CLUSTER ANALYSIS TECHNIQUES FOR EXPORT MARKET SELECTION

AGRICULTURAL EXPERIMENT STATION • DIVISION OF AGRICULTURE • OKLAHOMA STATE UNIVERSITY
BULLETIN B-796 JUNE 1991

Cluster Analysis Techniques for Export Market Selection

by

Tebogo Seleka

and

David M. Henneberry*

* Graduate Research Assistant and Associate Professor, Department of Agricultural Economics, Oklahoma State University. Research supported by Oklahoma Agricultural Experiment Station Hatch Project H-2102 and Regional Research Project RR2073. Comments of Dr. Brian Adam and Dr. Shida Henneberry are greatly appreciated. March 1991.

TABLE OF CONTENTS

Introduction to Cluster Analysis for Export Markets

The continuous increase in international trade has caused many companies to consider international opportunities. Selecting the right international market is crucial because it determines the company's probability of survival in the international environment. The first challenge faced by potential exporters is to select a group of countries based upon general characteristics. After reducing the number of countries based upon general characteristics, product specific evaluation may be made for final screening. A number of approaches to international market selection have been suggested in previous literature, and various empirical studies have been conducted to select potential markets for specific products. These approaches can be qualitative, quantitative or both.

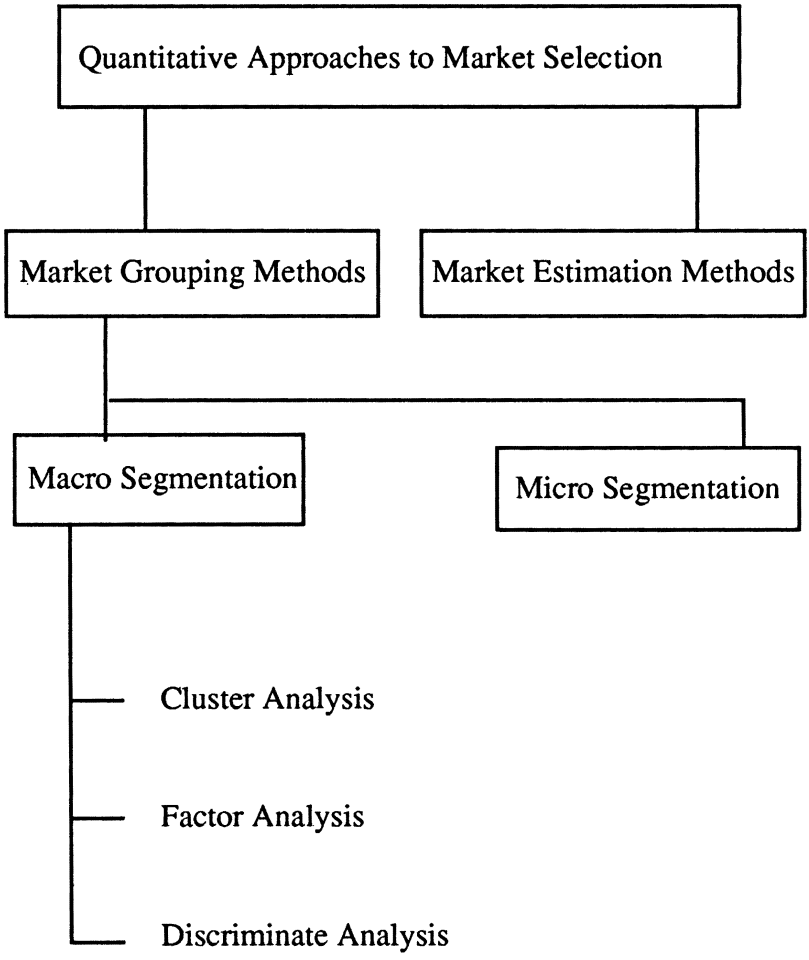
Quantitative approaches to market selection may be categorized into two broad groups: market grouping methods and market estimation methods (Papadopolous and Denis, 1988). Papadopolous and Denis (1988) further divide market grouping methods into two categories: macro segmentation techniques and micro segmentation techniques. Cluster analysis, factor analysis, and discriminate analysis fall within the category of macro segmentation. The classification of quantitative approaches is depicted by figure 1. This paper is limited to a discussion of cluster analysis which is among the most common of the macro segmentation methodologies.

A number of statistical techniques used to group objects, persons, stimuli, or concepts into homogeneous classes on the basis of their similarity are referred to as cluster analysis (Lorr, 1983). In the international marketing literature, cluster analysis has been used to segment international markets to aid companies in making marketing decisions. Usually a company intending to select an international market for entry faces a problem of evaluating the many countries of the world based upon varying characteristics. Cluster analysis can therefore be used to group countries so that those with similar characteristics are placed into a single group. Thus, the objective of cluster analysis is to group countries into clusters for the purpose of selecting a group or fewer countries the researcher wishes to investigate further. A further evaluation of countries in groups with desirable characteristics (using more direct techniques) is then made possible. Cluster analysis can therefore be used for screening purposes.

A cluster analysis study of international markets involves a sequence of the following steps:

1. Choice of countries and variables.
2. Assembling the data into a matrix.
3. Dimensional analysis of variables.
4. Conversion of variables into comparable units.
5. Selection of an appropriate similarity index and assessment of similarity between pairs of country profiles.

Figure 1: Quantitative Approaches to Market Selection



6 Selection and application of a clustering algorithm to the similarity matrix.

7. Computation of the mean profiles of each cluster and interpretation of results.

These steps are discussed in this paper. Emphasis has been put on the discussion of clustering techniques to give the reader an adequate understanding of the different techniques. Although past empirical work has shown that hierarchical clustering techniques are not entirely acceptable in international market selection, their discussion in this paper will help the reader understand cluster analysis. The paper also includes a section on determining the number of clusters. An example from previous studies is given as an illustration of the process of cluster analysis. Figure 2 illustrates the sequence of cluster analysis.

It is not the purpose of this paper to present new theoretical or empirical results. Instead, it is to clarify and provide further detail on some of the more critical areas encountered by international market researchers using the technique. Few examples exist within the available literature which fully explain the process of cluster analysis. When combined with the sources identified in the bibliography this paper is a useful reference for applied market analysts attempting to select export markets.

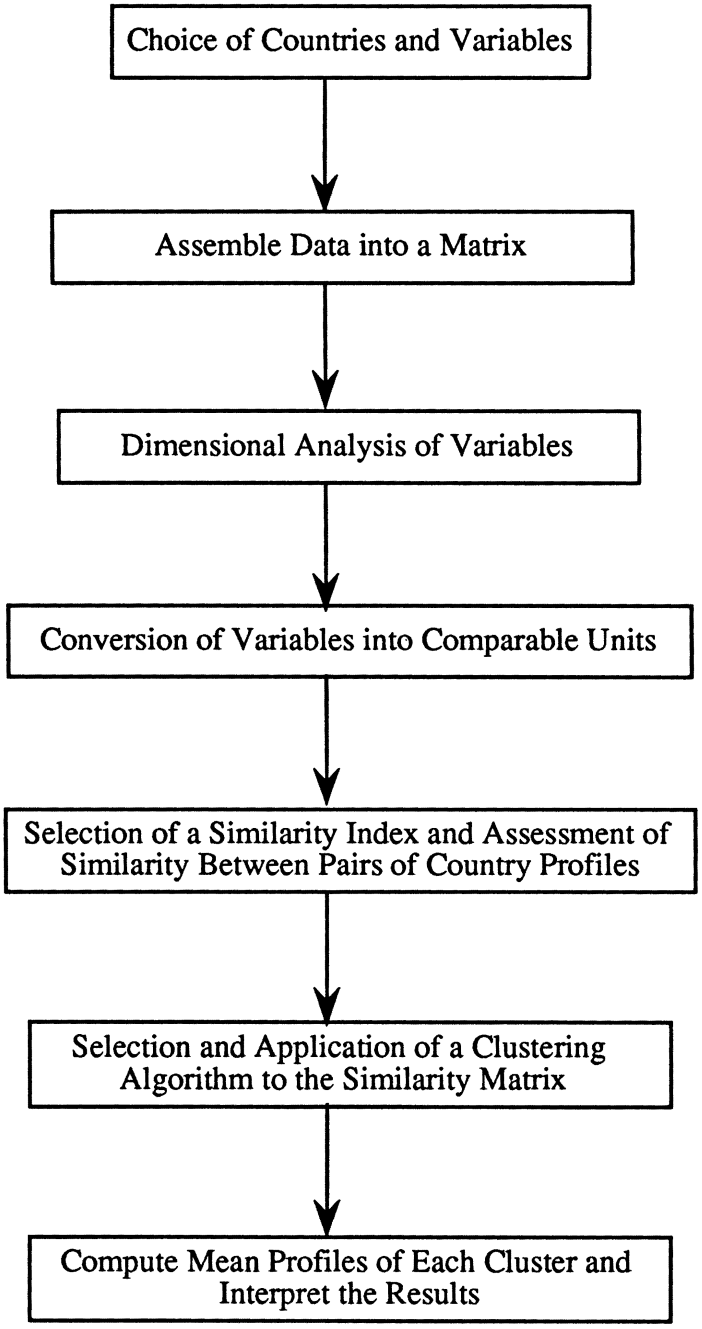
Choice of Countries and Variables

The first step in a cluster analysis of international markets is the choice of countries and variables by which to base the segmentation. When data are available on all variables of interest, it is advisable to include all countries in a cluster analysis study. Sampling is not necessary since the number of countries is small. The results of a cluster analysis of international markets cannot be applied to countries which were excluded from the original analysis. Therefore, starting with the largest available set of countries improves the usefulness of the final results.

The choice of variables has no mathematical or statistical guidelines. It is a reflection of the analyst's judgement of relevance for the purpose of the classification (Everitt, 1974). When selecting variables, the analyst needs to be product specific (Doyle and Gidengil, 1977). Although the need to be product specific when choosing variables has been stressed, general indicators of international business prospects have to be included in the analysis for a general comparison of the potential within different countries.

A company intending to segment international markets for a value added agricultural product may include variables which assess economic environment, import demand conditions, political conditions, and trade policy. The national economic environment may be assessed by such macroeconomic indicators as economic size (GNP or GDP), income levels (GNP per capita), real growth rate (percent change in per capita GNP), external dependance (ratio of foreign trade to GNP or/and ratio of foreign debt service to foreign exchange earnings), price

Figure 2: The Sequence of Cluster Analysis



levels (inflation), exchange rates and balance of payments (foreign exchange earnings minus foreign exchange expenditure). Assessment of import demand conditions may be general or product specific. Product specific variables may include annual imports of the product, current imports from the United States, imports from the United States as a percentage of total imports, and net imports (imports minus exports). To assess the political environment the company may include variables such as type of political system, country's relationship with the US, and government stability. Trade policy variables could be the presence of trade barriers and the extent of the barriers. Table 1 presents variables from previous empirical studies. The table is meant to give the reader an idea of variables that past researchers have used (listed variables do not necessarily reflect the authors' suggestions).

Assembling the Data into a Matrix

Following the selection of countries and variables, the analyst's task is to collect the data on all variables for all the countries/markets to be clustered. The collected data is then assembled into a matrix in which the rows represent countries and the columns represent variables. The raw data consists of an MxN matrix of measurements X, where

$$X = \begin{bmatrix} x_{11} & x_{12} \cdots x_{1N} \\ x_{21} & x_{22} \cdots x_{2N} \\ \vdots & \\ x_{M1} & x_{M2} \cdots x_{MN} \end{bmatrix}$$

and in which x_{ij} is the data on the jth variable for the ith country. Thus, the matrix is a set of M countries on which N variables have been recorded.

Dimensional Analysis of Variables

Often, the variables of interest are numerous and highly correlated. It is difficult to conduct and interpret an analysis that was carried out across a large number of variables which at the same time happen to be correlated. A dimensional analysis of variables is therefore required to reduce the number, redundancy and complexity of variables. A dimensional analysis places inter-correlated variables into a single cluster. Each cluster consists of a collinear subset of variables that are nearly independent from the definers of other variables

Table 1. Variables used in previous cluster analysis market studies

Variable	Author: Sethi	Author: Doyle and Gidengil	Author: Day, Fox and Huszagh
Demographic			
Past five year percentage increase in population		x	
Percentage of population in Agriculture	x		
Population	x	x	
Population density		x	
Urbanization (number of cities with over 100,000 population)	x		
Urbanization (percentage of population living in urban areas)			x
Education			
College, university, and professional school education per capita in population 15-64 years old	x		
Illiteracy among adults of 15 years and older	x		
Per capita school enrollment in population 15-19 years old	x		
Percentage of school enrollment of 14 year olds		x	
School enrollment per capita in population 5-14 years old	x		
University enrollment per 1,000 population		x	x
General Development Indicators			
Civil aviation passengers per kilometer	x	x	
Daily newspaper copies sold per 1,000 population		x	
Newspaper circulation	x		
Number of newspapers	x		
Number of passenger cars (per capita or per 1,000 population)	x	x	x
Number of radio sets (per capita or per 1,000 population)	x	x	x
Number of telephones (per capita or per 100 population)	x	x	
Number of television sets (per capita or per 1,000 population)	x	x	x
Total scheduled air passenger kilometers in 1980			x
Health			
Life expectancy	x		
Male life expectancy			x
Number of hospital beds	x		
Number of physicians per capita	x		
Number of population per hospital bed		x	x
Number of population per physician		x	x
Income			
GDP or GNP per capita	x	x	x
Personal income	x		
Industrialization			
Cement production in kg. per capita		x	
Consumption of printing and writing paper (kg. per capita)		x	
Electric energy production	x	x	
Electrical capacity (installed)			x
Energy consumption	x	x	x
Number of commercial vehicles per 1,000 population		x	
Steel consumption (kg. per capita)		x	
Macroeconomic Indicators			
Agriculture as a percentage of GDP		x	
Consumer price index	x		x
Exchange Rate			x
Government spending as a percentage of GDP			x
Manufacturing as a percentage of domestic product		x	
Wholesale and retail trade as a percentage of GDP		x	
Trade			
Civil aviation freight in tons per km.	x	x	
Exports as a percentage of imports			x
Total exports as a percentage of GNP or GDP	x	x	x
Total imports as a percentage of GNP or GDP	x	x	x

(Sethi, 1971). Two models are available for reducing redundancy of variables. These are principal component analysis and factor analysis. Readers may refer to numerous other sources for specific information on principal component and factor analysis.

The result of a dimensional analysis of variables is an Mxk matrix where k represents the number of variable clusters which are independent from each other. The resulting data matrix is given by:

Variable Cluster

Country	1	2	•	•	•	k
1	x_{11}	x_{12}	•	•	•	x_{1k}
2	x_{21}	x_{22}	•	•	•	x_{2k}
•	•	•	•	•	•	•
•	•	•	•	•	•	•
•	•	•	•	•	•	•
M	x_{M1}	x_{M2}	•	•	•	x_{Mk}

where each column represents measures of a single attribute over the different countries, and each row defines the profile of an individual country across attributes. K is usually smaller than the original n variables.

Conversion of Variables into Comparable Units

After the variables have been reduced to a manageable number through dimensional analysis, they need to be standardized to make them comparable. Variables expressed in raw data often vary in dispersion and in units of measurements, and analysis without conversion to a common metric (such as a standard score) may give implausible weights to some of them (Lorr, 1988). Variables are usually expressed in different scales. The four scales likely to be encountered in cluster analysis are nominal, ordinal, interval and ratio.

Variables on nominal and ordinal scales are often referred to as **qualitative** variables. A **nominal** scale assigns a meaningful measure of difference between two countries. In a nominal scale, a set of countries is partitioned into mutually exclusive subsets, meaning that members of a subset must be equivalent on the property being scaled. If two countries A and B are compared based on variable X, then either $X_A = X_B$ or $X_A \neq X_B$. An ordinal scale reflects the ordering of countries. In addition to distinguishing countries as with a nominal scale, ordinal scales distinguish between $X_A > X_B$ and $X_A < X_B$. It does not matter what numbers are assigned to ordered objects (countries),

so long as the higher numbers are assigned to members of a class which is "greater than" (Lorr, 1983).

Variables on interval and ratio scales are often called **quantitative** variables. An **interval** scale assigns a meaningful measure of difference between two countries. In addition to saying that $X_A > X_B$ one is able to say that country A is $X_A - X_B$ units different from country B. In an interval scale, equal units of measurement are used and a linear transformation is used because it preserves both the ordering of countries and the relative difference between them (Lorr, 1983). A linear transformation may be done by converting the raw data to standard scores with a mean of zero and a standard deviation of one. A **ratio** scale has the properties of an interval scale. In a ratio scale one is able to say a is X_a/X_b times greater than b. Ratio scales can be treated as interval scales.

Given a data set with different scales, the usual procedure is to standardize all dimensional attributes so that each scale has a mean of zero and a standard deviation of one. Properties not capable of further subdivision are treated as present or absent and then standardized (Lorr, 1983). Anderberg (1973) further explained the need to convert a data set to one scale if different scales are present. As an example, the analyst may wish to convert all the scales in a data set to interval. The reader wishing to pursue the subject of scale conversions is referred to Anderberg. Although some elementary statistics books state that different statistical procedures require the use of specific measurement scales (nominal, ordinal, interval, or ratio), an opposing view is held by many researchers who contend that indices of similarity such as distance and correlation may be computed even though the presence of interval scales cannot be demonstrated (Lorr, 1983). This latter view suggests that similarity measures such as distance and correlation coefficients can still be applied without the need to convert the scales to interval scale.

Selection of a Similarity Index and Assessment of Similarity

A similarity index is used to determine how similar markets are to one another. Some of the available literature suggests that the type of attributes determine the choice of a measure of similarity or difference. Interval scales may require the use of distance functions or correlation coefficients while nominal and ordinal scales may require the use of matching coefficients and ordinal rank respectively (Lorr, 1983). As was mentioned earlier, many researchers feel that distance measures and correlation coefficients may be used even though the presence of an interval scale cannot be demonstrated. The upcoming discussion presents distance functions and correlation coefficients as alternative measures of similarity between countries.

DISTANCE MEASURE

The difference between the attributes of any two countries can be thought of as the distance between data profiles for those countries. This distance is

computed as a square root of the sum of square difference in scores between the two countries over k variables. If the variables are represented by orthogonal (right-angled) axes, the distance can be calculated using the pythagorean theorem:

$$D^2_{ij} = \sum_{h=1}^{k} (x_{ih} - x_{jh})^2$$

The distance between country i and j is therefore defined as the square root of D^2_{ij}. As an example, suppose that the researcher wants to measure the distance between country 1 and 2 from the previous data matrix. From the data matrix the score profile for country 1 is (x_{11} x_{12} ...x_{1k}) and for country 2 is (x_{21} x_{22} ...x_{2k}). The distance between country 1 and 2 is evaluated by

$$D_{12} = \{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + \ldots + (x_{1k} - x_{2k})^2\}^{1/2}$$

Several distance measures have been suggested as a means of measuring similarity between profiles. These include the euclidean metric, the absolute or city block metric, and the Minkowski metric.

A distance function d(x,y) of pairs of points of a set E is a metric for E if it satisfies the following conditions.

i) $d(x,y) \geq 0$;

ii) $d(x,y) = 0$ if $x = y$;

iii) $d(x,y) = d(y,x)$;

iv) $d(x,z) + d(y,z) \geq d(x,y)$

The Euclidean metric, which is identical to the Pythagorean theorem, is probably the most commonly used and is defined as

$$d_{ij} = \left[\sum_{h=1}^{k} (x_{ih} - x_{jh})^2 \right]^{1/2}$$

where d_{ij} is the distance between country i and j, x_{ih} and x_{jh} are the values of the hth variable for the ith and the jth countries respectively.

The absolute or city block metric is defined as

$$d_{ij} = \sum_{h=1}^{k} |x_{ih} - x_{jh}|$$

and is the sum of the absolute values of the difference between countries for each profile element.

The Minkowski metric is a combination of the euclidean and the city block metrics. It is defined as

$$d_{ij} = \left[\sum_{h=1}^{k} |x_{ih} - x_{jh}|^r \right]^{1/r}$$

As can be seen from above, d_{ij} becomes the euclidean metric if r = 2, and becomes the city block metric if r=1.

Using distance as a measure of similarity (using raw scores) preserves information on the scatter, elevation and shape of the profiles* . The limitation of using raw scores (as mentioned earlier) is that implausible weights are often assigned to some of the variables. A remedy to this is to standardize variables. However, if the variables are standardized to deviation scores, information on elevation and scatter is lost.

Distance measurements are assembled into a distance matrix (D) where

$$D = \begin{bmatrix} d_{11} & d_{11} & \cdots & d_{1k} \\ d_{21} & d_{22} & \cdots & d_{2k} \\ \vdots & \vdots & & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{Nk} \end{bmatrix}$$

and d_{ij} is the distance between country i and j obtained by using the selected distance measure (Euclidean, etc.).

CORRELATION COEFFICIENTS

Correlation coefficients can also be used to measure similarity between profiles. These include (among others) the Pearson product-moment correlation, Cohen's coefficient and the Congruency coefficient.

---

* Elevation is the "mean of all scores for" a country, scatter is "the square root of the sum of squares of the entity's deviation score around its own mean", and "shape is the information remaining in the score set after removing elevation and equalizing scatter" (Lorr, 1983).

The Pearson product moment correlation is the most widely used and is defined as

$$Q_{ij} = \frac{\sum(x_{ih} - \bar{x}_i)(x_{jh} - \bar{x}_j)}{\left[\sum(x_{ih} - \bar{x}_i)^2\right]^{1/2}\left[\sum(x_{jh} - \bar{x}_j)^2\right]^{1/2}}$$

where the countries i and j are correlated and $\bar{x}_i$ and $\bar{x}_j$ represent the country means and the denominator terms represent the scatter. Subtracting the mean and dividing by scatter leads to a loss of information on both elevation and scatter, so the Pearson product moment correlation is based only on the shapes of the profiles.

Cohen's coefficient is defined as

$$r_c = \frac{\sum (x_{ih} - m)(x_{jh} - m)}{\left[\sum(x_{ih} - m)^2\right]^{1/2}\left[\sum(x_{jh} - m)^2\right]^{1/2}}$$

where m is the neutral point of any number of variables determined by addition of K new scales, each a reflection of the original scales. When m represents the actual means of x_{ih} and x_{jh}, then r_c becomes Q_{ij} and Cohen's Coefficient is the Pearson product moment correlation. A discussion of Cohen's coefficients is contained in Lorr (1983).

The Congruency coefficient is defined as

$$c = \frac{\sum x_{ih}x_{jh}}{\left[\sum x^2_{ih} \sum x^2_{jh}\right]^{1\backslash 2}}$$

When using this coefficient, information on elevation is retained since the country means are not subtracted while information regarding scatter is lost because of transformation to unit length.

The computed correlation coefficients are assembled in a similarity matrix S where

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1k} \\ s_{21} & s_{22} & \cdots & s_{2k} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ s_{N1} & s_{N2} & \cdots & s_{Nk} \end{bmatrix}$$

Selection and Application of a Clustering Algorithm

After the analyst decides on the similarity index (distance or correlation coefficients), measures similarity, and assembles the similarity measures (distance or correlation) into a matrix, the next step is to decide on the clustering method. This is followed by applying a clustering algorithm to the similarity matrix. Clustering techniques may be categorized into hierarchical techniques, optimization-partitioning techniques, density or mode seeking techniques, and clumping techniques. These techniques are discussed below, except for clumping techniques which are concerned with an overlap between clusters and are not individually discussed. Figure 3 illustrates the alternative cluster analysis techniques for international market selection.

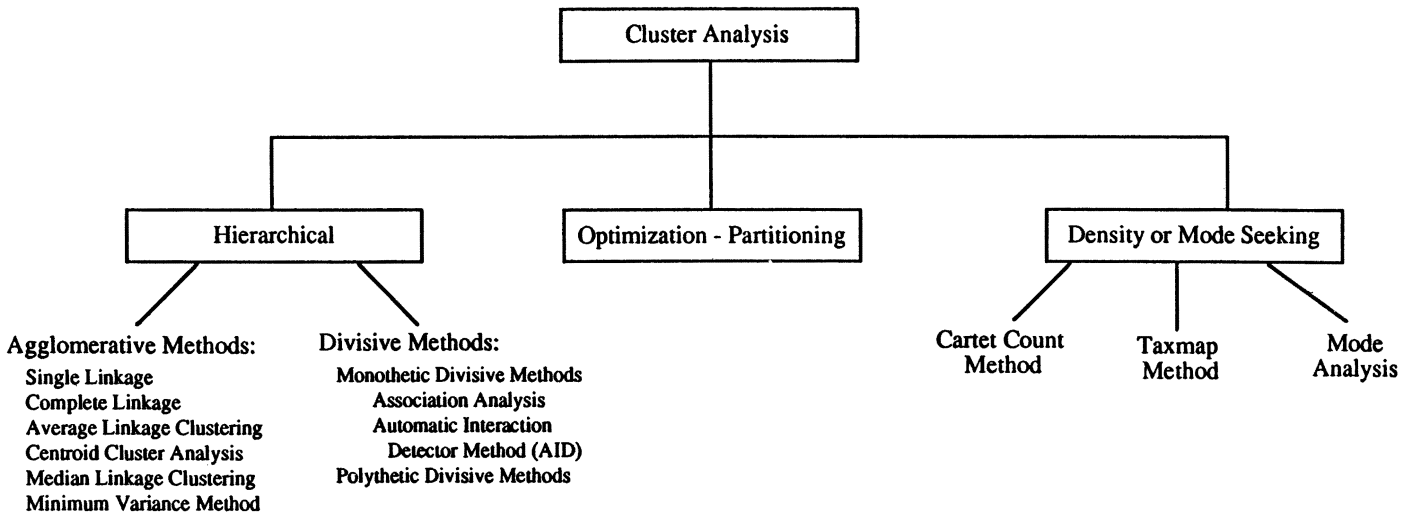
HIERARCHICAL CLUSTERING TECHNIQUES

Hierarchical clustering techniques construct a tree which depicts specified relationships among the countries, based on the similarity matrix (Anderberg, 1973). These techniques are subdivided into the agglomerative and divisive methods. Agglomerative methods develop a successive grouping of N countries into groups and divisive methods divide a set of N countries into finer partitions (Everitt, 1974).

Agglomerative Methods

The agglomerative methods begin fusion or merging of countries by the computation of the inter-country distance matrix from which the closest countries are merged into a single group, until a tree has been formed. These methods are classified into the single linkage method, complete linkage method,

Figure 3: Alternative Cluster Analysis Techniques for International Market Selection



average linkage clustering, centroid cluster analysis, median linkage clustering and minimum variance methods. The general distance formula for the agglomerative methods is

$$d_{hk} = Ad_{hi} + Bd_{hj} + Cd_{ij} + D|d_{hi} - d_{hj}|$$

where k is the index for the new group obtained by merging groups i and j; d is the measure of similarity ; h is a group other than the fused group; and A, B, C, and D represent parameters whose values vary with the method as shown below (Lorr, 1983; Mojena, 1977).

Method	A	B	C	D
Single Linkage	1/2	1/2	0	-1/2
Complete Linkage	1/2	1/2	0	1/2
Average Linkage	$n_i / n_i + n_j$	$n_j / n_i + n_j$	0	0
Centroid	$n_i / n_i + n_j$	$n_j / n_i + n_j$	$\frac{-n_i n_j}{(n_i + n_j)^2}$	0
Median	1/2	1/2	-1/4	0
Minimum Variance	$\frac{n_h + n_i}{n_h + n_k}$	$\frac{n_h + n_j}{n_h + n_k}$	$-n_h / n_h + n_k$	0

The agglomerative methods are now discussed individually.

Single Linkage. The single linkage, which is also known as the nearest neighbor method, can be used with both correlation coefficients and distance measures. The clustering procedure begins with the calculation of a distance or correlation matrix using the selected similarity measure.

When distance measures are used, the closest or most similar countries in a distance matrix are combined into a single group. The distance between groups is that between their closest members. This distance is defined as

$$d_{hk} = \min(d_{ik}, d_{jk})$$

where d_{hk} is the distance between the closest members of group h and k (Lorr, 1983). Alternatively, the distance can be derived from the general formula to yield

$$d_{hk} = 1/2(d_{hi}) + 1/2(d_{hj}) - 1/2|d_{hi}-d_{hj}|$$

To illustrate the single linkage method, Everitt (1974) began with a 5 country D1 matrix as follows:

$$D1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.0 & 2.0 & 6.0 & 10.0 & 9.0 \\ 2.0 & 0.0 & 5.0 & 9.0 & 8.0 \\ 6.0 & 5.0 & 0.0 & 4.0 & 5.0 \\ 10.0 & 9.0 & 4.0 & 0.0 & 3.0 \\ 9.0 & 8.0 & 5.0 & 3.0 & 0.0 \end{bmatrix} \end{array}$$

Examination of the matrix indicates that export markets 1 and 2 are the closest to each other with a distance of 2.0. These countries are fused to form group (12). The next procedure is to calculate another distance matrix (D2). The new off-diagonal elements of the matrix are obtained as follows:

$$d_{3(12)} = \min (d_{13}, d_{23}) = d_{23} = 5.0$$

$$d_{4(12)} = \min (d_{14}, d_{24}) = d_{24} = 9.0$$

$$d_{5(12)} = \min (d_{15}, d_{25}) = d_{25} = 8.0$$

The same results may be obtained by using the general formula.

$$d_{3(12)} = 1/2(d_{31}) + 1/2(d_{32}) - 1/2|d_{31} - d_{32}|$$

$$= 1/2(6.0) + 1/2(5.0) - 1/2|6.0 - 5.0|$$

$$= 5.0$$

$$d_{4(12)} = 1/2(d_{41}) + 1/2(d_{42}) - 1/2|d_{41} - d_{42}|$$

$$= 1/2(10) + 1/2(9.0) - 1/2|10 - 9.0|$$

$$= 9.0$$

$$d_{5(12)} = 1/2(d_{51}) + 1/2(d_{52}) - 1/2|d_{51} - d_{52}|$$

$$= 1/2(9.0) + 1/2(8.0) - 1/2|9.0 - 8.0|$$

$$= 8.0$$

Given the new calculations and the remaining elements of the D1 matrix, the new distance matrix (D2) is as follows:

$$D2 = \begin{array}{c} \\ (12) \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (12) & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.0 & 5.0 & 9.0 & 8.0 \\ 5.0 & 0.0 & 4.0 & 5.0 \\ 9.0 & 4.0 & 0.0 & 3.0 \\ 8.0 & 5.0 & 3.0 & 0.0 \end{bmatrix} \end{array}$$

The closest groups in the D2 matrix are 4 and 5 with a distance of 3.0. These are combined to form group (45), resulting in the calculation of the new distances as follows:

$$d_{(12)(45)} = \min (d_{14},d_{15},d_{24},d_{25}) = d_{25} = 8.0$$

$$d_{3(45)} = \min (d_{34},d_{35}) = d_{34} = 4.0$$

Alternatively, the general formula can be used to yield identical results as follows:

$$d_{(12)(45)} = 1/2(d_{(12)4}) + 1/2(d_{(12)5}) - 1/2|d_{(12)4} - d_{(12)5}|$$

$$= 1/2(9.0) + 1/2(8.0) - 1/2|9.0 - 8.0|$$

$$= 8.0$$

$$d_{3(45)} = 1/2(d_{34}) + 1/2(d_{35}) - 1/2|d_{34} - d_{35}|$$

$$= 1/2(4.0) + 1/2(5.0) - 1/2|4 - 5|$$

$$= 4.0$$

The new distances are assembled into another matrix (D3) given below:

$$D_3 = \begin{array}{c} \\ (12) \\ 3 \\ (45) \end{array} \begin{array}{c} \begin{array}{ccc} (12) & 3 & (45) \end{array} \\ \begin{bmatrix} 0.0 & 5.0 & 8.0 \\ 5.0 & 0.0 & 4.0 \\ 8.0 & 4.0 & 0.0 \end{bmatrix} \end{array}$$

The closest groups in D3 are (45) and 3 with a distance of 4.0. These are combined into a single group. The final step is fuse the group consisting of countries 3, 4, and 5 with that of 1 and 2 into a single group.

When the measure of similarity is correlation, the analyst has to find

$$r_{hk} = \max(r_{ik},r_{jk})$$

where r_{hk} is the degree of similarity between the two most similar countries in groups h and k (Lorr, 1983). The procedure for fusing groups is similar to the one described above, except that r_{hk} will be used instead of d_{hk} and that the highly correlated profiles are for the closest countries.

Complete Linkage. The complete linkage method, also known as the furthest neighbor method uses the distance between the most remote pairs of countries in clusters as the distance between them. The distance between clusters is therefore

$$d_{hk} = \max(d_{ik}, d_{jk})$$

where d_{hk} is the distance between the furthest countries in groups h and k. The alternative formula is

$$d_{hk} = 1/2(d_{hi}) + 1/2(d_{hj}) + 1/2|d_{hi} - d_{hj}|$$

The method begins by the assessment of a D1 matrix (presented earlier under the single linkage method section). Proceeding in an identical manner, it is apparent from the D1 matrix that the closest countries are 1 and 2 with a distance of 2.0. These are fused to form group (12). The new group then calls for the computation of the following elements for the D2 matrix:

$$d_{3(12)} = \max(d_{13}, d_{23}) = d_{13} = 6.0$$

$$d_{4(12)} = \max(d_{14}, d_{24}) = d_{14} = 10.0$$

$$d_{5(12)} = \max(d_{15}, d_{25}) = d_{15} = 9.0$$

Identical results can be obtained as follows:

$$d_{3(12)} = 1/2(d_{31}) + 1/2(d_{32}) + 1/2|d_{31} - d_{32}|$$

$$= 1/2(6.0) + 1/2(5.0) + 1/2|6.0 - 5.0|$$

$$= 6.0$$

$$d_{4(12)} = 1/2(d_{41}) + 1/2(d_{42}) + 1/2|d_{41} - d_{42}|$$

$$= 1/2(10) + 1/2(9.0) + 1/2|10.0 - 9.0|$$

$$= 10.0$$

$$d_{5(12)} = 1/2(d_{51}) + 1/2(d_{52}) + 1/2|d_{51} - d_{52}|$$

$$= 1/2(9.0) + 1/2(8.0) + 1/2|9.0 - 8.0)$$

$$= 9.0$$

This will result in the following matrix:

$$D2 = \begin{array}{c} \\ (12) \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (12) & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.0 & 6.0 & 10.0 & 9.0 \\ 6.0 & 0.0 & 4.0 & 5.0 \\ 10.0 & 4.0 & 0.0 & 3.0 \\ 9.0 & 5.0 & 3.0 & 0.0 \end{bmatrix} \end{array}$$

From the D2 matrix the closest groups are 4 and 5 with a distance of 3.0. The new distances are calculated as follows:

$$d_{(12)(45)} = \max(d_{14},d_{15},d_{24},d_{25}) = d_{14} = 10.0$$

$$d_{3(45)} = \max(d_{34},d_{35}) = d_{35} = 5.0$$

or

$$d_{(12)(45)} = 1/2(d_{(12)4}) + 1/2(d_{(12)5}) + 1/2|d_{(12)4} - d_{(12)5}|$$

$$= 1/2(10.0) + 1/2(9.0) + 1/2|10.0 - 9.0|$$

$$= 10.0$$

$$d_{3(45)} = 1/2(d_{34}) + 1/2(d_{35}) + 1/2|d_{34} - d_{35}|$$

$$= 1/2(4.0) + 1/2(5.0) + 1/2|4.0 - 5.0|$$

$$= 5.0$$

The distances are now assembled into a D3 matrix:

$$D_3 = \begin{array}{c} \\ (12) \\ 3 \\ (45) \end{array} \begin{array}{c} \begin{array}{ccc} (12) & 3 & (45) \end{array} \\ \begin{bmatrix} 0.0 & 6.0 & 10.0 \\ 6.0 & 0.0 & 5.0 \\ 10.0 & 5.0 & 0.0 \end{bmatrix} \end{array}$$

The next step is to fuse (45) with 3 into a single group. Finally, the group consisting of countries 1 and 2 is fused to that of countries 3, 4, and 5.

If the analyst decides to use correlation as a measure of similarity, then the formula becomes:

$$r_{hk} = \min(r_{ik},r_{jk})$$

where r_{hk} is the degree of similarity between the two most remote countries in groups h and k.

Average Linkage Clustering With the average linkage method, the distance between groups is defined as the average of the distance between all pairs of countries in the groups (Lorr, 1983). The formula for average linkage clustering is

$$d_{hk} = (n_i/n_k)d_{hi} + (n_j/n_k)d_{hj}$$

where $n_k = n_i + n_j$.

Everitt (1974) pointed out that the procedure can be used with both correlation coefficients and distance measures if the concept of an average measure is accepted. Furthermore, Everitt confirmed the findings by Lance and William (1967) that since the concept of average correlation coefficients is not entirely acceptable, the solution might be achieved as follows:

$$d_{hk} = \cos\left[\frac{1}{n_h n_k}\sum_{i,j}\cos^{-1} s_{hk}\right]$$

where d_{hk} is the similarity between clusters h and k, n_h and n_k are the number of countries in cluster h and k respectively, and s_{hk} is a single inter-individual measure.

Centroid Cluster Analysis When using the centroid method, the distance between clusters is defined as the distance between their centroids. The centroid is a set of variable means across the members of the group (Lorr, 1983). The centroid method uses squared distances. Groups are fused in terms of the distance between their centroids, those with the smallest distance being fused first (Everitt, 1974). The formula for centroid clustering is:

$$d_{hk} = (n_i/n_k)d_{hi} + (n_j/n_k)d_{hj} - (n_i n_j/n_k^2)d_{ij}$$

where $n_k = n_i + n_j$

Everitt (1974) provided an illustration of centroid cluster analysis for five countries on the basis of two variables. Suppose the initial data set is:

		Variable 1	Variable 2
Country	1	1.0	1.0
	2	1.0	2.0
	3	6.0	3.0
	4	8.0	2.0
	5	8.0	0.0

Given the data set, the inter-individual matrix is computed using the squared Euclidean distance to yield D_1 below.

$$D_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.0 & 1.0 & 29.0 & 50.0 & 50.0 \\ 1.0 & 0.0 & 26.0 & 49.0 & 53.0 \\ 29.0 & 26.0 & 0.0 & 5.0 & 13.0 \\ 50.0 & 49.0 & 5.0 & 0.0 & 4.0 \\ 50.0 & 53.0 & 13.0 & 4.0 & 0.0 \end{bmatrix} \end{array}$$

The values in the matrix are the squared Euclidean distances between countries. The following is an illustration of how the distance between country 1 and 3 (d_{13}) and that between country 2 and 3 (d_{23}) were calculated.

$$d_{13} = (1 - 6)^2 + (1 - 3)^2$$

$$= 29$$

$$d_{23} = (1 - 6)^2 + (2 - 3)^2$$

$$= 26$$

The first stage of the clustering procedure follows the calculation of the D_1 matrix. Examination of the matrix shows that countries 1 and 2 are the closest with a distance of one unit. These countries are combined into one group and the data set is reduced to

		Variable	
		1	2
Country	(12)	1.0	1.5
	3	6.0	3.0
	4	8.0	2.0
	5	8.0	0.0

Variables 1 and 2 for group (12) are the averages for countries 1 and 2. An illustration of the computation is given below.

Variable 1: $(1 + 1)/2 = 1$

Variable 2: $(1 + 2)/2 = 1.5$

Given the above data set, elements for the second distance matrix (D2) are computed. The procedure for calculating D_2 elements is the same as that explained above for D_1. Alternatively, the distances of the newly formed group can be calculated as follows:

$$d_{3(12)} = (n_1/n_{12})d_{31} + (n_2/n_{12})d_{32} - (n_1 n_2/n_{12}^2)d_{12}$$

$$= 1/2(19.0) + 1/2(26) - 1/4(1.0)$$

$$= 27.25$$

$$d_{4(12)} = (n_1/n_{12})d_{41} + (n_2/n_{12})d_{42} - (n_1n_2/n_{12}^2)d_{12}$$

$$= 1/2(50.0) + 1/2(49.0) - 1/4(1.0)$$

$$= 49.25$$

$$d_{5(12)} = (n_1/n_{12})d_{51} + (n_2/n_{12})d_{52} - (n_1n_2/n_{12}^2)d_{12}$$

$$= 1/2(50.0 + 1/2(53.0) - 1/4(1.0)$$

$$= 51.25$$

The D2 matrix is now

$$D_2 = \begin{array}{c} \\ (12) \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (12) & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.0 & 27.25 & 49.25 & 51.25 \\ 27.25 & 0.0 & 5.0 & 13.0 \\ 49.25 & 5.0 & 0.0 & 4.0 \\ 51.25 & 13.0 & 4.0 & 0.0 \end{bmatrix} \end{array}$$

Given the above matrix, the closest countries are 4 and 5. These are now combined into one group and the data set is reduced to the following. The calculation of the variables for a new group is the same as explained earlier.

		Variable	
		1	2
Country	(12)	1.0	1.5
	3	6.0	3.0
	(45)	8.0	1.0

The above data set yields the D_3 matrix shown below. The elements for the D_3 matrix can be directly computed using the centroid formula.

$$D_3 = \begin{array}{c} \\ (12) \\ 3 \\ (45) \end{array} \begin{array}{c} \begin{array}{ccc} (12) & 3 & (45) \end{array} \\ \begin{bmatrix} 0.0 & 27.25 & 49.25 \\ 27.25 & 0.0 & 8.0 \\ 49.25 & 8.0 & 0.0 \end{bmatrix} \end{array}$$

Examination of D_3 indicates that the smallest distance is between country 3 and a group of countries 4 and 5. They are fused together to form a three member group. The final stage is to fuse the two resulting groups to form a single group.

Median Linkage Clustering When centroid clustering is used to merge two groups that differ substantially in size, the centroid of the new cluster is very close to the larger group. The properties of the smaller group are then lost. Median clustering deals with this problem by making clustering independent of group sizes. With median clustering, the formula becomes

$$d_{hk} = 1/2d_{hi} + 1/2d_{hj} - 1/4d_{ij}$$

The clustering process is similar to the previous examples, except for a change of formula.

Minimum Variance Method The minimum variance method uses the error sum of squares (E.S.S.) to determine the fusion of countries. Initially, each country is regarded as a single member of a group with an error sum of squares of zero (Everitt, 1974). At each clustering, an E.S.S. of every possible pair of groups is determined. Countries in a pair of Groups with the minimum error sum of squares are combined into a single cluster. The step with the greatest increase in the E.S.S. indicates that the accuracy has been diminished by reducing the number of groups. With Ward's method, the error sum of squares is defined as

$$\text{E.S.S.} = \sum_{i=1}^{n} x^2_i - 1/n\ (\sum x_i)^2$$

where x_i is the score of the ith country (Everitt, 1974).

Divisive Methods

A cluster analysis with divisive methods begins by splitting a set of countries into two. Divisive methods are classified as either monothetic or polythetic.

Monothetic Divisive Methods Monothetic divisive methods are based on the possession or the lack of a specific attribute (Lorr, 1983). These techniques are usually applied to binary data. If a monothetic technique is applied to a data set with m attributes, there will be m potential divisions of the initial set, m-1 potential divisions of each of the two subsets formed, m-2 potential divisions of each of the four subsets formed from the second division, and so on (Everitt, 1974). Association analysis and the automatic interaction methods are examples of the monothetic divisive methods.

Association Analysis When using association analysis to cluster a set of data with m binary attributes and N countries, the first step is to divide the initial group into two sub groups on the basis of the presence or absence of one of the binary characters (T). One group will consist of countries which possess character T while the other group will consist of those countries which lack character T.

Association analysis begins with the computation of the m x m matrix of chi squared coefficients

$$x^2_{jk} = \frac{(ad - bc)^2 N}{(a + b)(a + c)(b + d)(c + d)}$$

where N individuals have been scored on m attributes, and a, b, c, and d are the cell counts in a four-fold table for attribute j and k (Everitt, 1974). This table will be in the form

		Variable j		
		1	0	
Variable k	1	a	b	a + b
	0	c	d	c + d
		a + c	b + d	5

where a is the number of countries which possess both character j and k, b is the number of countries which possess character k but lack character j, c is the number of countries which possess character j but lack character k, and d is the number of countries which lack both character j and k.

Next the division criterion is used to divide the data. The most commonly used criterion is basing the split on a variable k which makes $\sum_{j \neq k} x^2_{jk}$ a maximum.

Everitt (1974) illustrated the use of association analysis of five countries on the basis of three binary attributes, the data set for which is shown below:

		Variable		
		1	2	3
Country	1	0	1	1
	2	1	1	0
	3	1	1	1
	4	1	1	0
	5	0	0	1

Given the data set, the first procedure is to calculate the three chi-squared statistics between the three variables using the formula given above. The result is as follows:

(i) Variable 1 and 2

		Variable 2		
		1	0	
Variable 1	1	3	0	3
	0	1	1	2
		4	1	5

$x^2_{12} = 1.87$

(ii) Variable 1 and 3

		Variable 3		
		1	0	
Variable 1	1	1	2	3
	0	2	0	2
		3	2	5

$x^2_{13} = 2.22$

(iii) Variable 2 and 3

		Variable 3		
		1	0	
Variable 2	1	2	2	4
	0	1	0	1
		3	2	5

$x^2_{23} = 0.83$

From the above it is possible to compute the $\sum x^2_{jk}$ value as follows:

Variable 1: $x^2_{12} + x^2_{13} = 4.09$

Variable 2: $x^2_{21} + x^2_{23} = 2.70$

Variable 3: $x^2_{31} + x^2_{32} = 3.05$

The Maximum $\sum x^2_{jk}$ is 4.09 for variable 1, meaning that the first division will be based on the presence or absence of character 1. Basing the division on character 1 results in the following two groups.

	Group 1	Group 2
Country	(2,3,4)	(1,5)

Group 1 consists of those countries which possess character 1 while group 2 consists of countries which
lack character 1. The division then continues in an identical manner on group 1 and 2 separately.

Automatic Interaction Detector Method (AID) The automatic interaction detector method is another monothetic divisive method. AID consists of dividing a sample through a series of binary attributes into mutually exclusive subsets (Lorr, 1983). The method seeks optimal reduction in the unexpected sum of squares of the independent variables (Everitt, 1974). The splitting point selected is that which maximizes the between group sum of squares (B.S.S.) at any binary split. The split results in one group with a low criteria score and the other group with a high criteria score. At any split, B.S.S. is maximized.

$$B.S.S. = N_1N_2 / N(\bar{x}_1 - \bar{x}_2)^2$$

where N_1 and N_2 are the group sizes and $\bar{x}_1$ and $\bar{x}_2$ are the criteria means.

Begin by forming binary splits for each variable, then maximize BSS/TSS (the ratio of between group sum of squares to the total sum of squares) of the group to be split. The variable with the highest BSS/TSS ratio is then used to initiate the division of the parent group, and the process is repeated treating each resulting subgroup as a separate sample.

Polythetic Divisive Methods Unlike the monothetic divisive techniques which are based on a single attribute, polythetic divisive methods are based on all the attributes. To illustrate polythetic divisive techniques, Everitt (1974) used the method in which the measure of similarity is the Euclidean distance between each entity and the other entities in the group. As an illustration, suppose there are seven countries whose distance matrix is:

	1	2	3	4	5	6	7
1	0	10	7	30	29	38	42
2	10	0	7	23	25	34	36
3	7	7	0	21	22	31	36
4	30	23	21	0	7	10	13
5	29	25	22	7	0	11	17
6	38	34	31	10	11	0	9
7	42	36	36	13	17	9	0

Given the distance matrix, the first step is to divide the entities into two groups. The country used to initiate the division is that whose average distance from the others is maximum. The average distance for each country from the others is calculated as follows:

Country	Average distance from others
1	(10 + 7 + 30 + 29 + 38 + 42)/6 = 26
2	(10 + 7 + 23 + 25 + 34 + 36)/6 = 22.5
3	(7 + 7 + 21 + 22 + 31 + 36)/6 = 20.67
4	(30 + 23 + 21 + 7 + 10 + 13)/6 = 17.33
5	(29 + 25 + 22 + 7 + 11 + 17)/6 = 18.5
6	(38 + 34 + 31 + 10 + 11 + 9)/6 = 22.17
7	(42 + 36 + 36 + 13 + 17 + 9)/6 = 25.5

The above results indicate that country 1 has the maximum average distance from the others. The resulting groups are as follows:

Splinter Group	Main Group
(1)	(2,3,4,5,6,7)

Having identified the two groups above, the average distance of each country in the main group (2,3,4,5,6,7) to the country in the splinter group (1) and the average distance of each country in the main group to the others within the group are calculated. The table below gives these distances.

Country	Average distance to splinter group 1	Average distance to main group 2	Difference (2 - 1)
2	10.0	25.0	15.0
3	7.0	23.4	16.4
4	30.0	14.8	-15.2
5	29.0	16.4	-12.6
6	38.0	19.0	-19.0
7	42.0	22.2	-19.8

From the table, the maximum difference is 16.4 for country 3. Countries 3 and 1 are combined into a single group to generate two new groups as follows:

Splinter Group	Main Group
(1,3)	(2,4,5,6,7)

On the basis of these groups new average distances are calculated and assembled into a table as follows:

Country	Average distance to splinter group 1	Average distance to main group 2	Difference (2 - 1)
2	8.5	29.0	21.0
4	25.5	13.2	-12.3
5	25.5	15.0	-10.5
6	34.5	16.0	-18.5
7	39.0	18.7	-20.3

In the table, the maximum average distance is 21.0 for country 2. This country is then added to the splinter group to yield two new subgroups as follows:

(1,3,2) and (4,5,6,7)

Continuing with the analysis in the same manner generates the following table.

Country	Average distance to splinter group 1	Average distance to main group 2	Difference (2 - 1)
4	24.7	6.7	-18.0
5	25.3	11.7	-13.6
6	34.3	10.0	-24.3
7	38.0	13.0	-25.0

In the above table, all the differences are negative, meaning that each of countries 4, 5, 6, and 7 is closer to the main group than splinter group. For this reason, the analysis moves to a further division of subgroups (1,3,2) and (4,5,6,7). The procedure is the same as that employed in the initial division.

OPTIMIZATION - PARTITIONING TECHNIQUES

Like hierarchical clustering techniques, partitioning techniques produce a partition of objects. Unlike hierarchical techniques, partitioning techniques allow for the correction of a poor initial partition by relocating countries. A discussion of the partitioning process and clustering criterion follows.

The Process of Partitioning

The process of partitioning involves three distinct steps. These are initiating clusters, allocating countries to initial clusters, and reallocating some or all of the countries to other clusters after completion of the initial classification. These steps are individually discussed below.

Initiating Clusters Suppose the analyst intends to partition a set of countries into a predetermined number of clusters (k). To initiate clusters, the analyst has to identify k points. Several procedures for identifying k initial points have been suggested. MacQueen (1967) described a procedure in which the first k points in the data are selected as the initial k clusters. Everitt (1974) discussed Beale's method which initiates clusters with a value of k larger than necessary and sets centers regularly spaced at intervals of one standard deviation on each variable. The number of groups is then reduced until the clustering criteria (based on residual sum of squares) is satisfied.

Allocating Countries to Initial Clusters Identifying k initial clusters is followed by the allocation of each of the other countries to one of the k clusters. With MacQueen's method, each subsequent country is allocated to the cluster it is nearest to, on the basis of the euclidean metric. After each assignment the new mean for each cluster is computed.

Relocation of Countries With MacQueen's method, the cluster centroids for each group are fixed following the allocation of each country to one of the k initial clusters. Each country is then rechecked to see if it is nearer to any of the k clusters than the one it has been assigned to. Reallocation takes place when a country is closer to a different cluster. The process continues until no movement of a country is necessary, that is, when no further movement of a country improves the criterion being optimized.

Clustering Criteria

Many of the clustering criteria attempt to minimize the scatter or variation within clusters and to maximize the variation between clusters (Lorr, 1983). Many of these criteria are derived from the matrix equation.

$$T = W + B = X'X$$

where T is the total dispersion matrix, W is the pooled within group dispersion and B is the between group dispersion matrix.

In the equation,

$$W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)'(x_{ij} - \bar{x}_i)$$

$$B = \sum_{i=1}^{g} n_i \bar{x}_i' \bar{x}_i$$

where

x_{ij} is the jth observation vector (1xk) of the ith cluster,

$\bar{x}_i$ is the mean vector (1xk) of the ith cluster,

n_i is the number of observations in the ith cluster, and

g is the number of clusters (McRae, 1971).

Because of the fixity of T, minimization of W is equivalent to maximization of B.

Several criteria have been developed from the matrix equation. These include minimizing the trace of W, maximizing the ratio $|T|/|W|$ which is equivalent to minimization of $|W|$, and maximizing the trace of $W^{-1}B$. Reviews of these criteria are contained in Everitt (1974), Marriot (1971), McRae (1971) and Lorr (1983).

Everitt (1974) provides an illustration of the application of a partitioning technique on seven individuals (countries or export markets) on the basis of two variables. The data matrix is as follows:

		Variable 1	2
	1	1.0	1.0
	2	1.5	2.0
Country	3	3.0	4.0
	4	5.0	7.0
	5	3.5	5.0
	6	4.5	5.0
	7	3.5	4.5

Given the data matrix, the initial step is to partition the set into two groups such that each member of a group is nearer to the mean vector of that group than that of the other group (Everitt, 1974). Using the Euclidean measure, the two groups that are farthest apart are identified and used as a starting point. From the distance matrix (not shown) countries 1 and 4 are the furthest apart. The resulting two groups are as follows:

	Group 1	Group 2
Individual	1	4
Estimated Mean Vector	(1.0,1.0)	(5.0,7.0)

Everitt further stated that the other countries are examined in sequence and allocated to the groups whose mean vector is closest, and that the mean vector is recalculated every time a new country is added. This results in the following process.

Stage	GROUP 1		GROUP 2	
	Member	mv	Member	mv
1	1	(1.0,1.0)	4	(5.0,7.0)
2	1,2	(1.2,1.5)	4	(5.0,7.0)
3	1,2,3	(1.8,2.3)	4	(5.0,7.0)
4	1,2,3	(1.8,2.3)	4,5	(4.2,6.0)
5	1,2,3	(1.8,2.3)	4,5,6	(4.3,5.7)
6	1,2,3	(1.8,2.3)	4,5,6,7	(4.1,5.4)

From the table, the initial clustering of the two groups is as follows:

Group 1	Country 1, 2, and 3 Mean Vector = (1.8,2.3)
Group 2	Country 4, 5, 6, and 7 Mean Vector = (4.1,5.4)

The next step is to test whether each country is nearer to the mean vector of its own group than that of the other group. Country 3 was found to be closer to group 2 than 1. This country is then reallocated to group 2, resulting in the following groups:

Group 1	Country 1 and 2 Mean Vector = (1.2,1.5)
Group 2	Country 3, 4, 5, 6, and 7 Mean Vector = (3.9, 5.1)

The procedure could be continued identically in a further subdivision of groups 1 and 2.

DENSITY OR MODE SEEKING TECHNIQUES

The density or mode seeking techniques identify natural clusters in metric space. "Natural clusters are mutually exclusive subsets whose members are sufficiently related to each other so that recognitions of such clusters notably facilitates comprehension of the relations among all the items and permits good generalization or prediction about the attribute values of cluster members" (Carmichael et al., 1968). These techniques facilitate clustering by identifying high density areas in metric space. Techniques developed by Cattell and Coulter (1966), Carmichael et al. (1968), and Wishart (1969) constitute the density or mode seeking techniques.

The Cartet Count Method

The cartet count method was developed by Cattel and Coulter (1966). Countries (or export markets) are placed in a coordinate system. Convenient intervals are then taken on the coordinates to partition the space into cartets, which, in two-dimensional space, are squares. In multi-dimensional space, cartets are defined by hypercubes. Countries in each hypercube are counted. To locate clusters, a significantly high density count is set for a cube relative to the average total density. The limitation of this method is that there is no objective criteria for determining the number of clusters present.

The Taxmap Method

The Taxmap method was developed by Carmichael et al. (1968). This method compares the relative distance between points and then searches for continuous relatively densely populated regions of the space surrounded by continuous relatively empty regions (Everitt, 1974). The idea is that if there are natural clusters, the two countries which are closest belong to one of these clusters. Clustering is initiated by merging these countries. The next country for consideration is that which is closest to either of the two already merged. Termination of admission to a cluster comes if the prospective country is much farther away than the last country admitted. That is, "if there was a discontinuity in closeness" (Carmichael, 1968). Discontinuity in closeness is defined by a sudden drop in the average linkage.

Initial clusters are formed in a similar manner to the single linkage method, but there is a criteria for stopping the additions to the clusters. The following example of using this method is adapted from Everitt (1974). Assume five countries with the following similarity matrix.

$$S = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 1.0 & 0.7 & 0.9 & 0.4 & 0.3 \\ 0.7 & 1.0 & 0.8 & 0.5 & 0.4 \\ 0.9 & 0.8 & 1.0 & 0.4 & 0.2 \\ 0.4 & 0.5 & 0.4 & 1.0 & 0.7 \\ 0.3 & 0.4 & 0.2 & 0.7 & 1.0 \end{bmatrix} \end{array}$$

From the similarity matrix, the most similar countries are 1 and 3 with a similarity of 0.9. The initial step is to combine these two countries into a single group. The next step is to consider another country for inclusion into a group just formed. The country to be considered is that closest to either member of the group just formed. From the matrix country 2 is the closest to the group consisting of 1 and 3 with a similarity of 0.8 to country 3. Next the average similarity between the three countries is computed as follows:

$$(0.9 + 0.7 + 0.8)/3 = 0.8$$

When country 2 is included in the initial group consisting of countries 1 and 3, the decrease in similarity is 0.1 (0.9 - 0.8), and the measure of discontinuity is 0.7 (0.8 - 0.1). Lower measures of discontinuity indicate that the country should not be added to the cluster. If 0.5 was regarded as the determining point, country 2 with a discontinuity measure of 0.7 would be added to the group consisting of 1 and 3.

Having included country 2 to form a group consisting of countries 1, 3, and 2, the similarity matrix is examined for the next country closest to a member of the cluster. Countries 4 and 2 have the highest similarity of 0.5. Country 4 is then considered for inclusion into a group of 1, 3, and 2. Once again the average similarity of the four countries is computed as follows:

$$(0.9 + 0.7 + 0.4 + 0.8 + 0.5 + 0.4)/6 = 0.6$$

The decrease in similarity is 0.2 (0.8 - 0.6), and the measure of discontinuity is 0.4 (0.6 - 0.2). With minimum acceptable discontinuity set at 0.5, country 4 is not included into a group consisting of countries 1, 3, and 2. Country 4 is therefore used to initiate the second cluster. The merging then proceeds in an identical manner with country 4 as the initial basis for merging. Country 4 is closest to country 5 with a similarity of 0.7. These are combined into one group. Finally, there are two clusters as follows:

	Cluster 1	Cluster 2
Countries	1, 2, & 3	4 & 5

Clusters 1 and 2 can be subdivided using the procedure just described.

This method has a limitation in that "numerous parameters controlling the techniques" must be set by the analyst (Lorr, 1983).

Mode Analysis

Mode analysis was developed by Wishart (1969). This method searches for natural clusters by identifying dense and non-dense points. First the analyst selects the distance threshold (R) and the frequency threshold (K). To identify the initial clusters, a sphere with radius R is considered at each point, and the points falling within the sphere are counted. The points with spheres containing K or more other points are called dense points while those points with spheres containing less than K other points are referred to as non-dense points. Next, dense points are clustered by single linkage to represent the initial clusters. Each non-dense point is then reallocated to a suitable cluster on the basis of some criteria such as including a non-dense point in the cluster containing its nearest dense point (Wishart, 1969). A comprehensive discussion of mode analysis is provided by Wishart (1969).

Determining the Number of Clusters

Determining the number of clusters present is a task the researcher has to address with a practical knowledge of the need for market segregation. Professional judgement is an important factor in this decision. The discussion of this topic begins with hierarchical techniques and proceeds to other techniques.

HIERARCHICAL CLUSTERING TECHNIQUES

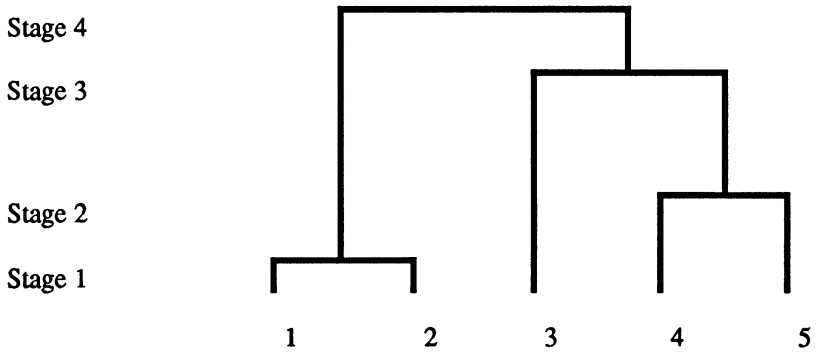
With hierarchical clustering techniques, there are no clear indicators to determine the best number of clusterss. Everitt (1974) suggested an examination of the dendogram for large changes between fusions as useful in the event clusters are required. For divisive techniques, various stopping rules such as the predetermined within group sum of squares or a minimum number of countries in the ultimate categories have been suggested.

Moreover, various stopping rules based on the distribution of the criterion (α) being maximized (minimized) have been suggested. Mojena (1977) evaluated two stopping rules with α defined as the standardized Euclidean distance. With these rules, "a significant change in α from one stage to the next implies a partition which should not be undertaken" (Mojena, 1977). One rule was found to yield stable results. For a similarity measure whose large values imply dissimilar groups, the rule states that a group level selected should satisfy

$$\alpha_{j+1} > \bar{\alpha} + ks_{\alpha}$$

where α_{j+1} is the value of the criterion in stage j+1; k is the standard deviate, $\bar{\alpha}$ and s_{α} are, respectively, the mean and the unbiased standard deviation of the α distribution. If the inequality is not satisfied at all values of α, the investigator should choose the stage j for which stage j+1 yields the largest standard deviate. The inequality is reversed if a large value of a similarity measure implies similar groups.

Suppose that the five country cluster example presented earlier yielded a dendogram as shown below:



If the inequality is satisfied at stage 2 + 1 (3) the analyst will have to stop clustering at stage 2, meaning that there are three clusters: (12), 3, and (45)

OPTIMIZING TECHNIQUES

With techniques that optimize a certain criterion, plotting the value of the criterion against the number of groups may indicate an appropriate number of clusters to be considered. If the criterion (W) is being maximized, a sharp increase may occur at the correct number of groups. If the criterion is being minimized, a sharp decrease occurs at the correct number of clusters.

Marriott (1971) pointed out that the criterion $g^2|W|$ may provide an answer to the number of natural clusters available. According to Marriot, the optimum subdivision is that which minimizes the criterion $g^2|W|$ where g is the number of groups and |W| is the variance - covariance matrix. For the details, the reader is referred to Marriot (1971).

DENSITY OR MODE SEEKING TECHNIQUES

The cartet count method partitions the multidimensional space into hypercubes and counts points in each cube. Natural clusters are located by setting a significant level for the cube relative to total average density. The clustering procedure for the taxmap method results in the identification of natural clusters. As was discussed under the taxmap method, the basis for clustering is a discontinuity measure. With mode analysis, natural clusters are initiated through identifying dense and non-dense points. Non-dense points are allocated to clusters containing their closest dense point (the procedure was explained previously under mode analysis).

Computation of Mean Profiles and Interpretation of the Findings

With clusters of countries, the researcher should calculate the mean profiles for each of the clusters on the basis of the dimensional variables (factor scores). Each cluster will have a score on each of the dimensional variables. To illustrate, Doyle and Gidengil (1977) had five market based factors: economic development, industrialization, distribution potential, urbanization, and trade. On the basis of these factors, mode analysis was used to detect eight clusters of countries. The mean factor scores were calculated for each of the eight groups. After these calculations, each of the eight clusters of countries had a single score on each of the five factors, meaning that clusters could be described in terms of the five factors. As an example, it was found that cluster 6 and 8 had means on factor 1 (economic development) that were higher than the overall means of all the countries. If economic development was an important factor in the selection, then these clusters would have been selected for further evaluation.

With these results (mean profiles), a researcher might further evaluate countries within a single cluster in order to decide which countries would be most favorable for export market development efforts. In some cases the company might want to evaluate countries in different clusters.

An Example from Previous Studies

Several cluster analysis studies have been conducted to segment international markets. An example is Doyle and Gidengil (1977). To sum up the process of cluster analysis for international market selection, this study is reviewed.

The study included eighty-five countries. Twenty six variables represented the market characteristics in each of the countries. The analysis began with factor analysis which reduced the data as well as correlation between variables. Factor analysis produced five groups of variables (factors). These were named economic development, industrialization, distribution potential, urbanization, and trade. Factor scores were then assigned to each country.

To measure similarity, Doyle and Gidengil used the cosine coefficient which is defined as

$$\cos \theta_{jk} = \frac{\sum_i x_{ij} x_{ik}}{(\sum_i x^2_{ij})^{1/2} (\sum_i x^2_{ik})^{1/2}}$$

where x_{ij} is the value of factor i for country j and x_{ik} is the value of factor i for country k. The cosine coefficient retains information on both magnitude and shape of the profiles. If instead the study used the product moment correlation coefficient, information on magnitude of the profiles would have been lost. The euclidean distance, on the other hand, overemphasizes some variables during squaring.

The study then proceeded to the choice of a clustering technique. Clustering techniques present problems in that they produce different results. Doyle and Gidengil (1977) decided against the use of hierarchical clustering techniques because they are unable to detect natural clusters, unless the clusters are clearly separated. These techniques were further criticized for yielding groups even though no natural clusters are available. Doyle and Gidengil therefore used mode analysis because of its ability to detect natural clusters. Eleven clusters of countries were produced and the mean profiles computed for each cluster.

Conclusion

Cluster analysis can be a useful tool in international market selection. The objective of cluster analysis is to group countries into clusters for the purpose of selecting a group the researcher wishes to investigate further. An in-depth analysis of countries in the selected group will lead to the selection of a small number of markets for business development.

This paper reviewed the process of cluster analysis with emphasis on the discussion of alternative clustering techniques (hierarchical, optimization-partitioning, and density or mode seeking). The results of cluster analysis will vary according to the similarity index chosen and the clustering technique used.

Hierarchical clustering techniques are limited because they produce partitions even though no natural groups exist (sometimes they cut across natural clusters). They do not allow for reallocation of countries which were initially misplaced. Optimization techniques frequently produce sub-optimal solutions because it is impossible to consider every possible partition. Density or mode seeking techniques also have limitations, although they attempt to identify natural clusters. The cartet count method may slice through natural clusters. With the Taxmap method, various parameters which control clustering are arbitrarily chosen by the analyst. Mode analysis is scale dependent and assumes spherical clusters. This poses problems if ellipsoidal clusters are present. Overall, hierarchical clustering techniques seem to have the most limitations. Figure 4 shows strengths and weaknesses of the different clustering techniques.

The analyst needs to make careful choices of both the similarity index and the clustering technique. Comparing results of alternative techniques may be useful to the analyst.

Figure 4: Strengths and Weaknesses of Cluster Analysis Techniques

Hierarchical Techniques		Optimization-Partitioning Techniques		Density or Mode-Seeking Techniques	
Strengths	Weaknesses	Strengths	Weaknesses	Strengths	Weaknesses
Appropriate for biological types of data for which a hierarchical structure can be assumed to exist	- Produce partitions even though no natural groups exist - May cut across natural clusters - No way to reallocate countries originally misplaced	Allows for reallocation of countries which were initially misplaced	Frequently produce sub-optimal solutions because every possible partition cannot be examined	Attempt to identify natural clusters	-The Cartet count method may slice through natural clusters -With the taxmap method, parameters are arbitrarily chosen by the analyst -Mode analysis is scale dependent and assumes spherical clusters

References

Anderberg, M.R. *Cluster Analysis for Applications.* New York: Academic Press, 1973.

Carmichael, D.W., George, J.A., and Julius, R.S. "Finding Natural Clusters." *Systematic Zoology*, 1968, 17, 144-150.

Cattell, R.B., and Coulter, M.A. "Principles of Behavioral Taxonomy and the Mathematical Basis of the Taxonomy Computer Program." *British Journal of Mathematical and Statistical Psychology*, 1966, 237-269.

Day, E., Fox, R., and Huszagh, S.M. "Segmenting the Global Market for Industrial Goods: Issues and Implications." *International Marketing Review*, 1988, 5, No. 3 Autumn, 14-27.

Doyle, P. and Gidengil, Z.B. "A Strategic Approach to International Market Selection", *Proceedings of the American Marketing Association, AMA*, Chicago, 1977, Ill, pp.230-4.

Everitt, B. *Cluster Analysis.* New York: John Wiley & Sons, 1974

Lorr, M. *Cluster Analysis for Social Scientists.* San Francisco, CA: Jossey - Bass, 1983.

MacQueen, J. "Some Methods for Classification and Analysis of Multivariate Observations". In L. M. LeCam and J. Neyman (Eds), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley: University of California Press, 1967, 1, pp 281-514.

Marriot, F.H.C. "Practical Problems in a Method of Cluster Analysis." *Biometrics*, 1971, 27, 501-514.

McRae, D.J. MICKA: A FORTRAN IV Iterative K-means Cluster Analysis Programme. *Behavioral Science*, 1971, 16, 423-424.

Mojena, R. "Hierarchical Grouping Methods and Stopping Rules: An Evaluation." *Computer Journal*, 1977, 20, 359-365.

Papadopolous, N. and Denis, J. "Inventory, Taxonomy and Assessment of Methods for International Market Selection". *International Marketing Review*, 1988, 5, No. 3 Autumn, 38-51.

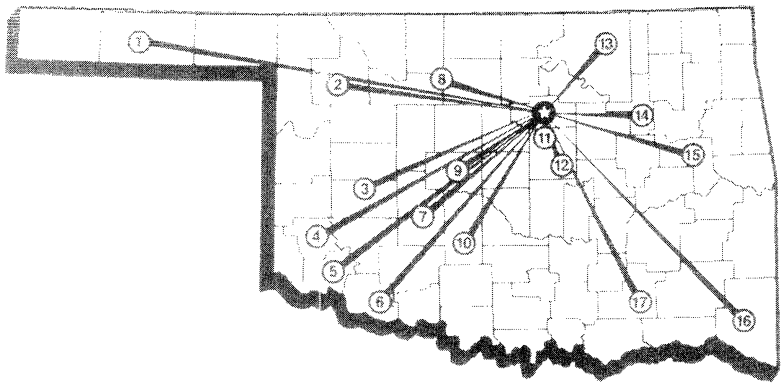
Robock, S.H. and Simmons, K. *International Business and Multinational Enterprises.* Homewood, Ill: Richard D. Irwin, 1983.

Sethi, S.P. "Comparative Cluster Analysis for World Markets." *Journal of Marketing Research*, 1971, 8, August, pp. 348-54.

Sethi, S.P. and Curry, D. (1973), "Variable and Object Clustering of Cross-Cultural Data: Some Implications for Comparative Research and Policy Formulations", *Multinational Business Operations*, Sethi, S.P. and Sheth, J.H. (Eds.), Good Year, Pacific Palisades, CA, pp. 31-61.

Wishart, D. (1969) Mode Analysis. *Numerical Taxonomy* (A.J. Cole, ed.), 282-308. New York, Academic Press.

THE OKLAHOMA
AGRICULTURAL EXPERIMENT STATION
System Covers the State



Main Station — *Stillwater and Lake Carl Blackwell*

1. Panhandle Research Station — *Goodwell*
2. Southern Great Plains Field Station — *Woodward*
3. Marvin Klemme Range Research Station — *Bessie*
4. Sandyland Research Station — *Mangum*
5. Irrigation Research Station — *Altus*
6. Southwest Agronomy Research Station — *Tipton*
7. Caddo Research Station — *Ft. Cobb*
8. North Central Research Station — *Lahoma*
9. Forage and Livestock Research Laboratory — *El Reno*
10. South Central Research Station — *Chickasha*
11. Agronomy Research Station — *Perkins*
    Fruit Research Station — *Perkins*
12. Pecan Research Station — *Sparks*
13. Pawhuska Research Station — *Pawhuska*
14. Vegetable Research Station — *Bixby*
15. Eastern Research Station — *Haskell*
16. Kiamichi Forestry Research Station — *Idabel*
17. Wes Watkins Agricultural Research and Extension Center — *Lane*

Issued by Oklahoma Agricultural Experiment Station serving people of all ages, social-economic levels, race, color, sex, religion and national origin. This publication is printed and issued by Oklahoma State University as authorized by the Dean of the Division of Agriculture. 500 copies have been printed and distributed at a cost of $1,299.46 for 1,175 copies. AC-874 10/1