Pakistan Journal of Applied Economics (1983) vol. II, no.1 (39–63).

QUANTITATIVE ANALYSIS OF WORLD DEVELOPMENT: A Cluster Analytic Approach

J. A. ZERBY and M. Habibullah KHAN*

A number of multivariate statistical techniques such as Principal Component Analysis, Factor Analysis, Discriminant Analysis, Cannonical Analysis and Multidimensional Scaling have so far been applied to the study of development. Some of these applications are examined and it is suggested that a new technique known as 'Cluster Analysis' can be used more effectively than the previous ones for analysing the nature and process of development. As an illustration, the technique is applied to group 108 countries on the basis of 120 socioeconomic indicators. The selected countries are also ranked on various social and economic scales by using the Wroclaw Taxonomic Method and the results provide a lot of important information about the structure of world development.

I. Introduction

During the last 20 years a number of multivariate statistical techniques have been applied to the study of economic development. These studies have generally adopted a broadly socioeconomic framework and have attempted to explain the development process in terms of a few mutually independent factors. Berry (1961) was one of the first to apply such techniques to obtain four basic development patterns using 43 socioeconomic indicators from a large number of countries. Adelman and Morris (1965, 1967, 1968a, 1968b, 1969 and 1974) used a number of multivariate techniques to analyse the interaction between various types of social and political change with the level of economic development. Some of these indicators were based upon official statistics issued by national and international agencies, but many were purely qualitative and were based upon subjectively assigned letter-grades reflecting expert opinions.

* The authors wish to acknowledge the assistance of the Division of Computing Research of the C.S.I.R.O., Australia, for the use of their clustering programs. The Wroclaw taxonomic analysis program was obtained from Princeton University.

The extensions of development analysis to include social and political indicators, together with the use of qualitative data, have substantially increased the scope and magnitude of the information to be analysed. With a larger number of variables from developing countries which were previously excluded, owing to a lack of official statistics, it is no longer possible to uncover development patterns by means of simple comparisons of raw percentages. The application of statistical techniques to condense and summarise the extended data base has become a necessity in such cases. The purpose of this paper is to discuss the various techniques which have been used in previous work and to suggest an additional one which, it is argued, should logically precede the other techniques in development analysis.

II. Dimension-Reducing Techniques

Principal Components

The basic method of compressing a large, multidimensional set of data into a set having fewer dimensions is known as principal components analysis. A geometric interpretation is likely to be comprehended more easily, and in order to keep the description within reasonably simple bounds three dimensions will be assumed for the original data set, requiring a reduction of one or at most two dimensions. The procedures involved are of course more general, and their relative advantages are only noticeable with data sets of dimensions much greater than three.

The data can be visualised as points plotted in an ordinary rectangular co-ordinate system. Each point denotes a specific country (so there are as many points as countries) and the values for each of the three indicators can be obtained as distances from the points to the three axes of the co-ordinate system. The entire set of plotted points constitutes the data space, and if the set were enclosed it would form a three-dimensional solid. Although the enclosing surface is likely to be irregular and uneven, the solid may be represented approximately as an ellipsoid.

The first step in principal components analysis is to find the line of longest length which can be passed through the ellipsoid. Having located it, the co-ordinate axes are then rotated so that one axis coincides with the line and is denoted as the major axis. If the points within the data space were projected onto the line it would represent the axis of maximum variance of the points with respect to the means of the indicators.[1] The second step

[1] Generally the original data are expressed in deviations from the respective means so that the origin of the initial co-ordinate system is located at the mean of all three indicator-values. The direction of maximum variance can then be described by a simple rotation.

is to locate the next longest line passing through the ellipsoid which is perpendicular (orthogonal) to the major axis. This (first) minor axis will reflect the maximum variance in the points of all lines which are perpendicular to the previously determined axis. Similarly, a third axis (second minor axis) can be found, and the variances of the three axes will sum to the total variance of the data set.[2] The purpose of the exercise, however, was to obtain less than three dimensions. This can be accomplished by disregarding the last minor axis since the proportion of the total variance it displays is less than that obtained from the others. Hopefully, the amount it adds to the total variance will be negligible. In any case, a representation of the three-dimensional data space by two dimensions introduces a bias, the magnitude of which depends upon the proportion of the total variance which is disregarded. The method of principal components assures only that the two-dimensional space selected is least biased of all possible two-dimensional representations obtained by a rigid rotation of the co-ordinate system.

An additional sacrifice in reducing the dimensions in the manner described is the loss of the original units of measures associated with the initial co-ordinate system. The units of measure in the rotated system can be obtained as a linear combination of the original units, and therefore represent a composite of dollars per capita, number per thousand of population, or whatever comprised the three indicators. Geometrically, the new scale of units will be proportional to cosines of the angle through which the co-ordinate system was rotated. Algebraically, the cosines of the new axes are the normalised characteristic vectors corresponding to the successively smaller characteristic roots[3] of the original mean-corrected data.

Factor Analysis

Although the mechanical aspects of factor analysis are similar to those of principal components, the basic approach is different. The procedure requires a model which specifies that the values of selected indicators are generated by a set of latent or unobservable factors which are smaller in number than the original set of indicators. Such generating factors are

[2] This is possible by virtue of the orthogonality property which ensures that no covariance exists between any two axes.

[3] If the original data are arranged in a matrix and that matrix is replaced by one containing the characteristic roots along the main diagonal (from top left to bottom right) with zeros elsewhere, then the new matrix will preserve two important characteristics of the first – the determinant and the trace. It is in this way that the total variance can be separated into discretely distinct segments. See Morrison (1978) for details.

referred to as common factors. Additionally, the model contains a specific factor which may be unique to the individual indicator. As a consequence, the total variance is not partitioned into successively smaller units, as it is with principal components, but rather, is separated into common-factor variance (communalities) and a specific-factor variance (specificity).

The different approach is motivated by a stronger desire to achieve factors which are capable of interpretation in terms of the original indicators than to minimise the bias introduced by discarding one or more of the original dimensions. A belief in a small set of generating factors is frequently considered to be justified in the social sciences, particularly in cases for which the variables or indicators have a considerable amount of inter-relatedness. In these cases, the bias is viewed more as a specification and estimation error than as a fault in reproducing the original characterisitcs.

Generally a satisfactory interpretation is possible only with a simple structure of factor influences.[4] This can be accomplished by rotating the factor axes or by permitting a departure from the orthogonal relationship by using oblique transformations. This requires that the inherent influence of any one factor falls predominantly on a few indicators, and that any one indicator is not equally influenced by all factors. Ideally, therefore, each factor is viewed as a linear combination of some but not all indicators and each indicator is closely associated with some but not all factors.

Principal Co-ordinate Analysis

Gower (1966) devised the method of co-ordinate analysis for the purpose of extending principal component analysis to include qualitative data and to permit the use of distance measures other than the standard Euclidean distance.[5] The new co-ordinate axes in component analysis are such that the sums of squared Euclidean distances of each point to its projections on the successive axes are minimised. The use of other distance measures[6] will yield solutions that only approximate the component axes, but such a result may be desirable. Other measures may highlight specific attributes of the data set, or may facilitate subsequent analysis with discriminant functions. Additionally, Rohlf (1972) has shown that principal co-ordinate analysis is less sensitive to missing observations.

[4] The concept of a simple structure was proposed by Thurstone (1945).

[5] The Euclidean distance between two points (j and h) in n-dimensional space is

$$\triangle_{jh} = \left[\sum_{i=1}^{n} (X_{ij} - X_{ih})^2\right]^{1/2}$$

where the Xs denote the units along each axis.

[6] Discussed in Sneath and Sokal (1973, pp.121-129).

Multidimensional Scaling

The basic similarity between two countries may be measured in terms of a distance, correlation coefficients or simply as a rank ordering of dissimilarities as evaluated subjectively by expert observers. A decision must then be made to represent the countries in a space of fewer dimensions than the original similarity (or dissimilarity) measures. If the positions of the countries in the reduced space are monotonically related to the observed, ordinal dissimilarities, the method is said to be an ideal method of ordination. Co-ordinates for the reduced space are obtained by iterative procedures which minimise the squared differences between the assigned co-ordinates and the co-ordinates which are necessary to maintain monotonicity. The procedures, referred to as miltidimensional scaling, were initiated by Kruskal (1964) and have been used primarily in cases for which qualitative indicators are predominant. Experiments have indicated[7] that the method appears to be better than principal components in balancing the small differences between similar countries and the large differences between separate groups, each of which contains several, similar countries. In other words, emphasis on the monotonic relationship between countries may be justified in situations involving wide disparities between groups relative to limited disparities within each group.

Discriminant Analysis

Linear discriminant functions represent the specific combination of indicators which optimally separates two groups of countries. The method is generally used to determine the extent of statistical justification for treating the two groups as separate entities, but can also be used to condense the original set of indicators into a smaller set which is linearly related to the former. The function will display maximum variance between groups relative to the combined variance within groups.

Geometrically, the discriminant function can be represented by plotting the group means (one mean for each indicator, the value of which is obtained from the data pertaining to all countries in the group) and describing the plane midway between the points (whose co-ordinates correspond to the respective means) and perpendicular to the line joining the points. The midway position assumes that the frequencies of members of the two groups are equal in terms of the overall population. The length, in discriminant function units, of the line joining the points having mean co-ordinates is

[7] See Rohlf (1970).

the square root of a distance[8] measure known as Mahalanobis' D^2.

A different discriminant function can be calculated for each pair of country groupings. In general the various functions will display different scales and different angles within the data space, though in each case the scale depends only upon three values: the reference score from each group mean and the midway point. The original indicators can therefore be represented by a weighted aggregate derived from the assumed homogeneity of the prescribed groupings. The aggregation procedure is equivalent to orthogonol transformations of the variables to sets in which each variable is independent and has unit variance.

III. Previous Applications

In a major study by Adelman and Morris (1967) a two-part, factor-analysis approach was applied. The first was viewed as a long run analysis, the objective of which was the estimation of a latent factor which 'generates' stages of economic development. Per capita G.N.P. together with 24 social and political variables were compiled from 74 developing countries. The desired factor was the one displaying the highest weight (factor loading) for G.N.P. per capita, and a composite index was derived from the estimated loadings of all variables which acquired significant weight with the chosen factor.[9] All countries were ranked according to their composite index and then separated into three groups. The boundary between groups was established subjectively.[10]

In the second part of the Adelman and Morris study, factor analysis was applied separately to each of the three groups for the purpose of determining the attributes which are common to each group. The results were interpreted as short run effects and were similar to those of previous applications.[11] For countries at the lowest socioeconomic grouping, the nature of the growth process requires both economic and social transformations. Social forces are typically the most important non-economic influence upon economic activity, and political factors do not exert a particularly strong systematic effect on economic growth. In contrast, at the highest of the three levels studied, the political factors are crucial to economic performance, while

[8] It is a Eculidean distance in a new data space in which the original axes are stretched and generally skewed so they (the original axes) may not appear at right angles.

[9] In addition to G.N.P. per capita, the variables comprising the index were ten measures of social change. Adelman and Morris (1967, p.168).

[10] In general, the gaps between groups were greater than those between members of the same set, so that the resulting groupings were reasonably obtained on the basis of the index. However, no objective criteria were expressed for the purpose of determing the boundaries.

[11] Adelman and Morris (1965) and Berry (1961).

social forces have little systematic effect. For countries at the intermediate position on the socioeconomic scale, the results were less conclusive with regard to a uniform set of social or political influences. The process of industrialisation emerged as a crucial element and it conveys both social and political aspects.

In subsequent studies, Adelman and her associates applied other multivariate techniques with results which were more or less consistent with their previous work. Discriminant analysis was used to identify the specific indicators that discriminate among three groups of countries (classified according to their overall development potential).[12] The results indicated that a single discriminant function of only four variables, namely, the degree of improvement in financial institutions, the degree of modernisation of outlook, the extent of leadership commitment to economic development, and the degree of improvement in agricultural productivity, can account for 97 per cent of the variance between groups. In a further study (1968b), they attempted to explain these four variables in terms of the remaining 35 economic and sociopolitical indicators using stepwise regression analysis. In another study, Adelman, Geier and Morris (1969) applied the technique of canonical correlation to estimate the relationship between a set of endogenously chosen variables, identified as instruments, and another set, defined as goals. In a more recent study[13] they used miltidimensional scaling to measure the level of development of 74 countries. Most recently, Syrquin (1978) analysed the groups of variables developed by Adelman and Morris by applying multidimensional scaling to determine the interaction between economic, social and political factors in the development process.

Adelman and Morris studies have been questioned by a number of economists both on methodological and empirical grounds.[14] Several issues have been raised, but only two will be briefly considered here. The multivariate techniques described above have apparent complexities arising from the calculation of characteristic roots and vectors or from iterative methods of solution. These complexities raise concerns that the results may not be intuitively acceptable and may be difficult to interpret, so that the old-fashioned, visual assessment of charts and graphs may be preferable. The second issue arises from the suspicion that the results of the dimension-reducing exercises are highly sensitive to the selection of indicators used in the analysis. The latter concern can be alleviated by substantially extending the number of indicators analysed, but such a decision taxes the ability

[12] Adelman and Morris (1968a).

[13] Adelman and Morris (1974).

[14] See especially Brookins (1970) and Rayner (1970).

of the researcher to recognise overall similarities with the use of charts and graphs.

All but a few of the Adelman-Morris results were easily interpreted and were relatively insensitive to variations in the method of scoring. The studies indicate that careful analysis will minimise the problems described above, but will not eliminate them. With any dimension-reducing technique, disagreement regarding the 'proper' ordering of countries is inevitable. When the estimates conflict with previous judgements, there is a tendency to discount the former as a specific phenomenon of the selected variables, their relative weights or the method of estimating the aggregate measure. Doubts remain as to the ability to achieve the same results with additional modifications to the data set or to the numerical procedures.

Similar problems have occurred for a somewhat longer period in the field of biology and other areas concerned with classification according to morphological detail. In these areas the use of numerical methods has become a necessity, and such methods have developed rapidly. Cluster analysis serves this objective and is intended primarily as a means of putting entities into distinct classes as opposed to arranging them in a continuous spectrum. The latter is referred to as ordination and describes most of the methods discussed previously in this paper. Putting entities into distinct classes assumes only that distinct classes exist; it entails no prior judgements as to why the members of a class are similar or why they are dissimilar to members of another class.

It is argued in this paper that classification or cluster analysis should logically precede ordination. It is easier to scan a small group for common features than a large group, particularly when numerical methods give some assurance that common features exist. After a satisfactory grouping has been obtained, further analysis is facilitated and most clustering programs which are now available for large computers contain a number of diagnostic aids which are either identical to or very similar to the methods previously described.

IV. Clustering Techniques

In the most general sense, cluster analysis is the procedure by which observations are sorted into groups such that the degree of 'natural association' is high among members of the same group and low among members of different groups.[15] Specific clustering procedures are differentiated on the basis of (a) the selected measure of similarity or association, and (b) the sorting strategy used to form the groups. In both cases a wide variety of

[15] For greater detail, see Anderberg (1973).

choices exists, each of which is capable of yielding a slightly different set of groupings.

When clustering individual entities, such as countries, the relationship between them is usually expressed as a distance, such as the Euclidean measure mentioned previously. Pairs of entities can then be formed on the basis of association by proximity within the multidimensional space. Consider an example of three entities or individuals, X_1, X_2 and X_3, each having six measured attributes. To avoid an implicit weighting[16] from differences in scale of the units of measurement, each attribute is standardised to a set of values having zero mean and unit variance. These values are indicated below as X_1, X_2 and X_3 respectively, showing that the larger magnitude of the last attribute is scaled equivalently to the others. The 'distance' between entity 1 and 2 can be found by summing the squared differences between the six pairs of elements under X_1 and X_2 and then taking the root of that sum, yielding $d_{12} = 3.2401$. Since $d_{23} = 4.200$ and $d_{13} = 2.8040$, the first and third entities can be accorded the greatest degree of similarity based

Original data			Standardised data		
X_1	X_2	X_3	X_1	X_2	X_3
4	2	1	1.0911	−0.2182	−0.8729
2	1	2	0.5774	−1.1547	0.5774
3	3	4	−0.5774	−0.5774	1.1547
9	12	7	−0.1325	1.0596	−0.9272
10	14	9	−0.3780	1.1339	−0.7559
120	150	110	−0.3203	1.1209	−0.8006

upon the shortest 'distance'. Group formation starts at the initial level by combining the most similar pair and then adding the next most similar individual, pair or group. When a group has been formed, it is treated as a single entity having attribute values which are based upon a selected measure.

For large numbers of entities, the magnitude of calculations can be exceedingly great. Not surprisingly, the use of high speed computers has

[16] Designating one variable, such as per capita G.N.P., to be more important than any other variable is explicit weighting and is a separate issue.

become indispensable, and the choice of clustering procedure is primarily a choice of computer algorithm. This involves a series of dichotomous decisions as described by Williams (1976, pp.76-83), the most important of which is the choice between hierarchical and nonhierarchical strategies. The former always optimises the route between the entire population and the set of individual entities of which it is composed. The route may be defined by progressive fusions beginning with individuals and ending with the required number of groups. This is known as the agglomerative procedure. Alternatively, the route may be established by progressive divisions, beginning with the population and decomposing it into individuals (divisive procedure). Nonhierarchical methods optimise the intra-group homogeneity, rather than an objective function which expresses the relationship between groups. It thus assures a more satisfactory definition of optimality, but is computationally more difficult.

Given the choice of clustering method, a sorting strategy is required in order to determine the precise way in which the algorithm proceeds. Seven sorting strategies have been described in the literature: (1) nearest neighbour (2) furthest neighbour (3) centroid (4) median (5) group average (6) sum of squares (7) flexible. In nearest neighbour sorting, the distance between two groups is defined as the distance between their closest element, one in each group. Clusters are joined at each stage by the single, shortest link between them and therefore the method is also known as single-linkage clustering. The least desirable feature of this sorting strategy is that it has a tendency to produce groups in a chain-like fashion. Furthest neighbour sorting is the exact antithesis of the nearest neighbour method, in that the distance between two groups is determined on the basis of the most remote pair of elements, one in each group. The method is known as complete-linkage clustering because all entities in a cluster are linked to each other at some maximum distances. In order to avoid the extremes introduced by either nearest neighbour or furthest neighbour strategies, Sokal and Michener (1958) developed group average sorting in which the distance between two groups is defined as the mean of all between groups inter-element distances. It provides relatively weak clustering and as a result did not receive much attention from the taxonomists. In centroid sorting, the distance between two groups is defined as the distance between group centroids (means) in a conventional Euclidean model. Median sorting is a variation of the centroid method in which equal weights are given to all centroids regardless of how many entities are in the respective clusters. Both centroid and median methods produce 'reversals' in the classification and therefore should be avoided when other methods are available. In the sum of squares strategy, which is due to Ward (1963), clustering is based on the minimum sum of squares within clusters resulting from each fusion.

Lance and Williams (1966 and 1967) generalised different sorting strategies into a uniform system and developed a new strategy which is called 'flexible' because the intensity of clustering in this sorting can be varied by altering the value of a single parameter. Starting with three groups, h, i and j, containing n_h, n_i and n_j elements, respectively, with inter-group distances denoted as d_{hi}, d_{hj} and d_{ij}, assume that i and j fuse to form a new group, k, with $n_k = n_i + n_j$ elements. The distance d_{hk}, which is required for later fusion decisions is expressed as a linear equation:

$$d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma |d_{hi} - d_{hj}| \qquad (1)$$

where the parameters α_i, α_j, β and γ determine the nature of the sorting strategy. Flexible strategy is defined by the following constraints:

$$\alpha_i + \alpha_j + \beta = 1; \quad \alpha_i = \alpha_j; \quad \beta < 1; \quad \gamma = 0. \qquad (2)$$

This strategy has all the desirable properties and produces fairly intense clustering. It gives the analyst the opportunity to experiment with the clustering patterns in order to obtain either a more discrete or a more continuous set of fusions. It also helps to overcome the decision-making problem associated with the wide range of sorting strategies by providing a convenient procedure for analysing the effect of small changes.

Since cluster analysis is concerned exclusively with the acquisition of a well differentiated set of groups from a finite population (neither individuals nor attributes are simple random samples in this analysis), conventional sampling theory and hypothesis testing are not applicable. However, to promote the interpretation of clustering results some aids are being used. Hierarchical tree (or dendrogram) illustrates the intergroup pattern, and diagnostic procedures show the contribution of various attributes in successive group formation (or splitting). The systematic relationship between attributes can then be studied by ordination analysis.[17]

The clustering procedure ranks different groups only on the basis of similarity. A 'higher' group is therefore not necessarily more developed. Since we are concerned with the measurement of socioeconomic development, it is desirable to obtain an independent ranking for the various groups of countries. This can be achieved with the use of the Wroclaw taxonomic

[17] The C.S.I.R.O. Division of Computing Research in Australia provides one of the most extensive set of cluster analysis programs in the world. A package of four programs, MULCLAS (gives a hierarchical agglomerative classification), GROUPER (diagnostic program which traces the contribution of different attributes over the entire sequence of fusions), GOWER (ordination program implementing principal co-ordinate analysis), and GOWECOR (diagnostic program showing the extent to which the eigenvectors reflect the original attributes), is held in a permanent library file called TAXON.

method[18] which creates a hypothetical 'ideal' country on the basis of the 'best' value for each indicator used. The aggregation of differences from the 'ideal' country is termed the measure of development (M.D.) and is scaled so that the value ranges between 0 (for the most developed country) and 1 (for the least developed country) in the analysis. The procedure provides an excellent method for international comparison of development and has been applied in a number of UNESCO studies.[19]

V. An Application of Cluster Analysis

Failure of growth-oriented strategies to improve the levels of living of the majority of the population in many underdeveloped countries, and increased evidence of the close interrelations between social and economic factors led to the evolution of the concept of socioeconomic development. The essence of this approach is that development is a single process involving the transformation of a whole social system, of which economic activities and relations are a part, for the achievement of specified goals.

Accordingly, this study extends recent work by UNRISD (1972) by analysing 120 socioeconomic indicators of which 66 are social measures and 54 are economic. A brief description of the data is given in the Appendix. An attempt was made to represent a broad spectrum of countries, with a final selection of 108. Missing observations, which represented about 15 per cent of the total, were estimated by means of a divisive clustering technique.

The selected countries were first ranked in order of levels of development and then clustered on the basis of similarities. Separate comparisons were made for social, economic and socioeconomic indices of development. Finally, the variables were analysed by diagnostic and ordination programs. Assumptions were required as to whether an indicator is a stimulant or retardant to development, the process of which involved subjective evaluation. Out of 120 socioeconomic indicators, we have considered 29 as negative factors (in the sense that all countries or at least most of the countries aim to improve upon the specific aspects conveyed by the respective indicators) and therefore the lower the value of the indicators, the more developed a country is in terms of socioeconomic development.[20]

The Wroclaw taxonomic analysis gave slightly different rankings for the social indicators compared with the economic indicators. The value of Spearman's rank correlation coefficient calculated between the two rankings

[18] Described in Gostkowski (1972) and Harbison (1970).

[19] A list of such studies is given in UNESCO (1974).

[20] These indicators are noted in the Appendix with an asterisk immediately preceeding the indicator number.

TABLE 1

Results of Wroclaw Taxonomic analysis

	Extreme positions		Coefficient of variation of M.D.		
Index	Highest	Lowest	108 Countries	1st 54 Countries	2nd 54 Countries
Social (66 indicators)	Czechoslovakia (M.D. = 0.5418)	Bangladesh (M.D. = 0.9358)	0.1161	0.1140	0.0242
Economic (54 indicators)	Singapore (M.D. = 0.7886)	Socialist Republic of Vietnam (M.D. = 0.9739)	0.0502	0.0467	0.0157
Socioeconomic (120 indicators)	Czechoslovakia (M.D. = 0.6426)	Nepal (M.D. = 0.9544)	0.0905	0.0849	0.0238

Notes: A modified version of the original program (developed by Princeton University) was used here.
M.D. denotes measure of development with the least-developed indicator in each set having a value of unity.

TABLE 2

Ten clusters using 120 socioeconomic indicators

Group No.	Member Countries
15	(Kuwait)
85	(Guyana)
98	(Albania)
181	(Iran, Iraq, Libya, Saudi Arabia)
196	(Bulgaria, Czechoslovakia, G.D.R., Hungary, Poland, Romania, Yugoslavia)
201	(Israel, Puerto Rico)
202	(Hong Kong, Singapore)
204	(Algeria, Brazil, Burma, Colombia, Costa Rica, Dominican Republic, Ecuador, Egypt, El Salvador, Fiji, Guatemala, Honduras, Jordan, Republic of Korea, Malaysia, Mauritius, Mexico, Mongolia, Nicaragua, Paraguay, Peru, Philippines, Sri Lanka, Syria, Thailand, Turkey, Socialist Republic of Vietnam, Yemen Democratic Republic)
205	(Argentina, Chile, Cuba, Cyprus, Greece, Jamaica, Lebanon, Panama, Portugal, Spain, Trinidad-Tobago, Uruguay, Venezuela)
206	(Afghanistan, Angola, Bangladesh, Benin, Bolivia, Botswana, Burundi, Central African Republic, Chad, Congo, Ethiopia, Gabon, Gambia, Ghana, Guinea, Haiti, India, Indonesia, Ivory Coast, Kampuchea, Kenya, Laos, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Morocco, Mozambique, Nepal, Niger, Nigeria, Pakistan, Papua New Guinea, Rwanda, Senegal, Sierra Leone, Somalia, Sudan, Tanzania, Togo, Tunisia, Uganda, Upper Volta, Yemen Arab Republic, Zaire, Zambia)

NOTE: Clustering method and sorting strategy: Hierarchical agglomerative with flexible strategy ($\beta = 0.25$) and standardised Euclidean distance.

is 0.83, which shows a close interrelationship between social and economic factors. Since no country within the sample group is close to the 'ideal' in all indices, it would appear that there is considerable variation within the indicators themselves and that no country is 'best' in all respects. It is evident from Table 1 that there is a relatively greater inter-country disparity on the basis of the social indicators. The results of the comparative analysis also show that the disparity is more pronounced in the case of countries at the higher development levels. For example, the disparity among the rich countries is about 300 per cent more than that among the poor countries in the combined index of socioeconomic development.

In order to determine the 'correct' position of a country it is necessary to compare its rank (based on distance from the 'ideal' country) with its cluster membership (based on similarities or actual distances), since similar countries may have significantly different rankings and vice versa. No generalised statements can therefore be made concerning a country's position by looking at its rank only. There is another important qualification. Clusters can be generated in a number of ways (by using different strategies) and there is no unique 'optimal' set of clusters. Thus, any hypothesis made concerning a country's position and its path to development should be treated tentatively.

Three sets of clusters were obtained with the three sets of indicators and Table 2 reproduces the results obtained with the entire set of socioeconomic indicators. The groups are listed in the order in which they were formed and the numbers used for the identification of groups have only ordinal meaning. The clusters are clearly of different sizes. The first three are single-membered groups, indicating that the respective countries have no close match. Since three other groups have four or less members, the vast majority of countries have no close match. Since three other groups have four or less members, the vast majority of countries are grouped in only four of the ten clusters. Of all the regions, Africa is the most homogeneous, comprising the bulk of the membership of group 206. Although the selected European nations tend to be split apart from the Third World, the grouping is not entirely regional. Three European countries share group 205 with Latin American and Asian countries.

The relationship among the groups is shown in the partial dendrogram in Figure 1, by continuing the fusion process until only one group exists. The ten clusters are indicated by number at the bottom of the diagram. Of these groups, numbers 98 and 205 were the first to merge. On the right-hand side, the four Middle East countries fused with 28 countries, having no dominant region, which are generally treated as moderately underdeveloped countries. The additional fusions are indicated by nodes on the diagram, with decreasing degrees of similarity associated with the higher positions on the chart.

The left-hand side shows countries with a higher level of development and the resulting mergers tend to be less homogeneous than is the case for those on the right-hand side.

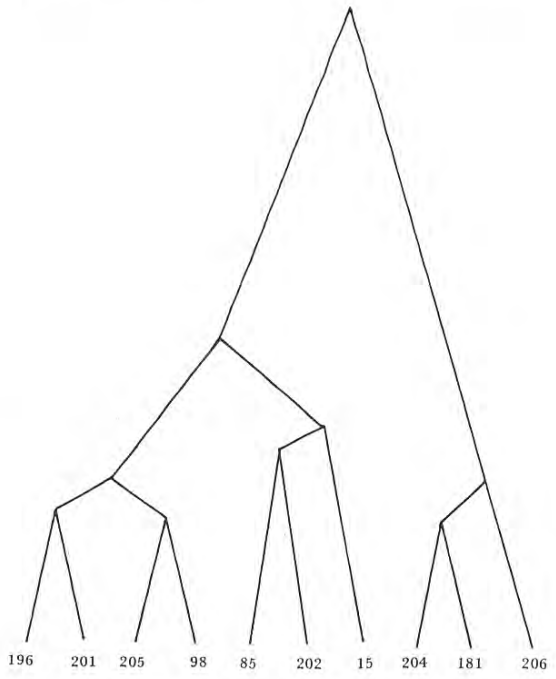

Figure 1

Partial dendrogram with 120 socioeconomic indicators.
(Numbers at bottom indicate reference numbers.)

In order to investigate the variability of the influence attributed to different types of factors, the ten clusters were ranked in order of the level of development by utilising the results of the Wroclaw taxonomic method. These ranks are shown in brackets at the bottom of Figure 2, which reproduces the dendrogram shown previously. Given the ranking of cluster, it is then possible to analyse the relative contribution of the entire set of social indicators compared with the economic indicators, with respect to the differences in the groups. This is accomplished with the diagnostic program GROUPER which gives the contribution of each indicator in the fusion of groups. The indicators which contribute least to the fusion must contribute most to the separation. The latter values were converted into percentages of the total influence attributed to the social factors and are shown in Figure 2 at each split.

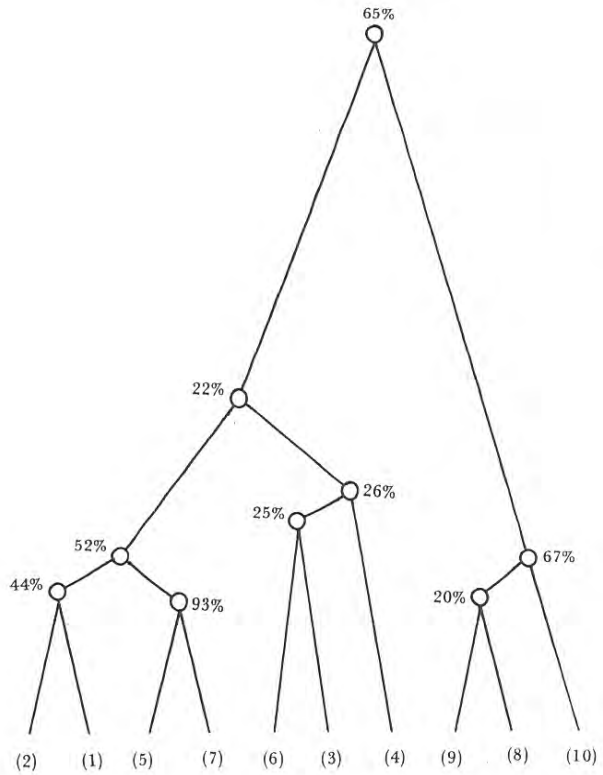

Figure 2

Partial dendrogram with 120 socioeconomic indicators. (Numbers at bottom indicate ranking and percentages show the contribution of the social indicators.)

The general pattern shows the relatively greater importance of the social indicators for the less developed groups (i.e., groups with a low ranking), with more importance attached to the economic indicators for the more developed groups. This result is consistent with the conclusion of other comparative studies referred to earlier in this paper, but the pattern is less uniform than was previously suggested. Specifically, the social indicators contribute only 20 per cent of the division into groups 204 and 181 which rank ninth and eighth, respectively (therefore among the least developed); while social factors contribute 93 per cent of the division into groups 205 and 98 which rank fifth and seventh, respectively (therefore close to the average).

The ordination results show that the first principal co-ordinate accounts for 34 per cent of the total sum of squares and the first seven co-ordinates capture only 61 per cent of total variation. This indicates that the sample space tends to be spherical and cannot be adequately represented by a small number of dimensions. The first principal co-ordinate consists of the following 20 variables in order of importance: (1) per capita dwellings, (2) expectation of life at birth, (3) per capita number of TV sets, (4) consumption of newsprint, (5) percentage of population living on agriculture, (6) per capita circulation of daily newspapers, (7) crude birth rate, (8) index of industrial production, (9) percentage of population that is literate, (10) salaried and wage earners as a percentage of economically active population, (11) percentage of female population that is literate, (12) consumption of newsprint, (13) share of non-agricultural population, (14) per capita electricity consumption, (15) percentage of population engaged in transport, storage and communications, (16) number of telephones per 100 population, (17) use of tractors per 1,000 hectare arable land, (18) percentage of population living in urban areas, (19) G.N.P. per capita, and (20) production of books. The construction of an aggregate index of socioeconomic development with these 20 indicators, however, would not be practical since they reflect only 34 per cent of the total variation in the data.

The clusterings obtained by using social and economic indicators separately show a high degree of association between the two types of variables. The similarity coefficients between the two clusterings, using the Rand[21] criteria, is 0.83 (with a maximum of 1.0). This result is consistent with the earlier result using the Wroclaw analysis. The diagnostic program shows that educational and demographic factors are more important at early stages while cultural and political factors gain in importance at later stages. With the economic indicators, different trends were observed in different regions. For example, transport and communications appear to have played a vital role at early stages in the development of African countries, while international trade was more important at the early phase of development in Asia.

VI. Conclusion

The comparative analysis used in this paper has yielded findings that lead to the formulation of some structural hypothesis. First of all, there is a wide range of disparity among the Third World countries and among regions of those countries. Secondly, the disparity on the social index is considerably greater than that on the economic index. Thirdly, the disparity

[21] W.M. Rand (1971).

among the relatively advanced nations is much greater than that among the poorer nations. Fourthly, the relative importance of various factors varies from stage to stage, so that social factors appear to be more important at early stages and economic factors become prominent at the later stages. Finally, there is a high degree of association (mainly nonlinear) between social and economic factors.

Such comparative studies of socioeconomic development can be useful in several ways. In setting growth targets for a particular country, planners may consider levels of development in a neighbouring country or in a group of comparable countries. Cluster analysis provides the planner with a list of countries which are similar, and ranking procedures can then be used to give relative positions within the cluster and among clusters. The results can also provide some hypothetical planning targets as an initial point in choosing development goals. A target value can be estimated for any indicator for any country by averaging values for all those countries with (1) a relatively higher M.D. and (2) which are within the same cluster. A similar approach can also be used for regional planning within a specific country.

The scope of this study is necessarily limited and further work is needed to carry out similar analyses over time. The analytical techniques used in the study have certain limitations. Like correlation analysis they cannot entirely avoid the problem of collinearity since the selected indicators are clearly related to one another. Additionally, all of the indicators have been treated equally in the sense that the indicators were not weighted in order of *a priori* importance. The second limitation is more controversial, largely because cluster analysts have been reluctant to establish general rules regarding variable selection and weighting. There is limited evidence[22] to suggest that with moderately intercorrelated sets of variables, as generally exists in socioeconomic studies, larger sets tend to even out the *a priori* weights and produce more stable clusters. In other words, individual decisions tend to have less impact on the results with a larger number of socioeconomic indicators. It does not, however, satisfy fully the potential lack of generality in any specific set of estimates. It is unlikely that a 'natural grouping' of developing countries exists, or that a 'proper' ordering of countries can be found to satisfy all development analysis objectives. More experimental evidence is needed, and it is suggested here that the cluster analytic approach is an excellent method of compiling such evidence.

University of New South Wales
National University of Singapore

[22] Zerby and Khan (1980).

Appendix

List of Indicators

A. Demographic, Social and Political

I. Demographic Indicators

1. Population density per sq. km.
*2. Annual rate of growth of population.
3. Percentage of population living in urban areas.
4. Population in localities of 20,000 and over as a % of total population.
*5. Average size of private household.
*6. Crude birth rate per 1,000 population.
*7. Crude death rate per 1,000 population.
*8. Infant mortality rate.
9. Expectation of life at birth (average of male and female).
*10. Dependency ratio (children aged under 15 plus persons aged 65 and over as % of the age group 15-64).
*11. Child dependency ratio (children aged under 15 as % of the age group 15-64).
12. Crude marriage rate per 1,000 population.
*13. Crude divorce rate per 1,000 population.

II. Health and Nutritional Indicators

14. Hospital beds per 10,000 population.
15. Doctors per 10,000 population.
16. Dentists per 10,000 population.
17. Pharmacists per 10,000 population.
18. Nurses per 10,000 population.
19. Midwifery personnel per 10,000 population.
*20. Death rate due to infections and parasitic diseases per 100,000 population.
21. Dietary energy supply per capita daily kilo-calories.
22. Grams protein consumed per capita per day.
23. Total calorie consumption as % of requirement.
24. % contribution of animal protein to total intake of protein.
*25. Consumption of calories derived from cereals and starchy roots as % of total calories consumed.

III. Educational Indicators

26. Percentage of literacy of adult population (15 plus).
27. Percentage of female literacy (15 plus female population).
28. First level enrollment ratio (as % of the age group 5-14).
29. Second level enrollment ratio (as % of the age group 15-19).
30. Third level enrollment ratio (as % of the age group 20-24).

31. % of females in the first level.
32. % of females in the second level.
33. % of females in the third level.
*34. Student/teacher ratio (number of students per one teacher) at the first level.
*35. Student/teacher ratio (number of students per one teacher) at the second level.
*36. Student/teacher ratio (number of students per one teacher) at the third level.
37. Proportion of second level enrollment in vocational education.
38. Proportion of third level enrollment in agricultural courses.
39. Proportion of third level enrollment in medical courses.
40. Proportion of third level enrollment in science and engineering courses.
41. Public expenditure on education as % of GNP.
42. Total stock of scientists, engineers and technicians per 10,000 population.
43. Total stock of scientists, engineers and technicians engaged in research and experimental development per 10,000 population.
44. Expenditure for research and experimental development as % of GNP.
45. Production of books (number of titles by subject per 10,000 population).

IV. Housing Indicators

46. Average size of dwelling (rooms per dwelling).
*47. Average number of persons per room.
48. Dwellings with toilet (and type) as % of all dwellings.
49. Dwellings with piped water as % of all dwellings.
50. Dwellings with electricity as % of all dwellings.
51. Dwellings constructed per 1,000 population.
52. Index number of construction activity (1970 = 100).

V. Cultural Indicators

53. Circulation of daily general-interest newspapers per 1,000 population.
54. Circulation of non-daily general-interest newspapers per 1,000 population.
55. Consumption of newsprint per inhabitant (kilograms).
56. Consumption of printing paper (other than newsprint) and writing paper per inhabitant (kilograms).
57. Cinema seats per 1,000 population.
58. Annual cinema attendance per inhabitant.
59. Number of radio sets per 1,000 population.
60. Number of T.V. sets per 1,000 population.

VI. Political Indicators

*61. Defence expenditure as % of GNP.
*62. Military personnel per 1,000 population.
63. Voting participation: voter turnout as % electorate.
64. Political stability index (average tenure of a national executive/ruling group).

*65. Death from political violence per one million population.
*66. Ethnic and linguistic fractionalization.

B. Economic Indicators

I. Agricultural Indicators

*67. % of total population living on agriculture.
68. Arable land per person in agriculture (hectare/capita).
*69. Percentage contribution of agriculture in G.D.P.
70. Index number of per capita total agricultural production (1961-65=100).
71. Use of tractors per 1,000 hectare arable land.
72. Use of chemical fertilizers per 1,000 hectare arable land (in metric tons).

II. Industry

73. Index of industrial production (general index, 1970 = 100).
74. % of total economically active population engaged in industrial activity.
75. % contribution of industrial activity in G.D.P.
76. % contribution of manufacturing in G.D.P.
77. Per capita energy consumption (total commercial energy) in kilograms per capita.
78. Per capita electricity consumption (total industrial and public) in kwh.
79. Per capita steel consumption (kilograms/capita).

III. Labour

80. % of total population economically active.
81. % of females in total economically active population.
82. Share of non-agricultural population in total economically active population.
83. Salaried and wage-earners as % of total economically active population.
*84. General level of unemployment.
*85. Degree of industrial unrest (total working days lost as a % of total economically active population in industrial activity).

IV. Transport and Communications

86. % of economically active population engaged in transport, storage and communication.
87. Passenger railway kilometers per capita.
88. Railway net ton kilometers per capita.
89. Motor vehicles (passenger and commercial) per 1,000 population.
90. Total road network per 100 population.
91. % of roads paved.
92. Civil aviation: passenger km per capita.

93. Civil aviation: total ton-km per capita.
94. International tourist travel: Tourist receipts per capita (in U.S. dollars).
95. Domestic mail (received and sent) per capita.
96. Foreign mail (received and sent) per capita.
97. Domestic telegram (sent) per capita.
98. Foreign telegram (sent) per capita.
99. Number of telephones per 100 population.

V. International Trade

100. Total value of exports per capita (in U.S. dollars).
101. Total value of imports per capita (in U.S. dollars).
102. Exports as % of GNP.
103. Imports as % of GNP.
104. Average annual growth rate of exports.
*105. % contribution of agriculture in total value of exports.
106. % of contribution of manufacturing in total value of exports.
*107. Exports concentration index.
108. Exports diversification index.
*109. Index of export fluctuations.
110. Terms of trade (average 1971-75, 1970 = 100).

VI. General

111. GNP per capita (at market prices) in U.S. Dollars.
112. GNP at parity prices.
113. Annual growth rate of GNP per capita.
114. Government final consumption expenditure as % of GDP.
115. Private final consumption expenditure as % of GDP.
116. Gross fixed capital formation as % of GDP.
*117. Total per capita receipt of foreign aid (official development assistance from DAC countries through bilateral agreements and through multilateral institutions in U.S. dollars).
*118. Total per capita receipt of foreign capital (direct investment and other long-term private capital in SDR's).
*119. Annual rate of inflation (average for 1971-75).
*120. Gini index of income inequality.

*Denotes negative factors or factors tending to retard socio-economic development.

References

Adelman, I., N. Geier and C. T. Morris, 1969, Instruments and goals in economic development, American Economic Review, 59: 409-426.

Adelman, I., and C. T. Morris, 1965, A factor analysis of the inter-relationship between social and political variables and per capita gross national product, Quarterly Journal of Economics, 79: 555-578.

Adelman, I., and C. T. Morris, 1967, Society, politics and economic development: A quantitative approach, Baltimore: Johns Hopkins Press.

Adelman, I., and C. T. Morris, 1968a, Performance criteria for evaluating economic development potential, Quarterly Journal of Economics, 82: 260-280.

Adelman, I., and C. T. Morris, 1968b, An econometric model of socio-economic and political change in underdeveloped countries, American Economic Review, 58: 1184-1218.

Adelman, I., and C. T. Morris, 1974, The derivation of cardinal scales from ordinal data: An application of multidimensional scaling to measure levels of national development, in W. Sellekaerst, ed., Economic development and planning: Essays in honour of Jan Tinbergen, London: Macmillan, 1-39.

Anderberg, M. R., 1973, Cluster analysis for applications, New York: Academic Press.

Berry, B. J. L., 1961, Basic patterns in economic development, in N. Simsberg, ed., Atlas of Economic Development, Chicago: University of Chicago Press.

Brookins, O. T, 1970, Factor analysis and gross national product: A comment, Quarterly Journal of Economics, 84: 648-650.

Gostkowski, Z., ed., 1972, Towards a system of human resources indicators for less-developed countries, Ossolineum: Polish Academy of Sciences.

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

Harbison, F. H., J. Maruhric and J. R. Resnick, 1970, Quantitative analysis of modernisation and development, Princeton: Princeton University Press.

Khan, M. H., 1981, A comparative study of socioeconomic development: Asia, Asian Profile, 9: 309-319.

Khan, M. H., and J. A. Zerby, 1981, The socioeconomic position of Pakistan in the Third World, Pakistan Development Review, 20: 347-365.

Khan, M. H., and J. A. Zerby, 1982, The socioeconomic positions of India, Pakistan and Bangladesh in the Third World, South East Asian Economic Review, 3: 85-100.

Kruskal, J. B., 1964, Multidimensional scaling by optimising goodness of fit to a nonmetric hypothesis, Psychometrika, 29: 1-27.

Lance, G. N., and W. T. Williams, 1966, A generalised sorting strategy for computer classification, Nature,: 212-218.

Lance, G. N., and W. T. Williams, 1967a, A general theory of classificatory

sorting strategies, I. Hierarchical system, Computer Journal 9: 373-380.

Lance, G. N., and W. T. Williams, 1967b, A general theory of classificatory sorting strategies, II. Clustering systems, Computer Journal, 10: 271-277.

Morrison, D. F., 1978, Multivariate statistical methods, New York: McGraw-Hill.

Rand, W. M., 1971, Objective criteria for the evaluation of clustering methods, Journal of the American Statistical Association, 66: 846-850.

Rayner, A. C., 1970, The use of multivariate analysis in development theory: A critique of the approach adopted by Adelman and Morris, Quarterly Journal of Economics, 84: 639-647.

Rohlf, F. J., 1970, Adaptive hierarchical clustering schemes, Systematic Zoology, 19: 58-83.

Rohlf, F. J., 1972, An empirical comparison of three ordination techniques in numerical taxonomy, Systematic Zoology, 21: 271-280.

Sneath, P. H. A., and R. R. Sokal, 1973, Numerical taxonomy San Francisco: W. H. Freeman.

Sokal, R. R., and C. D. Michener, 1958, A statistical method of evaluating systematic relationships, University of Kansas Science Bulletin, 38: 1409-1438.

Syrquin, M., 1978, The application of multidimensional scaling to the study of economic development, Quarterly Journal of Economics, 92: 621-639.

Thurstone, L. L., 1945, Multiple factor analysis, Chicago: University of Chicago Press.

U.N.E.S.C.O., 1974, Social indicators: Problems of definition and of selection, Report no. 30, Paris.

U.N.R.I.S.D., 1972, Contents and measurement of socioeconomic development, New York: Praeger.

Ward, J. H., 1963, Hierarchical grouping to optimise an objective function, Journal of the American Statistical Association, 58: 236-244.

Williams, W. T., ed., 1976, Pattern analysis in agricultural analysis, Melbourne: C.S.I.R.O. and Elsevier Scientific Publishing Co.

Zerby, J. A., and M. H. Khan, 1980, Selection and weighting of attributes in cluster analysis: An economic development example, School of Economics, Discussion paper, University of New South Wales.