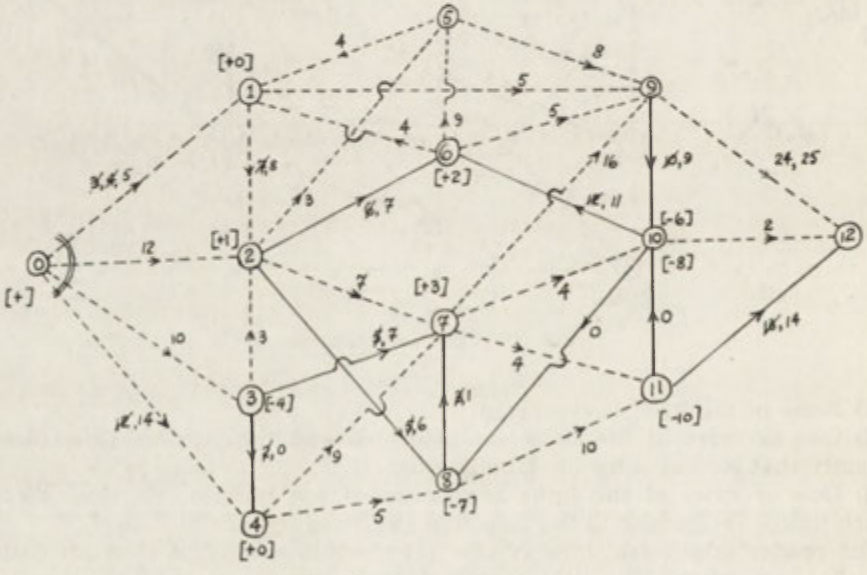


Fig. 19. Final adjustment

pattern [+], [+0], [+1], [+2], [−6], [−10], [+11]. Thus this chain is a case (3) type and the flows are changed accordingly. The same procedure is applied to all chains until we arrive at Fig. 19 where the links (⓪ ①), (⓪ ②), (⓪ ③), and (⓪ ④) are all saturated so that starting with [+] on ⓪ no sign can be put on ①, ②, ③, or ④ without exceeding the capacity of the associated links (⓪ ①), (⓪ ②), (⓪ ③) and (⓪ ④). These sets of links which separate the nodes upon which [+] or [−] can be located from the nodes on which neither [+] nor [−] can be located, constitute a cut (and in fact the minimum cut). The

total flow in such a cut is equal to the maximum flow the network can accommodate. The maximal flow in the network we have so far been studying is $5+12+10+14 = 41$. We must emphasize that the cut need not have all (or any) of its component links directly incident on the origin. Figure 20 and Fig. 21

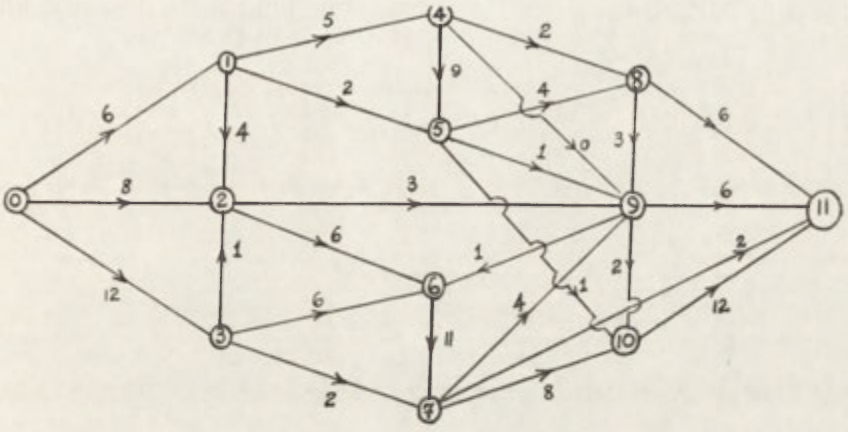


Fig. 20.

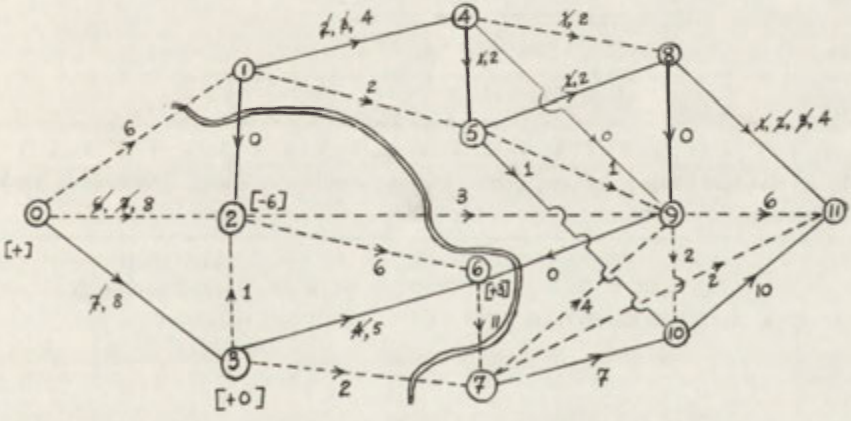


Fig. 21.

provide an example to illustrate this last point. The wavy line defines the cut which consists of those links directly intersecting the cut. Notice that in this case only one of the links is directly incident on the origin. Thus the maximum flow in this network is $6+0+3+11+2 = 22$.

III-2 ONE ORIGIN-ONE DESTINATION MANY INTERMEDIATE NODE FLOW PROBLEM

The network flow problem discussed in this section is fundamental to many sequential decision problems and in particular has all the basic essential properties of a dynamic programming problem. In the network (Fig. 22) displayed below, the number encircled indicate nodes; arrows indicate direction of links between nodes and the number beside each link is the flow cost, assuming links are uncapacitated. The problem we wish to solve is to find the sequence

of links through which commodities must move from the origin (0), i.e., supply center, to the destination ⑩ so that total flow cost is as low as possible. We shall establish the solution in stages, working backwards from the destination to the origin. The principle underlying the solution is the naive-sounding, yet fundamental principle of dynamic programming, namely that any path which is optimal for any stage must be included in the total optimal path linking the origin to the destination. We proceed stage-wise as in Table 1:

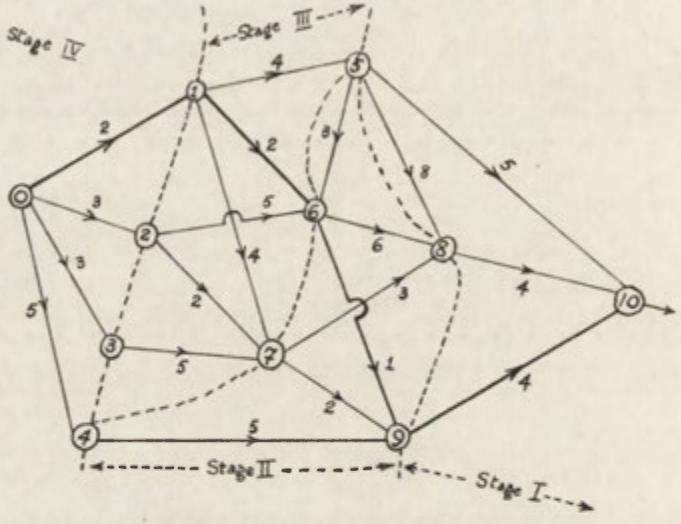


Fig. 22.

TABLE 1. Solution stages for the one origin-one destination-many intermediate nodes flow problem

	STAGE I Flow cost to destination ⑩	
From Node	⑤	5
	⑧	4
	⑨	4

	STAGE II Flow cost to destination ⑩ via nodes				Optimal Node
	⑤	⑥	⑧	⑨	
From Node ④	∞	∞	∞	9	⑨
⑤	5	13	12	∞	⑤
⑥	∞	∞	10	5	⑨
⑦	∞	∞	7	6	⑨

	STAGE III Flow cost to destination ⑩ via nodes				Optimal Node
	④	⑤	⑥	⑦	
From Node ①	∞	9	7	10	⑥
②	∞	∞	10	8	⑦
③	∞	∞	∞	11	⑦

	STAGE IV Flow cost to destination ⑩ via nodes			Optimal Node
	①	②	③	
From Node ⑩	9	11	14	①

Thus the solution consists of the following sequence of nodes (and the implied links) ⓪, ①, ⑥, ⑨, ⑩.

The solution procedure may be described as follows: One first partitions the network into stages. Stage I consists of all nodes that are connected over one link to the destination; Stage II consists of all nodes that are connected over two links to the destination, etc. The computation for Stage I involves mainly the listing of the flow cost from node ⑤ to ⑩, node ⑧ to ⑩ and node ⑨ to ⑩. For Stage II, we calculate the cost of flow from ④ directly through ⑤ to ⑩, through ⑥ to ⑩, through ⑧ to ⑩, and through ⑨ to ⑩. We use the symbol "∞" to indicate that two nodes are not directly linked. However, a unit flow from ④ through ⑨ into ⑩ costs $\$(5+4) = \$9$. Hence, we enter ⑨ in the ④ ⑨ cell of the table describing Stage II. We follow the same procedure for nodes ⑤, ⑥ and ⑦. Next, for each row (in the table representing Stage II) we identify the cell with the smallest cost and we indicate the column number corresponding to that cell and transfer that number to the extreme right under Optimal Node. Thus (in Stage II) the value 9 is the lowest in row ④ and that value corresponds to column ⑨.

By substituting distance for cost in the procedure outlined above, we may derive a solution for the minimum path problem for every pair (x, y) of points in the given network.

Call the origin x and the destination y. We label the vertices of the network as follows: x has the label zero; a vertex r has the label t if the least number of steps that must be traversed from x to get to r is t. Define $\Omega_x(t) = \{r \mid r$ is at least t steps away from $x\}$ Then $\Omega_x(t+1) = \{r \mid r \in \Omega(t)$ and a member of $\Omega_x(t)$ is linked to $r\}$. We continue in this fashion until we reach $y \in \Omega_x(n)$ and $\in \Omega_x(n-1)$. Then n is the shortest distance between x and y.

SECTION IV

In the earlier sections, problems of flows on networks, and problems of optimal network structure, have been treated separately. The relative simplicity thus achieved enabled us to obtain a variety of interesting results for networks consisting of a fairly large number of given points. Quite often, in reality, the problem of finding the optimal network structure may involve finding the network connecting a set of points so that the construction and flow costs are minimized simultaneously. Unfortunately, the introduction of differential construction and flow costs introduces at least two fundamental problems which quite severely restrict the solution, interpretation and aplication of this aspect of network theory. First, by combining both cost, we introduce considerable distortion into the uniform surface with which the second section of this paper dealt. A result of this distortion is a sharp increase in the number of network alternatives that must be examined to arrive at an optimal solution. Secondly, we run into a problem of incompatible dimensions (since flow cost is really measured per unit time per unit time) all of whose solutions to date (see Quandt, 1960) are largely artificial and prevent an unambiguous interpretation.

Bearing these two issues in mind, we shall proceed to formulate the next problem following the general structure provided initially be Friedrich (1956) and Beckmann (1952) and later elaborated by Werner (1968). However, the methods (i.e., the arguments) we employ in this section are more compact and, we hope, much simpler than those of the authors just mentioned.

Notation:

μ — flow cost per unit distance per unit flow,
f_{ij} = total flow between source A_i and destination A_j,
c = construction cost per unit length (in terms of capital cost by unit capacity per time).

Given the above variables, we proceed with a statement of our basic problem as follows:

Problem: Given are the points A_1, A_2, A_3, each in R^2. Find the point A_1^* through which A_1, A_2, and A_3 can be connected so that the network consisting of links $\overline{A_1A_1^*}$, $\overline{A_2A_1^*}$ and $\overline{A_3A_1^*}$ is of least flow plus construction costs.

Solution: One of the simplest approaches to the solution of the above problem is to set up a diagram such as Fig. 23 in which the coordinate system is arranged so that $A_1 = (0, 0)$, $A_2 = (x_2, 0)$, $A_3 = (x_3, y_3)$ and $A_1^* = (x, y)$. The total cost per unit length along the link $\overline{A_1A_1^*} = c+\mu f_{13}$; along the link $\overline{A_2A_1^*} = c+\mu f_{23}$; and along the link $\overline{A_3A_1^*} = c+\mu(f_{13}+f_{23})$.

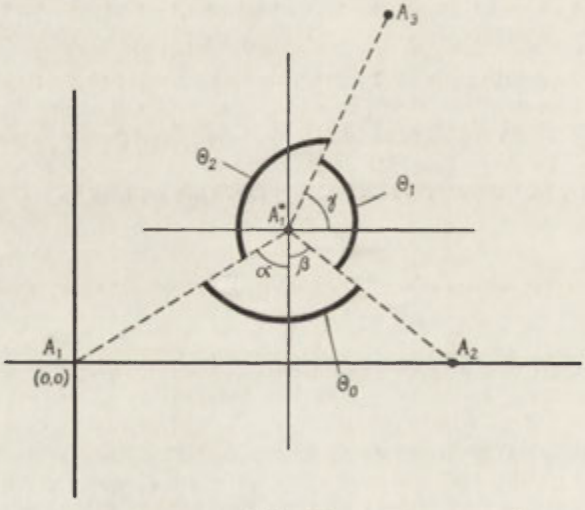


Fig. 23.

Put $t_1 = c+\mu f_{13}$; $t_2 = c+\mu f_{23}$; and $t_0 = c+\mu(f_{13}+f_{23})$. Thus, if T is the total flow and construction cost for the network in Fig. 23, we have

$$T(x, y) = T(A_1^*) = t_1(x^2+y^2)^{1/2} + t_2\big((x_2-x)^2+y^2\big)^{1/2} + t_0\big((x_3-x)^2+(y_3-y)^2\big)^{1/2}.$$

The point (x, y) for which T is minimum satisfies the following partial differential equation (which constitute necessary conditions):

$$\frac{\partial T}{\partial x} = t_1\frac{x}{(x^2+y^2)^{1/2}} - t_2\frac{x_2-x}{\big((x_2-x)^2+y^2\big)^{1/2}} - t_0\frac{x_3-x}{\big((x_3-x)^2+(y_3-y)^2\big)^{1/2}} = 0, \quad (1)$$

$$\frac{\partial T}{\partial y} = t_1\frac{y}{(x^2+y^2)^{1/2}} + t_2\frac{y}{\big((x_2-x)^2+y^2\big)^{1/2}} - t_0\frac{y_3-y}{\big((x_3-x)^2+(y_3-y)^2\big)^{1/2}} = 0. \quad (2)$$

Equations (1) and (2) above are equivalent to the following:

$$t_1 \sin\alpha - t_2 \sin\beta - t_0 \cos\gamma = 0, \tag{1'}$$

$$t_1 \cos\alpha - t_2 \cos\beta - t_0 \sin\gamma = 0. \tag{2'}$$

To simplify the relations, we first add (1′) times $\sin\beta$ to (2′) times $(-\cos\beta)$ to obtain

$$t_1 \sin\alpha \sin\beta - t_2 \sin^2\beta - t_0 \sin\beta \cos\gamma - t_1 \cos\alpha \cos\beta - t_2 \cos^2\beta + t_0 \cos\beta \sin\gamma) = 0,$$

or

$$t_1(\sin\alpha \sin\beta - \cos\alpha \cos\beta) - t_2(\sin^2\beta + \cos^2\beta) - t_0(\sin\beta \cos\gamma - \cos\beta \sin\gamma) = 0,$$

i.e.,

$$-t_1 \cos(\alpha+\beta) - t_2 - t_0 \sin(\beta-\gamma) = 0,$$

or

$$\sin^2(\beta-\gamma) = (1/t_0^2)[t_2 + t_1 \cos(\alpha+\beta)]^2.$$

Next we add (1′) times $\cos\beta$ to (2′) times $\sin\beta$ to obtain

$$t_1 \sin\alpha \cos\beta - t_2 \sin\beta \cos\beta - t_0 \cos\beta \cos\gamma + t_1 \cos\alpha \sin\beta + t_2 \cos\beta \sin\beta - t_0 \sin\beta \sin\gamma = 0,$$

or

$$t_1(\sin\alpha \cos\beta + \cos\beta \sin\gamma) - t_0(\cos\beta \cos\gamma + \sin\beta \sin\gamma) = 0,$$

or

$$t_1 \sin(\alpha+\beta) - t_0 \cos(\beta-\gamma) = 0,$$

or

$$\cos^2(\beta-\gamma) = \frac{t_1^2}{t_0^2} \sin^2(\alpha+\beta). \tag{2''}$$

Now add (1″) to (2″) to obtain

$$1 = \frac{1}{t_0^2}[t_2^2 + t_1^2 \cos^2(\alpha+\beta) + 2t_1 t_2 \cos(\alpha+\beta) + t_1^2 \sin^2(\alpha+\beta)]$$

or

$$t_0^2 = t_2^2 + t_1^2 + 2t_1 t_2 \cos(\alpha+\beta),$$

which yields

$$\cos(\alpha+\beta) = \cos\theta_0 = \frac{t_0^2 - t_1^2 - t_2^2}{2t_1 t_2}.$$

Thus the necessary conditions may be expressed in terms of angles as the following set of equations:

$$\cos\theta_i = \frac{t_i^2 - t_j^2 - t_k^2}{2t_j t_k}, \tag{3}$$

where $j \equiv (i+1) \bmod 3$, $k \equiv (i+2) \bmod 3$ and $0 \leqslant i \leqslant 2$.

SUFFICIENT CONDITIONS

To establish sufficiency, one checks that the Hessian matrix

$$H(x, y) = \begin{vmatrix} \frac{\partial^2 T}{\partial x^2} & \frac{\partial^2 T}{\partial x \partial y} \\ \frac{\partial^2 T}{\partial y \partial x} & \frac{\partial^2 T}{\partial y^2} \end{vmatrix}$$

is positive definite. A direct computation of elements of the above Hessian is quite space consuming. However, that H is positive definite follows from the fact that the function $T(x, y)$ is a strictly convex function, a fact also guarantees the uniqueness of A_i^*.

APPLICATIONS OF THE BASIC RESULT

The basic result just derived is used in solving multiple point network problems via the principles of decomposition and conjunction which we elaborate as follows. It is convenient to discuss the principles under two categories. In the first category, we already have a network but would like to investigate the possibilities for reducing flow and construction costs. In the second category, we have only a pattern of points that we wish to connect by a network of minimum flow and construction costs. It should be clear that these two categories are substantially different in at least two respects: proportion of new nodes that can be created, and hence the number of alternative network structures that may emerge. The steps involved in effecting the reduction are now briefly outlined.

Category I (Figs. 24a–24c)

Step 1. Select nodal points of the network (e.g., A^* in Fig. 24a).

Step 2. Group the links which converge on A^* into pairs, as for example, links ($\overline{A^*A_1}$, $\overline{A^*A_2}$), ($\overline{A^*A_3}$, $\overline{A^*A_4}$), and ($\overline{A^*A_5}$, $\overline{A^*A_6}$).

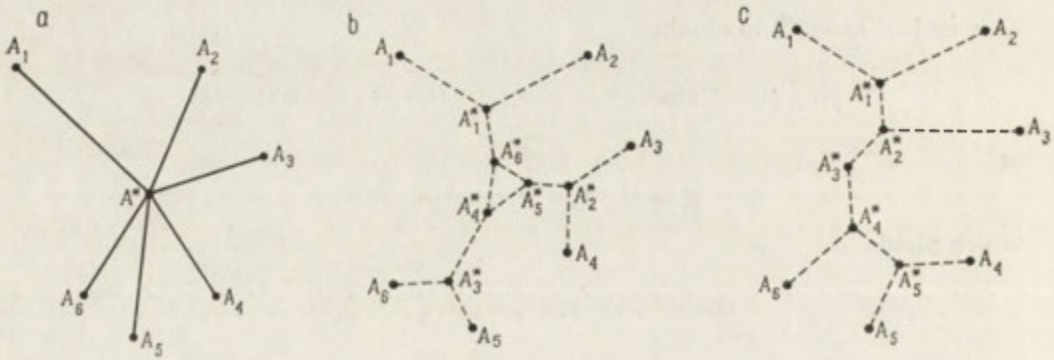


Fig. 24. Category 1

Step 3. Each pair of links, say ($\overline{A^*A_i}$, $\overline{A^*A_j}$), in the group identified is replaced by a triple of links ($\overline{A_iA_{ik}^*}$, $\overline{A_jA_{jk}^*}$, $\overline{A^*A_{ik}^*}$) so that A_{ik}^* is an additional node through which flows from two branches may be conjoined into a single branch $\overline{A^*A_{ik}^*}$.

Step 4. Consider any non-empty set $\{A^*_{i_v} A_{i_k}\}_{k=1}^{3}$ of links of conjoined flows which in consequence of the application of Step 3 intersect at a common point $A^*_{i_v}$. (There is no reason why any such point must exist. Fig. 24c is a decomposition of Fig. 24b in which no such point occurs.) In case such a point occurs, it is replaced by a triangle so that $A^*_{i_v}$ is replaced by three new points.

Step 5. Consider the set $\{A^*_{i_k}\}_{k\in I}$ of new points. Take the first element $A^*_{i_1}$ of this set, and consider the links, say $\overline{A_{i_1} A^*_{i_1}}$, $\overline{A_{i_2} A^*_{i_1}}$ and $\overline{A^*_{i_2} A^*_{i_1}}$, which are incident on the point $A^*_{i_1}$. Fix all other points of the decomposed network (i.e., the network obtained by applying Steps 1 to 4). Next apply the basic result to locate $A^*_{i_1}$ so as to minimize the flow and construction costs of the network linking the points A_{i_1}, A_{i_2} and $A^*_{i_2}$.

Step 6. Repeat step 5 for every $A^*_{i_k}$ until every such point has assumed a location from which a shift away increases the total costs. Such locations are attainable in view of the fact that Step 5 systematically reduces costs which have a lower bound.

CATEGORY II

In this second category, all of the steps used in Category I apply, together with the following additional principle which one applies during the decomposition stage. Namely, every given point is the terminal of two conjoined flows. Thus for three points, Fig. 25a shows the most general decomposition. In this threepoint case, there are exactly 12 possible solutions, each representing different topologies. Figure 25b contains four possible solutions (the triangle plus three *V*-solutions), and Fig. 25c contains eight possible solutions

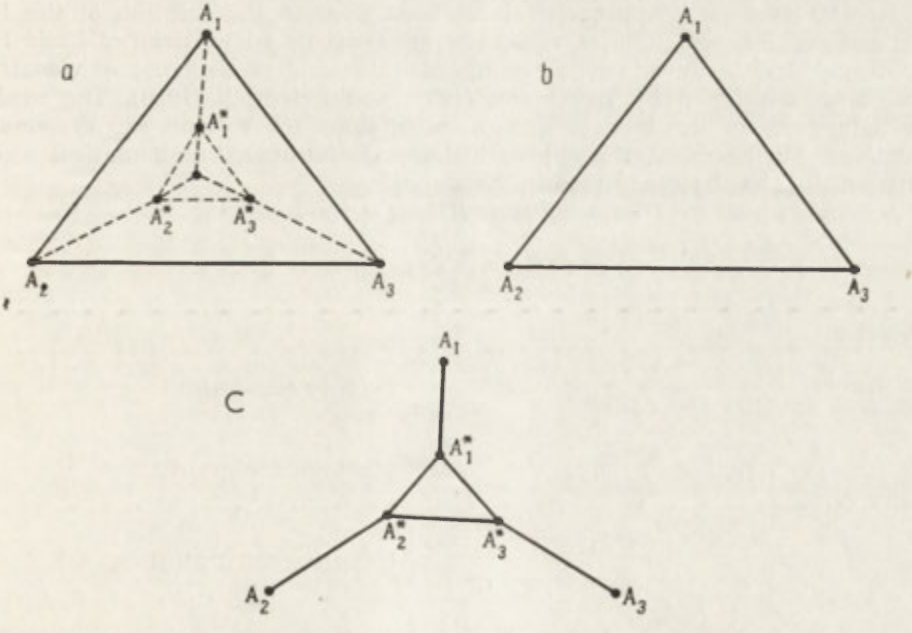


Fig. 25. Category 2

(a Y-solutions; three solutions depending on which A_i^* merges with its corresponding A_i; three solutions depending on which pair A_i^*, A_{i+1}^* merges with its corresponding A_i, A_{i+1}; and Fig. 25c itself). Thus, it can be seen that the introduction of flow costs abruptly increases the number of admissible topologies for optimal solutions.

Finally, we shall briefly discuss the problem of link addition to networks. From the point of view of application in planning situations, this problem is undoubtedly as important as any we have mentioned. It is, however, far more complex for at least two reasons. First, we require an understanding of practically all of the foregoing problems before we can even adequately characterize the important dimensions of the link addition problem. Secondly, it possesses ties with the notions of systems effect of transportation improvements, which (effects) are still scarcely understood and in any case are extremely difficult to measure cardinally. Consequently, all the difficulties we encounter in the previous sections are compounded here.

The question one seeks to answer is: Given that the capacity of a network needs to be increased, which of the existing links in the network should be improved and which new links should be created in order either (1) to provide maximum stimulus to productive activities and have enough network capacity to cope with new increases in flow or (2) to provide least (suitably defined) cost additional capacity to accommodate expected flow increases or (3) to have some combination of (1) and (2) via the concept of cost effectiveness? The formulation of a suitable model for the realization of any of the above objectives runs into difficulties because (a) the effects of capacity addition are not localized but are rather spread throughout the system and we possess no general theory for tracing them out completely and (b) the number of alternatives grows very fast compared with the increase in number of links; (for example, in a problem involving n links there are 2^n alternative solutions to consider; for four links this means 32 alternatives!).

To date two main approaches have been used in the solution of the link addition problem: calculus of variations and various adaptations of basic Hitchcock problem in linear programming. Models using the calculus of variations have been developed by Beckmann (1952) and Friedrich (1956). The models are hampered by the lack of general algorithms for solving the systems of equations. Furthermore, the approach demands far more mathematical sophistication than has been assumed in this paper.

A simple linear programming formulation is the following.

$$\text{Min} \sum_i \sum_j f_{ij} x_{ij},$$

subject to:

$$-\sum_j x_{ij} \geqslant -s_i \qquad \text{(supply constraint)}$$

$$\sum_i x_{ij} \geqslant d_j \qquad \text{(demand constraint)}$$

$$\left.\begin{aligned} -x_{ij} &\geqslant -c_{ij} - k_{ij} \\ -x_{ji} &\geqslant -c_{ji} - k_{ji} \end{aligned}\right\} \qquad \text{(capacity constraint)}$$

$$-\sum_i \sum_j k_{ij} \phi_{ij} \geqslant -T \qquad \text{(budget constraint)}$$

where r_{ij} = unit flow cost from i to j; x_{ij} = flow from i to j; c_{ij} = existing capacity of ij-links; k_{ij} = added capacity in $i-j$ link; ϕ_{ij} = unit construction (capital) cost on $i-j$ link and T – available capital budget.

The objective function is designed to minimize only flow cost and hence to avoid the problem of dimension referred to earlier. The model can be complicated in several directions, examples of which may be found in Garrison and Marble (1958), Quandt (1960) and Werner *et al.* (1968). Finally, as Roberts and Funk (1964) have reported, even this simple model can be very hard to solve because of the vast number of numerical iterations required to arrive at a final result.

CONCLUSIONS

It is important to stress three points in concluding this paper. First, only very few network concepts have been applied to geographic problems. Among these, graph theory seems to be the most frequently used and even in this case we still lack empirical interpretations which are consistent with the underlying theory. Secondly, we have hitherto been able to solve only special cases of the network problems that are of fundamental interest to us. Finally, and in consequence of the first two points, we have striven in this paper on the one hand to expose the theoretical properties of the network concepts so that interpretations might consistently be checked against the theory and on the other hand we have indicated issues that must be resolved if the more general cases of the spatial problems (identified) are to be solved.

Northwestern University, Evanston, Illinois

BIBLIOGRAPHY

Akers, S. B. Jr., 1960, The use of the Wye-delta transformations in network simplification, *Oper. Res.*, 8, 311–323.

Alao, N., 1970, A note on the solution matrix of networks. *Geogr. Anal.*, 2, 83–88.

Avondo-Bodino, G., 1962, *Economic applications of the theory of graphs*, New York, Gordon & Breach.

Beckmann, M. J., 1952, A continuous model of transportation, *Econometrica*, 20, 643–60.

Beckmann, M. J., 1967, Principles of optimum location for transportation networks, in Garrison & Marble (eds.), *Quantitative Geography*, Vol. 1, Northwestern University Studies in Geography, 13, Evanston, Northwestern University Press.

Berge, C., 1962, *The theory of graphs and its applications*, New York, Wiley.

Berry, B. J. L., 1960, An inductive approach to the regionalization of economic development, University of Chicago, *Dep. of Geography, Research Paper*, 62, 78–107.

Bott, R. and Mayberry, P. J., 1954, Matrices and Trees, in O. Morgenstern (ed.), *Economic Activity Analysis*, New York, Wiley.

Busacker, R. G. and Saaty, T. L., 1965, *Finite graphs and networks: An introduction with applications*, New York, McGraw-Hill.

Cockayne, E. J., 1962, On the Steiner problem, *Canad. Math. Bull.*, 10, 431–450.

Cockayne, E. J. and Melzak, Z. A., 1968, Steiner's problem for set terminals, *Quart. Appl. Math.*, 26, 213–218.

Cox, K. R., 1965, The application of linear programming to geographic problems, *Tijdschr. Econ. Soc. Geogr.*, 56, 228–236.

Dantzig, G. B., 1962, *Linear programming and extension*, Princeton, N. J., Princeton University Press.

DeMar, R. F., 1968, The problem of the shortest network joining *n* points, *Math. Mag.*, 225–231.

Ford, L. R. and Fulkerson, D. R., 1962, *Flows in networks*, Princeton, Princeton University Press.

Friedrich, P., 1956, Die Variationsrechnung als Planungsverfahren der Stadt- und Landesplanung, *Veroff. Akad. Raumforsch. Landespl.*, 32.

Garrison, W. L., 1960, Connectivity of the interstate highway system, *Papers, Reg. Sci. Ass.*, 6, 121–37.

Garrison, W. L., and Marble, D. F., 1958, Analysis of highway networks: A linear programming formulation, *Proc. Highw. Res. Board*, 37, 1–17.

Garrison, W. L., and Marble, D. F., 1964, Factor analytic study of the connectivity of a transportation network, *Papers, Reg. Sci. Ass.*, 12, 231–238.

Garrison, W. L., and Marble, D. F., 1965. A prolegomenon to the forecasting of transportation development, Technical Report 65–35. Prepared for the U. S. Army.

Gilbert, E. N. and Pollak, H. O., 1968, Steiner minimal trees, *Siam J. Appl. Math.*, 16, 1–29.

Gould, P. R., 1960, *The development of the transportation pattern in Ghana*, Northwestern University Studies in Geography, 5, Evanston, Northwestern University Press.

Hagget, P. and Chorley, R. J., 1969, *Network analysis in Geography*, London, Arnold.

Harary, F., 1969, *Graph theory*, Reading, Addison-Wesley.

Kansky, K. J., 1963, *The structure of transportation networks*, Chicago, University of Chicago Press.

Kaufman, A., 1967, *Graphs, dynamic programming and finite games*, New York, Academic Press.

Kissling, C. C., 1969, Linkage importance in a regional highway network, *Canad. Geogr.*, 13, 113–127.

Kuhn, H. W., 1967, On a pair of dual nonlinear programs, in J. Abadie, *Nonlinear programming*, Amsterdam, North-Holland.

Kuhn, T. E., 1962, *Public enterprise economics and transport problems*, Berkeley, University of California Press.

Loo-Keng, H. *et al.*, 1962, Applications of mathematical models to wheat harvesting, *Chinese Math.*, 2, 77–91.

Maranzana, F. E., 1964, On the location of supply points to minimize transportation costs, *Oper. Res. Quart.*, 15, 261–70.

Melzak, A. A., 1961, On the problem of Steiner, *Canad. Math. Bull.*, 4, 143–148.

Metzger, R. W., 1958, *Elementary mathematical programming*, New York, Wiley.

Miehle, W., 1966, Link-length minimization in networks, *Oper. Res.*, 14, 279–91.

Nystuen, J. D. and Dacey, M. F., 1961, A graph theory interpretation of nodal regions, *Papers, Reg. Sci. Ass.*, 7, 29–42.

Orden, A., 1956, The transshipment problem, *Manag. Sci.*, 2, 276–85.

Pitts, F. R., 1965, A graph theoretic approach to historical geography, *Prof. Geogr.*, 17, 15–20.

Quandt, R. E., 1960, Models of transportation and optimal network construction, *J. Reg. Sci.*, 2, 27–45.

Roberts, P. O. and Funk, M. L., 1964, Toward optimum methods of link addition in transportation networks, Report, M. I. T. Department of Civil Engineering.

Scott, A. J., 1967, A programming model of an integrated transportation network, *Papers, Reg. Sci. Ass.*, 19, 215–222.

Shimbel, A., 1953, Structural properties of communication networks, *Bull. Math. Biophys.*, 15, 501–7.

Shimbel, A., 1954, Structure in communication nets, *Proc. of the Symp. on Information Networks*, Brooklyn Polytech. Inst.

Shimbel, A. and Katz, W., 1953, A new status index derived from sociometric analysis, *Psychometrika*, 18, 39–43.

Warntz, W., 1957, Transportation, social physics and the law of refraction, *Prof. Geogr.*, 9, 2–7.

Werner, C., 1968, The role of topology and geometry in an optimal network design, *Papers, Reg. Sci. Ass.*, 21, 173–89.

Werner, C., 1969, Networks of minimum length, *Canad. Geogr.*, 13, 47–189.

Werner, C. *et al.*, 1968, A Research Seminar in Transportation Geography, in F. Horton (ed.), *Geographic studies of urban transportation and network analysis*, Northwestern University Studies in Geography, 16, Evanston, Northwestern University Press.

Whitin, T. M., 1954, An economic application of matrices and trees, in O. Morgenstern (ed.), *Economic activity analysis*, New York, Wiley.

Yeates, M., 1963, Hinterland delimitation: A distance minimizing approach, *Prof. Geogr.*, 15, 7–10.

CONTENTS OF VOLUMES

GEOGRAPHIA POLONICA

Vol. 1. 11 papers devoted to the present status of geography in Poland and 3 papers giving the results of research. List of Polish geographers, geographical institutions and geographical periodicals, 262 pp., 20 Figures, 1964 $6.15 (out-of-print)

Vol. 2. 34 papers prepared by Polish geographers for the XXth International Geographical Congress in London, July 1964, 259 pp., 91 Figures, 1964, $8.65

Vol. 3. Problems of Applied Geography II. Proceedings of the Second Anglo-Polish Seminar at Keele — Great Britain, September 9—20 1962, Co-edited by the Institute of British Geographers. 21 papers by British and Polish geographers, 274 pp., 69 Figures, 1964 $19.10

Vol. 4. Methods of Economic Regionalization. Materials of the Second General Meeting of the Commission on Methods of Economic Regionalization, International Geographical Union, Jablonna — Poland, September 9—14, 1963. Reports, communications and discussion, 200 pp., 6 Figures, 1964, $5.40

Vol. 5. Land Utilization in East-Central Europe. 17 case studies on land use in Bulgaria, Hungary, Poland and Yugoslavia, 498 pp., 104 Figures, 16 colour maps, 1965, $16.95

Vol. 6. 14 papers prepared by Polish geographers for the Seventh World Conference of INQUA ni U.S.A., September 1965, 150 pp., 86 Figures, 1965, $4.35

Vol. 7. 10 papers on the geography of Poland, mostly dealing, with the economic-geographical problems of Poland, 132 pp., 46 Figures, 1965, $3.60

Vol. 8. Aims of Economic Regionalization. Materials of the Third General Meeting of the Commission on Methods of Economic Regionalization IGU, London, July 23, 1964. Report and 5 papers, 68 pp., 7 Figures, 1965, $1.65

Vol. 9. Colloque de Géomorphologie des Carpathes. Materials of the geomorphological symposium held in Cracov and Bratislava, September 17—26, 1963. Report, 7 papers, 2 summaries, 118 pp., 22 Figures, 1965, $2.90

Vol. 10. Geomorphological Problems of Carpathians II. Introduction and 6 papers by Rumanian, Soviet, Polish, Hungarian and Czech geographers, 172 pp., 68 Figures 1966, $4.55

Vol. 11. 11 papers prepared by Polish geographers dealing with the history of Polish geography, Polish studies on foreign countries and different economic-geographical questions concerning Poland, 154 pp., 36 Figures, 1967, $3.90

Vol. 12. Formation et l'Aménagement du Reseau Urbain. Proceedings of the French-Polish Seminar in urban geography. Teresin, Poland, September 20—30, 1965. Papers by French and Polish geographers, discussion, 298 pp., 51 Figures, 1967, $7.25

Vol. 13. 9 papers embracing different fields of both, physical and economic geography, all of which have been devoted to methodological problems and research techniques, 130 pp. 4 Figures, 1968, $3.90

Vol. 14. Special issue for the 21st International Geographical Congress in New Delhi, 968, 43 papers prepared by Polish geographers: 24 dealing with physical and 19 with economic and human geography. List of geographical institutions in Poland and Polish geographers, 406 pp., 89 Figures, 1968, $11.90

Vol. 15. Economic Regionalization and Numerical Methods. The volume contains the final report on the activities of the IGU Commission on Methods of Economic Regionalization, as well as a collection of 8 papers by American, Canadian, Soviet and Polish authors, 240 pp., 54 Figures, 1968, $6.05

Vol. 16. 11 papers dealing with research problems and techniques in both economic and physical geography, 136 pp., 27 Figures, 1969, $3.45

Vol. 17. Special issue prepared for the 8th Congress of the International Union for Quaternary Research Paris, 1969. 28 papers by Polish authors, including studies in stratigraphy and neotectonics (6), geomorphology and paleohydrology (10), paleobotany (3), sedimentology (5), archeology (4), 428 pp., 122 Figures, 1969, $10.80

Vol. 18. Studies in Geographical Methods. Proceedings of the 3rd Anglo--Polish Geographical Seminar, Baranów Sandomierski, September 1—10, 1967, 260 pp., 54 Figures, 1970, $7.35

Vol. 19. Essays on Agricultural Typology and Land Utilization, 20 papers presented at the meeting of the Commission on World Agricultural Typology of the IGU, held 1968 in New Delhi, 290 pp., 97 Fugures, 1970, $7.25

Vol. 20. 9 papers on various aspects of both physical and economic geography, including urbanization, international trade, changes in rural economy, industrial development, urban physiography and hydrographic mapping, 183 pp., 69 Figures, 1972, $4.75

Vol. 21. 10 papers dealing with selected problems of economic growth, transportation, cartographic methods and theory, climatology and geomorphology, 147 pp., 82 Figures, 1972, $3.90

Vol. 22. 15 papers prepared for the XXIInd International Geographical Congress in Montreal, August 1972, 205 pp., 43 Figures, 1972, $5.95

Vol. 23. Present-day Geomorphological Processes. Issue prepared for the 22nd International Geographical Congress by the IGU Commission on Present-day Geomorphological Processes, 180 pp., 82 Figures, 1972, $4.85

Vol. 24. Geographical aspects of urban-rural interaction. Proceedings of the 4th Anglo-Polish Geographical Seminar, Nottingham, September 6–12, 1970, 256 pp., 76 Figures, 1972, $7.35

Subscription orders for the GEOGRAPHIA POLONICA should be placed with FOREIGN TRADE ENTERPRISE ARS POLONA — RUCH

Warszawa, Krakowskie Przedmieście 7, Poland

Cables, ARSPOLONA, Warszawa

TABLE 4. The behaviour matrix scores on dyadic factor I
The result of the application of factor analysis to a dyadic matrix of Polish commodity flows, by value in 1958

Destination voivodship / Origin voivodship	Białystok	Bydgoszcz	Gdańsk	Katowice	Kielce	Koszalin	Kraków	Lublin	Łódź	Olsztyn	Opole	Poznań	Rzeszów	Szczecin	Warszawa	Wrocław	Zielona Góra	Total	rank
Białystok	0	−2.9725	−3.2169	−1.5293	−3.0773	−3.5063	−2.8661	−2.4166	−3.1898	−2.5975	−3.4658	−3.2080	−3.4082	−2.8833	−0.3371	−3.1821	−3.5157	−45.3725	17
Bydgoszcz	−3.0131	0	−0.7875	4.4609	−2.7201	−2.4739	−1.7191	−2.5961	−1.0426	−2.6983	−2.2865	0.1056	−3.2968	−2.1974	−0.0475	−1.4096	−2.9363	−24.6583	6
Gdańsk	−2.8093	−0.7012	0	−1.8872	−3.3851	−2.2842	−3.1950	−3.3489	−3.1217	−2.1196	−3.2967	−2.4683	−3.2966	−2.9294	−1.9540	−2.8193	−3.3801	−42.9966	15
Katowice	−1.0274	3.0888	4.7860	0	2.7946	−1.7137	20.3550	0.5127	4.3752	−1.0046	11.9505	5.5255	2.2827	−0.3101	11.5167	8.3166	−0.9341	70.5144	1
Kielce	−3.2136	−2.2783	−2.8734	5.0005	0	−3.4644	2.3929	−1.9735	−0.5470	−2.9074	−2.5165	−2.1762	−2.0541	−3.1895	0.5303	−1.5741	−3.1383	−23.9828	5
Koszalin	−3.4416	−1.6716	−2.0344	−0.5470	−3.0936	0	−3.3764	−3.4623	−2.8177	−3.4837	−3.2360	−0.7236	−3.4582	−2.1115	−1.4903	−2.9591	−3.3111	−41.2181	14
Kraków	−2.5366	−1.2804	−0.9781	20.9774	1.4251	−2.7482	0	3.0014	1.2412	−2.7059	0.6194	2.6344	4.1012	−1.7520	3.7296	1.3208	−2.6324	24.4169	4
Lublin	−2.5624	−2.1228	−2.2994	−1.2636	−2.9375	−3.3342	−2.6840	0	−2.5086	−2.9983	−3.3140	−1.4856	−1.9112	−3.0516	−0.3105	−2.7253	−2.7177	−38.2267	12
Łódź	−3.2981	−2.9927	−3.1607	−1.4705	−2.7062	−3.4649	−2.7212	−3.2495	0	−3.4330	−3.3006	−1.5769	−3.4034	−3.3287	−1.5667	−2.3533	−3.3327	−45.3591	16
Olsztyn	−2.2890	−1.5473	−1.7432	−0.8538	−2.9113	−3.3816	−3.0127	−2.1458	−2.5998	0	−3.3493	−2.6022	−3.4647	−3.1813	1.0131	−3.1415	−3.3861	−38.5965	13
Opole	−2.7539	−0.4908	−0.2270	22.0963	−1.3957	2.4389	2.2781	−1.2071	0.3016	−2.7856	0	4.2663	−1.2983	−0.1535	4.0815	3.3043	−1.8825	26.5726	3
Poznań	−2.9433	−1.0083	−2.2229	1.0056	−1.9620	−2.7746	−1.6307	−2.3383	−1.2717	−2.7814	−2.5276	0	−3.0633	−2.4752	2.9859	−0.8764	−1.2539	−25.1381	7
Rzeszów	−3.2712	−2.9665	−2.7026	−0.9614	−1.9796	−3.2833	2.1085	−1.3026	−2.6820	−3.2095	−2.7286	−2.0657	0	−3.2308	−1.8624	−2.4543	−3.3283	−35.9203	10
Szczecin	−3.1863	−1.7201	−2.4806	−1.1427	−3.0318	−1.4160	−2.9172	−3.1196	−1.3657	−3.1869	−3.0560	0.0107	−3.3746	0	0.6118	−2.5976	−1.9973	−33.9699	9
Warszawa	−2.3550	−1.4721	−2.5877	−0.2635	−2.5761	−3.3268	−2.6065	−0.8404	−2.1981	−2.5524	−3.2263	−0.8970	−3.0403	−3.0843	0	−2.4911	−3.2986	−36.8162	11
Wrocław	−2.9769	0.3565	−1.6812	17.5776	0.2248	−1.9012	4.5666	−1.0460	3.5608	−2.9518	2.3677	5.3229	−0.5260	0.7479	3.8556	0	0.1586	27.6559	2
Zielona Góra	−3.4628	−2.4890	−2.8674	0.4255	−3.0662	−3.2803	−2.6435	−3.2930	−2.2002	−3.1322	−2.8941	1.1849	−3.3210	−2.4532	0.2492	2.3314	0	−30.9119	8
Total	−45.1405	−22.2683	−27.0770	61.6248	−30.3980	−39.9147	2.3285	−28.8256	−16.0661	−44.5481	−24.2604	1.8468	−32.5328	−35.5839	21.0052	−13.3106	−40.8865	−314.0072	
rank	17	7	9	1	11	14	3	10	6	16	8	4	12	13	2	5	15		

TABLE 5. The behaviour matrix scores on dyadic factor II
The result of the application of factor analysis to a dyadic matrix of Polish commodity flows, by value in 1958

Destination voivodship / Origin voivodship	Białystok	Bydgoszcz	Gdańsk	Katowice	Kielce	Koszalin	Kraków	Lublin	Łódź	Olsztyn	Opole	Poznań	Rzeszów	Szczecin	Warszawa	Wrocław	Zielona Góra	Total	rank
Białystok	0	−0.2806	−0.4876	1.4558	−0.3395	−0.6316	−0.2405	−0.1630	−0.4987	0.1151	−0.5266	−0.1994	−0.5703	0.6342	0.2982	−0.1841	−0.6033	−2.2219	8
Bydgoszcz	−0.4592	0	0.7038	−1.2095	−0.5583	0.3744	−0.7042	−0.3220	−0.5674	−0.3051	−0.1105	2.1922	−0.6046	−0.0016	0.3496	−0.3230	−0.5407	−2.0861	7
Gdańsk	−0.3271	0.1075	0	−0.5006	−0.5978	0.5017	−0.4786	−0.5688	−0.4496	0.0749	−0.5211	0.3105	−0.6047	−0.2793	−0.1900	−0.6303	−0.5557	−4.7090	12
Katowice	−1.2811	−2.7656	−3.1200	0	−2.0120	−1.0726	−6.4865	−0.9433	−2.6100	−1.3748	−5.5319	−3.3757	−1.8299	−1.7089	−5.7192	−4.6583	−1.5428	−46.0326	17
Kielce	−0.6289	−0.7321	−0.6901	0.0851	0	−0.6285	−1.2320	−0.4219	−0.8618	−0.6162	−0.5698	−0.4758	−0.7062	−0.6826	−1.2017	−0.5290	−0.3796	−10.2711	14
Koszalin	−0.6016	0.5192	0.4599	1.1912	−0.3851	0	−0.5499	−0.5972	−0.1565	−0.6325	−0.3976	2.2882	−0.5939	0.6898	0.1992	−0.0319	−0.4040	0.9973	3
Kraków	−0.6721	−0.6806	−0.6752	−4.4351	−1.4431	−0.5513	0	−1.4185	−1.0593	−0.5825	−1.3171	−1.6437	−1.2057	−0.9738	−1.5154	−1.1596	−0.6297	−19.9627	16
Lublin	−0.8762	−0.5855	−0.8960	1.6460	−0.3996	−0.6828	0.2051	0	0.6477	−0.7569	−0.2763	1.3308	−0.4845	−0.7566	−0.7782	0.6632	0.6695	−1.3303	5
Łódź	−0.5565	−0.3589	−0.6000	0.6800	−0.4695	−0.6284	−0.7179	−0.5618	0	−0.6231	−0.5608	1.7995	−0.6273	−0.4965	0.4747	−0.0726	−0.5294	−3.8485	10
Olsztyn	0.0713	0.4541	1.1292	2.1598	−0.3002	−0.4783	−0.3233	−0.0429	−0.2410	0	−0.4485	0.6757	−0.6009	−0.0938	1.4556	−0.2291	−0.4935	2.6942	2
Opole	−0.4526	−1.2878	−0.5973	0.1764	−0.8408	0.3589	−0.2628	−0.0176	−0.5561	−0.6133	0	−1.2571	−0.8694	−1.3543	0.9405	0.0102	−0.8212	−7.4383	13
Poznań	−0.4769	0.8194	−0.2432	1.4800	0.0039	−0.4412	−0.1011	−0.2928	0.1143	−0.4409	−0.0136	0	−0.5459	−0.3662	6.7983	0.7000	0.4581	7.4522	1
Rzeszów	−0.5981	−0.5474	−0.5520	0.0850	−0.4242	−0.6699	1.8091	−0.3353	−0.5162	−0.5961	−0.4144	−0.0939	0	−0.6006	−0.4087	0.0200	−0.5730	−4.3557	11
Szczecin	−0.5737	−0.3506	−0.4937	0.1353	−0.5327	0.0670	−0.4719	−0.5475	0.0131	−0.5863	−0.4138	1.4090	−0.5824	0	0.5214	−0.2252	0.3380	−2.2940	9
Warszawa	−0.7659	0.5554	−0.7498	2.1329	−0.6463	−0.5379	−0.3442	−0.8752	−0.3409	−0.4515	−0.5919	3.2384	−0.5812	−0.1735	0	0.3529	−0.4555	−0.2342	4
Wrocław	−0.6525	−0.7826	−0.7515	−4.0680	−0.1326	−0.7088	−0.7070	−0.4673	−0.2010	−0.6864	−0.5549	0.5475	−0.6455	−1.6345	0.4969	0	−0.8338	−11.7820	15
Zielona Góra	−0.6357	−0.2543	−0.3287	1.1594	−0.4256	−0.6248	−0.2521	−0.5675	0.0191	−0.5233	−0.3768	1.4430	−0.5837	−0.4042	0.6771	0.1318	0	−1.5463	6
Total	−9.4868	−6.1704	−7.8922	2.1737	−9.5052	−6.2941	−10.8578	−8.1426	−7.2583	−8.5989	−12.6256	8.1292	−11.6361	−8.2024	2.3983	−6.1650	−6.8966	−106.9708	
rank	13	5	9	3	14	6	15	10	8	12	17	1	16	11	2	4	7		

TABLE 6. The behaviour matrix scores on dyadic factor I
The result of the application of factor analysis to a dyadic matrix of Polish commodity flows, by value in 1966

Destination voivodship / Origin voivodship	Białystok	Bydgoszcz	Gdańsk	Katowice	Kielce	Koszalin	Kraków	Lublin	Łódź	Olsztyn	Opole	Poznań	Rzeszów	Szczecin	Warszawa	Wrocław	Zielona Góra	Total	rank
Białystok	0	−1.7659	−1.2971	−1.1592	−2.3468	−2.6252	−2.2408	−1.7603	−2.2862	−0.8109	−2.2155	−2.1863	−2.3006	−2.4414	−0.7031	−2.2634	−2.3117	−30.7114	11
Bydgoszcz	−2.1971	0	1.6142	3.0122	−1.5665	−1.3208	0.6668	−2.1197	0.0373	−0.7025	−0.8641	1.2602	−2.1105	−1.3875	1.2606	−0.5431	−1.7634	−6.7239	4
Gdańsk	−2.3494	−0.8954	0	−1.0027	−2.5474	−1.8078	−2.1769	−2.6229	−2.2926	−1.0466	−2.6268	−1.9033	−2.6191	−2.4720	−1.5678	−2.3267	−2.4918	−32.7492	13
Katowice	−1.7258	1.9965	1.0289	0	1.4670	−1.9404	11.8524	−0.8312	3.6732	−1.0508	8.7946	2.8932	0.2661	0.2934	4.0303	8.9390	−1.2182	38.4682	1
Kielce	−2.5662	−2.1103	−2.5520	1.6350	0	−2.6122	−0.3581	−1.5367	−1.2171	−2.1800	−2.1582	−1.8760	−2.2485	−2.3182	−0.9760	−1.9151	−2.3803	−27.3699	9
Koszalin	−2.5248	−1.9648	−2.1746	−2.0104	−2.5174	0	−2.3594	−2.6292	−2.3692	−2.6254	−2.5086	−1.6125	−2.6170	−2.0843	−2.3355	−1.9637	−2.4429	−36.7397	16
Kraków	−1.5812	0.0122	−1.6983	7.9578	−0.0542	−2.3265	0	−0.2643	0.1227	−2.1247	1.6142	−0.7072	2.4582	−2.0350	0.9524	−0.0074	−2.0786	0.2401	2
Lublin	−2.1069	−1.8561	−1.8411	−1.3397	−1.9702	−2.3065	−1.2005	0	−2.3764	−2.0097	−2.2210	−0.6959	−0.0550	−2.4213	−0.7589	−1.9126	−2.2665	−27.3383	8
Łódź	−2.6313	−1.4726	−2.0132	−0.9576	−2.1388	−2.2692	−2.1072	−2.5485	0	−2.4892	−2.3264	−0.6635	−2.4185	−2.2008	−1.6339	−2.0943	−2.5577	−32.5227	12
Olsztyn	−1.5386	−2.4211	−1.5491	−2.2177	−2.6001	−2.5702	−2.1248	−2.5201	−2.5945	0	−2.5010	−2.1389	−2.4061	−2.6882	−1.9181	−2.4792	−2.6646	−36.9263	17
Opole	−2.3069	−0.9648	−2.2142	7.3926	−1.6525	−1.7452	0.0841	−1.2911	−1.4373	−2.2488	0	1.6896	−1.3683	−1.2397	−0.4873	1.3611	−1.8150	−8.2437	5
Poznań	−2.5898	0.8790	−1.8382	0.0438	−1.6730	−1.6357	−1.2159	−2.4890	−0.6963	−2.4991	−1.6444	0	−2.3760	−1.5856	−1.3965	−0.2296	−0.2616	−21.2079	7
Rzeszów	−2.3242	−1.6697	−1.7323	−1.5962	−2.0506	−2.6540	−1.1941	−1.2012	−1.5251	−2.3966	−2.5213	−1.5275	0	−2.3259	−1.6259	−1.6788	−2.3928	−30.4192	10
Szczecin	−2.6598	−2.0988	−2.0116	−1.8094	−2.4816	−1.2213	−2.3948	−2.5445	−2.4647	−2.6546	−2.4631	−1.2815	−2.5380	0	−2.3801	−2.3112	−1.7522	−35.0672	14
Warszawa	−1.5367	−1.0612	−0.3981	3.4070	−2.1558	−2.3382	−1.3867	−2.1445	−1.8498	−2.1928	−2.1932	0.1294	−2.2328	−0.8675	0	−1.6051	−2.3374	−20.7634	6
Wrocław	−2.5376	−0.6742	−1.5060	6.1196	−1.5974	−1.5506	1.0615	−2.0030	−0.6686	−2.3250	2.9833	1.9707	−1.7913	−0.6235	−0.4029	0	0.9881	−2.5569	3
Zielona Góra	−2.4995	−2.3176	−2.3582	−1.8908	−2.6094	−2.5565	−2.2872	−2.4824	−2.3843	−2.7082	−2.4887	−1.1302	−2.9527	−2.3625	−2.5038	−1.0165	0	−36.1885	15
Total	−35.6758	−18.3848	−22.5409	15.5903	−28.4947	−33.4803	−7.3816	−30.9886	−20.3319	−32.0649	−15.3402	−7.7797	−26.9501	−28.7600	−12.4465	−12.0466	−29.7466	−346.8229	
rank	17	7	9	1	11	16	2	14	8	15	6	3	10	12	5	4	13		

TABLE 7. The behaviour matrix scores on dyadic factor II
The result of the application of factor analysis to a dyadic matrix of Polish commodity flows, by value in 1966

Destination voivodship / Origin voivodship	Białystok	Bydgoszcz	Gdańsk	Katowice	Kielce	Koszalin	Kraków	Lublin	Łódź	Olsztyn	Opole	Poznań	Rzeszów	Szczecin	Warszawa	Wrocław	Zielona Góra	Total	rank
Białystok	0	−0.2018	−0.6714	−0.2399	0.0687	0.1847	0.0540	−0.0894	−0.0214	−0.7571	0.0244	−0.1150	0.1009	0.0589	−0.6124	−0.0309	−0.0538	−2.3015	16
Bydgoszcz	−0.0799	0	−1.5363	−2.4174	−0.3702	−0.2616	−0.1135	−0.0561	−1.4324	−0.1091	−0.7259	−1.7976	−0.0967	−0.4846	−1.5324	−0.9733	−0.3352	−12.3222	17
Gdańsk	−0.0004	−0.7433	0	0.1581	0.3127	−0.0174	0.2797	0.2432	0.2091	−0.1208	0.2692	0.0026	0.2530	0.1935	0.1120	0.2355	0.2233	1.6080	13
Katowice	0.9315	3.6990	3.5088	0	6.4000	0.9296	11.7153	2.7189	5.4081	1.6341	8.0365	5.1416	2.6556	3.2140	7.0569	13.1752	1.8879	78.1130	1
Kielce	0.3256	0.5433	0.3149	1.3773	0	0.7648	0.3772	0.6610	0.5738	0.3835	0.3355	0.4119	0.3960	0.4653	0.6252	0.8392	0.3630	8.7575	4
Koszalin	0.2979	−0.0055	0.0684	0.1727	0.1671	0	0.6364	0.2488	0.1449	0.2361	0.1586	−0.0807	0.2227	0.0427	0.2253	0.0369	0.1288	2.7011	10
Kraków	0.3933	0.8073	0.4532	6.2360	1.9128	0.4358	0	1.9109	1.0329	0.3953	2.4579	0.4717	0.8557	0.6521	1.1210	1.0696	0.6427	20.8482	2
Lublin	0.0334	0.0825	0.2181	−0.1535	0.1421	0.2161	−0.1230	0	1.1845	0.1617	0.1412	−0.4604	−0.4199	0.2131	0.0742	0.1746	0.1080	1.5927	14
Łódź	0.2377	0.1888	0.3290	0.4074	0.2672	0.2490	0.2187	0.2659	0	0.7336	0.1829	−0.3458	0.2669	0.5301	0.1949	0.3202	0.2905	4.3370	7
Olsztyn	−0.1543	0.1734	0.1548	0.0624	0.2195	0.2164	−0.0171	0.2400	0.2264	0	0.4809	−0.0614	0.1561	0.2237	0.0037	0.1814	0.2021	2.3080	11
Opole	0.3847	0.3092	0.3420	1.6008	1.7493	0.4862	0.1728	0.8774	0.1668	0.2965	0	−0.4504	0.1260	0.6730	1.4605	0.0666	0.4572	8.7186	5
Poznań	0.2969	−1.0676	0.2391	−0.0701	0.1874	−0.0470	−0.0362	0.4210	−0.1333	0.2804	0.0894	0	1.3238	0.1564	0.1591	0.2989	−0.1770	1.9212	12
Rzeszów	0.1836	0.2214	0.0701	0.3399	0.0522	0.2673	−0.0291	−0.2797	0.2978	0.1514	0.2254	0.2284	0	0.5756	0.0297	0.7406	0.1542	3.2288	9
Szczecin	0.2367	0.1054	0.1863	0.2481	0.4101	−0.0384	0.3545	0.2360	0.2109	0.2276	0.1420	−0.0241	0.2563	0	0.5922	0.3321	−0.1058	3.3699	8
Warszawa	0.0250	−0.3384	−0.5949	−1.6420	0.5366	0.1752	−0.0755	0.2681	0.2687	0.2579	0.2070	−1.0781	0.3018	−0.5195	0	0.7778	0.1521	−1.2782	15
Wrocław	0.2980	1.0640	0.5518	2.3547	1.4295	0.5911	0.8122	0.5414	0.6345	0.5903	0.5078	0.9748	0.6117	0.8025	1.0329	0	3.5231	16.3203	3
Zielona Góra	0.4432	0.3175	0.1979	0.0561	0.2763	0.2872	0.1742	0.2766	0.2386	0.2745	0.2207	0.1591	0.3497	0.2188	0.2839	1.0901	0	4.8644	6
Total	3.8529	5.1552	3.8318	8.4906	13.7613	4.4390	14.3986	8.4840	9.0099	4.6359	12.7535	2.9766	7.3596	7.0156	10.8267	18.3345	7.4611	142.7868	
rank	15	12	16	7	3	14	2	8	6	13	4	17	10	11	5	1	9		

Cena zł. 34.—