2002 World Cultures 13(1): 22-38

CLUSTER ANALYSIS IN CROSS-CULTURAL RESEARCH

Bruce A. Maxwell
Department of Engineering, Swarthmore College, Swarthmore, PA 19081; Maxwell@swarthmore.edu
Frederic L. Pryor*
Department of Economics, Swarthmore College, Swarthmore, PA 19081; Fpryor1@swarthmore.edu
Casey Smith
Swarthmore College, Swarthmore, PA 19081

* Corresponding Author

This essay has three purposes. The first is to present a relatively non-technical description of cluster analysis. The second is to describe a computer program available on the World Wide Web, which allows such a statistical technique to be carried out in a very simple way. The third is to show how this approach can be used with cross-cultural data to extract similarities and differences between societies in a systematic fashion. Although the example used focuses on the economic systems of foragers, the methodology is also applicable to a wide variety of other cross-cultural research problems.
?

1. INTRODUCTION

Although considerable cross-cultural data are available - for instance, the 1700 series for the Standard Cross-Cultural Sample published by *World Cultures* - such information has been underutilized. Part of the problem is that so much information is available that it is difficult to discern patterns in a sufficiently objective manner to allow others to be able to replicate the results.

One traditional way to reduce the dimensionality of the data is to use some variant of principle component analysis, a technique that permits us to determine which traits are related. Nevertheless, if we wish to determine which societies are the most similar or different using the results of the derived principle components, difficulties begin to arise because, according to one factor, two societies may be very different while, according to another factor, they may be quite similar. Other analytic problems arise because the results may depend upon whether we employ a standard principle component analysis, where, in effect, each factor is removed before the next factor is derived (thus deriving orthogonal factors) or some type of varimax technique in which the factors may be related.

Cluster analysis approaches the problem of determining similarities and differences among societies more directly, namely by determining the multidimensional distances between various societies and then picking out those groups of societies within which the distances are relatively small. This statistical technique has been used in a wide variety of data analysis and pattern recognition applications, and a number of clustering techniques exist, the most common ones being the K-means and hierarchical clustering algorithms (MacQueen 1967; Johnson 1967). The k-means

cluster analysis explicitly divides the data into a set of k groups by trying to minimize intra-cluster variance and maximize inter-cluster variance by using an iterative algorithm; hierarchical clustering is a step-wise process that merges the two closest data points or group of data points at each step. A hierarchical clustering process creates a tree structure with each data point as a leaf at the top of the tree and all of the data points as a single group at the bottom. The hierarchical clustering algorithm can generate any number of groups simply by stopping the step-wise process—in other words, cutting the tree--at the appropriate number of branches (subsets). Although clustering techniques have been employed by some social scientists for analyzing cross-cultural differences (for instance, Schneider 1999; or Divale 1997), we have not found any recent uses for analyzing ethnographic data.

Those interested in using cluster analysis face three hurdles: most descriptions of cluster analysis are highly technical; most available programs (of which we are aware) are difficult to use; and the manner in which the technique can be applied to cross-cultural data and the way in which the data can be interpreted, are far from clear.

The purpose of this essay is to reduce these barriers. To provide concrete understanding on the use of cluster analysis, we explore in considerable detail a specific example, namely determining economic systems among foraging societies from an examination of ten parameters of property and exchange relationships. The flow of the argument below is straightforward: We turn first to a more extensive description of cluster analysis. In the following section we examine the ethnographic problem that serves as the example. In the third section we discuss briefly the actual program, providing additional notes on its use in an appendix. In the remaining sections we explore the results, showing how the results can be understood and some of the common pitfalls of interpretation.

2. THE ABCs OF CLUSTER ANALYSIS

The purpose of cluster analysis is to identify subsets of a data set that contain similar points. Replacing these subsets by their aggregate properties, such as means and standard deviations, creates a compact representation of the data set as a set of clusters. The cluster properties can then be used for comparative data analysis.

There are two general types of cluster analysis: unsupervised and supervised clustering. Unsupervised clustering imposes no *a priori* assumptions about where the natural clusters are. Inputs given by the user include: the choice of variables, the relative weight of each variable, and the total number of clusters. It represents a bottom-up approach to the analysis, and is most commonly used in exploratory analysis and in developing new typologies of complex data sets (MacQueen 1967).

Supervised clustering, on the other hand, uses a set of set of example data points to classify the rest of the data set. In addition the user still determines the variables, relative weights, and number of clusters. To execute supervised clustering, the user must know *a priori* where the clusters are in

the data space. The most common use of supervised clustering is to classify new data using an existing typology; for instance, out-of-sample data can be used to determine the degree to which the derived clusters have more general applicability than the original sample.

In the example used below, we first carry out an unsupervised cluster analysis using data from societies relying 75 percent or more on foraging for subsistence. In this part of the exercise we are trying to discover if there is a natural typology of economic systems. We then use this typology to execute a supervised clustering on the economic systems of societies not in the initial cluster set that rely 55 to 75 percent on foraging.

One of the most difficult issues when dealing with multidimensional data points is how to define when two points are similar. In a homogeneous multi-variable space, such as the 3D space we inhabit, a useful measure is the Euclidean distance. For the N-dimensional data points x and y, the Euclidian distance D_E is:

$$1) \qquad D_E = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}$$

For any heterogeneous data set, defined here as data points whose dimensions do not have similar variances, the Euclidean distance should not be used. Instead, a scaled Euclidean distance D_S should be used.

$$2) \qquad D_S = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2 / \sigma}$$

The similarity measure D_S simply divides the squared difference of each dimension by the variance of that dimension. While this does not take into account correlations between dimensions, it does ensure distances in each dimension are statistically similar.

Note that these measures of similarity must be modified if there is missing data, a problem discussed by Maxwell and Buddemeier (2002). Alternative measures of distance, including measures that are not based on Euclidean measures, can also be specified, as discussed in the same source.

Once a user has determined the appropriate variables and an appropriate distance measure, the next step in the cluster analysis is to determine how many clusters should be identified. Although for a sample with 44 societies in our example, it is certainly possible to have 44 clusters (every society as a unique cluster), the user would gain no new information. One technique that can assist users in selecting an appropriate number of clusters is based on information theory. The technique tests a range of clusterings with differing numbers of clusters and calculates the binary description length for each result. The description length balances the number of clusters with the error in the representation. The error is simply the sum of the squared distances of each point to its nearest cluster mean. The error is maximized when there is a single cluster, and minimized (zero) when there are as many clusters as there are data points.

It turns out that the minimum description length [MDL] is the point at which the number of clusters balances the representational error. If the number of clusters is increased beyond the balance point, the marginal gain in information from the increased number of clusters does not produce a worthwhile reduction in the representation error. Rissanen (1989,2001) provides a more technical description of the MDL technique that, in essence, is a 20th century version of Occam's Razor--entities should not be multiplied unnecessarily. A user can examine the results of the MDL analysis visually using a graph, and the low point of the graph shows the optimal range for the number of clusters.

In some cases, the description length of k clusters is quite similar to that of k-1 or k+1. In the case of a relatively small sample, it is generally advisable to select the smaller number of clusters. In the example used below, for instance, the optimal number of clusters is 6, but 5 clusters are almost as good and, as a result, it was chosen (Pryor 2003a).

The final step in the cluster analysis is to run the clustering algorithm. The k-means unsupervised clustering algorithm is best suited for the purposes of generating an initial or exploratory cross-cultural analysis (MacQueen 1967). The hierarchical clustering algorithm may also be appropriate for exploratory analysis; however, the two approaches do not generally produce the same results for a given number of clusters. The k-means algorithm, since it directly calculates the k clusters, has more flexibility in identifying the clusters than hierarchical methods. Whichever method is used, the clustering results can be visualized in several ways: mapped onto longitude and latitude, or mapped into the data space. Visualizing the data using geographical location, for instance, shows that one cluster is located primarily in the North Pacific and polar regions. In this case, the diffusion of traits across societies seems likely. Other clusters are spread all over the world, however, suggesting that diffusion is less likely.

3. THE ETHNOGRAPHIC EXAMPLE

The ethnographic problem to be used as an example is drawn from Pryor (2003a), who was trying to determine whether it is possible to define different economic systems of foraging (hunting, gathering, or fishing) societies. The data base consisted of all societies in the Standard Cross-Cultural Sample, and most of the data used in this analysis were collected by him from various ethnographic sources. A foraging society is defined as any society directly obtaining 75 percent or more of its food from hunting, gathering, or fishing.

A looser definition (for instance, 55 to 75 percent reliance on foraging) was not used because of the fear that the foraging economic systems would be 'contaminated' by so much agriculture or animal husbandry. Nevertheless, as shown below, the use of such out-of-sample points is important for interpreting the results that are obtained. More details about the sample or the

data, and the reasons for selecting the particular identifying dimensions are discussed in Pryor (2003a) and need not concern us here.

The major problem in defining the clusters is determining the dimensions by which the clusters are to be identified. In economics, systems are usually defined in terms of property and exchange relationships, but at this point a serious problem arises. If the analyst sets up only a few criteria for distinguishing one economic system or another, the risks of subjectivity are high since the criteria selected may not capture key systemic differences between the various societies. Alternatively, the analyst can let the data speak for themselves by determining significant clusters of societies that have similar economic systems. Such an atheoretic approach, however, also has its risks: the data may be difficult to interpret if a large number of variables are used and/or if many clusters are obtained. Keeping this tradeoff in mind, the following defining dimensions are used:

1. Distribution dimensions: Two of the distribution dimensions - wealth inequalities and food sharing - concern particular aspects of sharing, a protean concept which covers both one-way transfers and two-way exchanges. The remaining two dimensions define several other types of distribution. More specifically, the four distribution dimensions are:

a. Important inequalities of wealth
b. Extent of food sharing
c. The presence of a significant amount of trade or barter
d. The presence of significant taxation or tribute paid to the political leader

2. Property dimension: Property or ownership is defined as the exclusive use of tangible or intangible assets, a relationship that is socially enforced. Four of the six different types of property cover ownership of land, people (slaves), food inventories, and intangible powers from which income can be gained. Two of the types of ownership reflect particular aspect of the accumulation of property.

a. Exclusive possession of land
b. The existence of private food stocks with an exclusive owner
c. The occurrence of slavery, either at the pinpointed date or in the past
d. The existence of important intangibles with which the owner gains appreciable income
e. The occurrence of a economically significant bridewealth
f. A significant inheritance of goods, rather than destruction upon death of the owner

The statistical problem can now be simply stated: Do the 44 societies fall into meaningful clusters using these ten dimensions? And, more immediately, is there a program which can be mastered in a relatively small amount of time to can carry out such an analysis. Both answers are affirmative.

4. THE COMPUTER PROGRAM

The clustering program, LOICZView was written as part of the Land-Ocean Interactions in the Costal Zone Project [LOICZ], which is a component of the International Geosphere-Biosphere Programme. The clustering program, however, has uses far beyond the purpose for which it was originally designed. It is now on the world wide web and can be accessed by anyone (www.palantir.swarthmore.edu/loicz/). There is a guest account for users interested in exploring the program. There is also an account set up for anthropologists who wish to use it for their own research under the alternative usernames *anthro1* to *anthro5*. To obtain the password, please contact the program maintainer at maxwell@swarthmore.edu.

An instruction manual for the program is provided in the appendix of this essay. LOICZView, described in greater detail by Maxwell and Buddemeier (2002), is a graphical user interface to a set of software tools that permits users to carry out both principle component analysis and cluster analysis. For the faint of heart, the program also contains instructions supplementing those provided in the user manual below.

In addition to any data sets already loaded into the program, users can load their own data. It is simplest to assemble the data on a spreadsheet, including, if possible, latitude, longitude, society name, and the different variables to define the cluster. More exact details on the file format are supplied in the appendix and within the program. Once the data set is complete, the file should be saved as a text file (*.txt) with the data delimited either by tabs or commas. The data set loads into the program using a standard web upload interface.

Users start an unsupervised cluster analysis by first specifying the number of clusters to be calculated. Users who wish to obtain guidance on the appropriate number of clusters for their data set should use the MDL tab to execute an analysis based on minimum description length as described above. Once the number of clusters is specified, all that is necessary is to select the unsupervised clustering technique to begin the analysis. When undertaking exploratory analysis, unsupervised clustering minimizes the imposition of the user's views on how the societies should be grouped. For unsupervised clustering, the program implements the k-means clustering algorithm, an iterative technique initialized using a random seed.

The random initialization may cause small differences in the final results. The program takes this into account, and tries to find an optimal solution by running the clustering algorithm with numerous initial conditions and keeping the result with the lowest representation error. With some data sets, there is little or no variation in the final results; in others, there may be significant differences.

To test the variability of a particular data set, the user should run the clustering program several times to see how the results differ, or if they differ at all. In the example under examination, this problem seemed particularly acute; so six runs were made, with each run taking the best result from 100 different initial conditions. In each of these six runs, the five clusters found by the

program appeared roughly the same. Nevertheless, six of the 44 societies moved from cluster to cluster. For five of these wayward societies we placed the society in the cluster where it was most often found. In the sixth case, a society with heavy Western influence (Slave Indians of northern Canada), we placed it in a cluster, which, on the basis of additional criteria, it seemed best to fit. In all these cases, however, a footnote is necessary to discuss these human interventions into the program's calculations.

Once the clustering is complete, LOICZView provides several methods for viewing the results. If the longitude and latitude are given, it can present the data on a map with color indicating cluster membership. If all the points of a given color are in one portion of the map, then diffusion of traits has likely occurred. In the example under discussion, it turns out that diffusion probably occurred in some clusters, but not in others, which are scattered all over the globe. LOICZView also allows the data and clustering results to be viewed in a three-dimensional data space and rotated so that different interrelations of the particular clusters can be seen. For those less visually oriented, it also calculates the multidimensional distance of every cluster with every other cluster and presents it as a simple matrix.

The final numerical results can be viewed by clicking on the "View" tab, then tapping on the "Source" button, and finally pressing the "Tag" button. The data, which can be copied and put on a spreadsheet, contain a column labeled "Tag," which designate the cluster.

5. INTERPRETATION OF THE RESULTS OF THE UNSUPERVISED CLUSTER ANALYSIS

Table 1 (next page) presents the results, along each of the ten dimensions for the unweighted averages of the societies in each cluster. We also present the average Carneiro (1970) measure of cultural complexity, an indicator totally independent of the dimensions used to define the economic system, to provide more perspective.

The results can be quickly summarized. One cluster, designated classic foragers, has a much lower level of cultural complexity than the others and, as shown by the various characteristics in the table, has all of the stereotypical (i.e., communal) characteristics a foraging society is expected to have. As a result, the averages for the total sample are presented both including and excluding the foragers.

Of the remaining four clusters the results also allow a clear interpretation. One cluster, classic designated transitional foragers, possess most of the communal characteristics of the classical foragers, but to a lesser extent. The other three systems have different mixes of different types of inequality. Specifically, one has much greater socio-economic inequality; the second has much greater political inequality (the presence of political leaders who collect taxes by obtaining a part of the foraged game, fish, or gathered plants of others in the society); and the third has much greater inequality of intangibles (for instance, income and/or wealth are gained

through receiving payment for curing, songs, or other magic or from a bridewealth paid to the family of an eligible daughter).

Another way of looking at the relationships between the five clusters requires calculating the multidimensional distance between each, and the results are shown in Table 2 (next page). The two closest clusters are the classic and the transitional. The three furthest clusters are between the classic foragers and the clusters with various types of important inequalities. This is further justification of our calculating the sample averages both including and excluding the classic foragers.

Table 1: Results of the Cluster Analysis: Means for the Ten Variables for Each Cluster

Foraging systems	Classic	Transitional	Unequal politically	Unequal socio-economic-ally	Unequal in intangibles	Total	Total excluding Classic
No. of societies	10	10	6	10	8	44	34
Avg. level of cultural complexity	11.1[a]	23.6[b]	40.5	38.2	35.0	28.4	33.6
Land	0.96[a]	2.04	1.20	1.64	1.20	1.44	1.58
Food Storage	0.20[a]	1.00[b]	2.33	1.60[b]	1.13	1.39	1.74
Slavery	0.20[a]	0.20[b]	0.50	2.00[b]	0.88	0.77	0.94
Intangibles	0.90[a]	2.90	3.67	2.30[b]	3.88[b]	2.59	3.09
Bridewealth	0.50	0.00[b]	0.67	0.20[b]	3.50[b]	0.89	1.00
Inheritance	1.10[a]	1.40[b]	2.00	3.30[b]	2.13	1.98	2.24
Wealth inequality	0.20[a]	1.10[b]	2.17	3.20[b]	2.88	1.84	2.32
Food sharing	2.52	3.11[b]	2.22	2.06	1.86[b]	2.39	2.35
Market/ barter	1.00a	2.10	2.83	2.45	2.38	2.08	2.40
Taxation/ tribute	0.00[a]	0.00[b]	4.00[b]	0.00[b]	0.00[b]	0.55	0.71

Note: All of the property and distribution dimensions are measured on a scale running from 0 (low or unimportant) through 4 (high or important). This common scaling allows the results of one dimension to be compared with another. For the classic foragers, the superscript 'a' indicates that the average is significantly different (at the .05 level of confidence) from the sample average. For the non-classic foragers, the superscript 'b' indicates that the average is significantly different (at the .05 level of confidence) from the sample average excluding the classic foragers.

The results are difficult to interpret for the remaining clusters. The socio-economically unequal are closest to the intangibly unequal, but the latter is also equally distant from the transitional group. The politically unequal cluster is closest to the transitional and the intangibly unequal clusters, although the reverse is not the case. These mixed results also parallel the results obtained from the averages of the Carneiro measure of cultural complexity - the three economic systems with important inequalities stand at roughly the same level. Although the transitional foragers have a significantly lower level of complexity than the remaining three, these three, in turn, have roughly the same level on this index.

Table 2: Multidimensional Distances between Clusters

	Classic	Transitional	Unequal socio-economically	Unequally politically	Unequal in intangibles
Classic	0	0.6	1.5	2.0	1.6
Transitional	0.6	0	1.2	1.3	1.0
Unequal socio-economically	1.5	1.2	0	1.4	1.0
Unequal politically	2.0	1.3	1.4	0	1.3
Unequal in intangibles	1.6	1.0	1.0	1.3	0

Interpretation of any type of cluster analysis raises a knotty question, are these the only clusters that can be isolated? In the context of the illustration, are these the only economic systems of foragers? Four different answers can be given, each focusing on a different facet of the problem:

Number of clusters: As noted above, any number of clusters can be calculated. Based on certain criteria of optimality based on information theory, we selected five in order to reduce problems of interpretation. Nevertheless, other numbers of clusters generally divide or combine the clusters derived for this example, as noted above.

Dimensions of the clusters: If other dimensions are used to define economic system, the clustering would probably be different. Obviously, the results of the cluster analysis depend on the data that are entered and, if certain social criteria such as kinship terminology or

particular cultural variables are added, different clusters might result. This kind of problem can be easily handled, however, by looking at the correlations between societies with particular economic systems and these other variables. For instance, once the classic foragers are removed from the sample, there is no significant correlation between the size of the foraging band and the economic systems (Pryor, 2003, Table 2).

Sample size: Because the sample includes just 44 societies, it may not have included societies with much different types of economic systems. Given the manner in which the authors of the SCCS tried to diversify the cultural areas represented in the sample, however, this possibility appears low.

Sample bias: The sample of foraging societies includes only those that were primarily foragers at the time of the ethnographic report on which the data were based. But many societies in the rest of the SCCS, which were originally foraging, advanced to agriculture and, therefore, were not included in our sample. Moreover, these might have had much different economic system, which allowed them to advance, while the societies in the sample did not have such systemic characteristics remained stuck at a particular developmental level. Because these agricultural societies in their foraging stage are omitted, a problem of sample bias arises and circumventing such difficulties raises some thorny issues. One possible approach is to look at out-of-sample points and examine the economic systems of societies which are much more reliant on agriculture for subsistence to see how they match up against the foraging societies in the sample. One such statistical technique is the use of supervised clustering.

6. AN EXPERIMENT WITH A SUPERVISED CLUSTER ANALYSIS

A supervised cluster analysis can be calculated either in terms of "archetype averaging" or "k-nearest neighbor." Archetype averaging represents each cluster as a single means and standard deviations, identical to the method used in the k-means clustering algorithm. For most applications, this method is appropriate as its results are directly comparable to the results of an unsupervised clustering. The k-nearest neighbor method, on the other hand, represents each cluster using multiple means. This is appropriate when it is known *a priori* that the clusters possess complex shapes in the data space that are not representable as multidimensional ellipsoids. For this discussion we use the latter technique. As it turns out, the two techniques yield almost the same results.

The SCCS contains 13 societies where subsistence directly from foraging accounts for 55 to 75 percent of all consumed food, which, because they lie between foraging and agricultural societies, we label intermediate societies. This sample is not very representative since it includes three Polynesian/Melanesian societies and five Amazonian societies.

Which of the five economic systems defined above do any of these intermediate societies resemble

the most? Once such relative multidimensional distances are determined, we can find out if they also share the same pattern of systemic characteristics as the societies relying more on foraging? Table 3, which is set out in the same manner as Table 1, allows direct comparisons between the two results. Because the sample of intermediate societies is so small, tests of statistical significance have little meaning and we can only interpret the data impressionistically.

Table 3: Results of the Cluster Analysis for the Intermediate Societies

Foraging systems	Classic	Transitional	Unequal politically	Unequal socio-economic-ally	Unequal in intangibles	Total	Total excluding Classic
No. of societies	2	3	3	3	2	13	11
Avg. level of cultural complexity	22.0	53.7	195.0	73.7	119.0	96.1	109.5
Land	0.75	1.67	3.00	2.00	2.50	2.04	2.27
Food Storage	0.00	1.33	2.00	2.00	2.00	1.54	1.82
Slavery	0.00	1.33	1.00	2.67	0.00	1.15	1.36
Intangibles	2.50	2.67	3.67	4.00	4.00	3.38	3.55
Bridewealth	0.00	0.00	0.33	0.00	4.00	0.69	0.82
Inheritance	1.00	1.33	3.33	3.67	4.00	2.69	3.00
Wealth inequality	0.50	0.67	2.67	2.33	3.00	1.85	2.09
Food sharing	2.31	2.13	2.56	3.06	2.60	2.55	2.59
Market/ barter	1.25	1.67	1.50	1.00	3.25	1.65	1.73
Taxation/ tribute	0.00	0.00	4.00	0.67	0.00	1.08	1.27

Note: All of the property and distribution dimensions are measured on a scale running from 0 (low or unimportant) through 4 (high or important). This common scaling allows the results of one dimension to be compared with another. The variables are defined exactly with the original codings in the Appendices. We use the k-nearest neighbor technique for calculating the supervised cluster; when the archetype-averaging technique is used, the results differ for only one society, the Omaha Indians, who are placed among the transitional, rather than the unequal-socio-economically group.

Comparing the sample averages, the societies with more agriculture and animal husbandry have considerably more land ownership, slavery, intangible wealth, inheritance, and taxation; surprisingly, hey seem to feature less market and/or barter. Such results, combined with the results of the Carneiro indices of cultural complexity show that these societies with less reliance on foraging and more reliance on agriculture and animal husbandry are, in a real sense, at a higher stage of economic development.

The general outlines of the five economic systems appear roughly the same (with some exceptions noted below), but because of the small number of intermediate societies, we cannot be completely sure. The two societies with a classic foraging system have, like those reported in Table 1, the most communal characteristics but with one exception - their food sharing is somewhat lower than societies with other economic systems. In most respects the societies with transitional foraging systems lie somewhere in between the classic and the other three economic systems. The unequal-politically societies also reveal most of the same patterning of characteristics as those of the foraging societies. It is with the unequal socio-economically and the societies with marked inequalities in intangibles where differences arise: in both economic systems, intangible wealth is very important; the intangibly unequal actually shows more wealth inequality and market exchange than the socio-economically unequal systems. This is because, as societal complexity and the division of labor increases, healers in almost all societies begin to collect a fee for their services. As a result, it is the presence of bridewealth that now defines inequality in intangibles unequal societies and which led to supervised cluster to arrive at the results presented in the table. More clarity on these issues can be gained once we separate those societies with a significant reliance on fishing, but this exercise must be left for another essay (Pryor, 2003-b).

Despite the unrepresentative nature of the sample of intermediate societies, they appear to share many of the systemic characteristics as those societies relying more on foraging. This, in turn, suggests that the economic system per se does not seem to the major barrier to agriculture and, furthermore, that it seems unlikely that those societies, which have moved further away from foraging and toward agriculture/animal husbandry, had economic systems essentially different from those which we have isolated. Given the nature of the data with which we are working, such guarded generalizations are about all that we can venture.

7. SOME BRIEF CONCLUSIONS ABOUT CLUSTER ANALYSIS

Cluster analysis is one of several statistical tools that can assist users in discovering patterns in data. In cross-cultural studies it shows which societies are most similar to others along certain dimensions. The derived typology, in turn, can provide a useful starting point for further analysis. In the example provided, knowledge of the economic system may provide some useful insights into certain cultural and social characteristics that have hitherto seemed unrelated to the economic institutions of the society. Like other types of pattern recognition techniques, interpretation of the results raises a number of serious problems. Nevertheless, it also takes us a considerable distance in

resolving difficulties from other types of analyses. For instance, Galton's problem - whether two societies are really different or whether they are simply variants of the same society with slight changes due to diffusion can be easily tackled by looking at the geographic location of societies in the same cluster.

The LOICZView program, originally designed for a quite different purpose, lends itself easily to cluster analysis in cross-cultural research. Because the program is still being revised to increase its use for many different purposes, we welcome your suggestions for its improvement.

8. REFERENCES

Carneiro, Robert

1970 Scale Analysis, Evolutionary Sequences, and the Rating of Cultures. In *A Handbook of Method in Cultural Anthropology*. (Raoul Naroll and Ronald Cohen, eds.). Garden City: New York: Natural History Press. Pp. 834-872.

Divale, William

1997 Developing Cross-Cultural Mental Illness Classifications from Symptoms. Paper delivered at the 26th Annual Meeting of the Society for Cross-Cultural Research, San Antonio, Texas.

Johnson, Stephen C.

1967 Hierarchical Clustering Schemes. *Psychometrika* 32:241-255.

Maxwell, Bruce A., and Robert W. Buddemeier

2002 Coastal Typology Development with Heterogeneous Data Sets. *Journal of Regional and Environmental Change*, forthcoming.

MacQueen, James B.

1967 Some Methods for Classification and Analysis of Multivariate Observations. In, *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press. Pp. 281-297.

Pryor, Frederic L.

2003a Economic Systems of Foragers. forthcoming.

2003b The So-called Agricultural Revolutions – with Comparisons to the Industrial Revolution. forthcoming.

Rissanen, Jorma

1989 *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific Publishing Company.

2001 Information, Complexity and the MDL Principle. In, *Cycles, Growth and Structural Change: Theories and Empirical Evidence*. (Lionello F. Punzo, ed). New York: Routledge. Pp. 339-351.

Schneider, Andreas

1999 Emergent Clusters of Denotative Meaning. *Electronic Journal of Sociology* 4, No.:2. [http://www..sociology.org/vol004.002/Schneider.html].

9. APPENDIX: A HANDBOOK FOR LOICZVIEW

This is a seven-step process to use LOICZView to perform cluster analyses like those in this essay. LOICZView is available online at "http://www.palantir.swarthmore.edu/loicz/".It was created by Bruce Maxwell of the Swarthmore College Engineering Department and Casey Smith, a student at Swarthmore College; and it is free for academic use. As indicated on the initial webpage, researchers may obtain a password by e-mailing maxwell@swarthmore.edu with a request..

A. Database Format

Data sets are arranged into rows and columns. Each row is an object, such as a society (as in this paper). Each column is a variable that describes the objects. The first row is the "header." It contains the variable names, which should be enclosed with quotation marks. The remaining rows are objects. The first five rows of a data set might look like:

"Cell ID"	"Longitude"	"Latitude"	"!Meta1"	"@supervised"	"Var1"	Var2
1	-19.83	20.58	Kung	2	14.57	2.23
2	7.75	81.25	Vedda	1	25.41	18.53
3	9.00	-83.25	Bribri	-9999	12.34	2.24
4	-23.50	-58.50	Lengua	-9999	24.92	22.63

LOICZView requires that the first three columns be specific variables. The first column must be some type of identification consisting of a unique integer for each object (the society number, for instance). The next two columns are the longitude and latitude. These are used as plotting axes when visualizing the data. These values must be given in decimal form. For example, 37' 15" would be given as 37 + 15/60 = 37.25. A sign convention is used such that positive latitude is for the Northern Hemisphere while negative latitude is for the Southern Hemisphere, and positive longitude indicates east longitude (as measured from the prime meridian) while negative longitude indicates west longitude.

After the longitude and latitude comes any number (zero or more) of optional columns for meta data. Meta data is any information about an object that would be useful after the clustering is completed (such as the name of the society), but should not be used for clustering. Meta data columns can be any alphanumeric string. Meta data columns are identified by preceding the variable name on the header row with an exclamation point ("!"). "!Meta1" in the example above is a meta data column with the names of the societies.

After the meta columns comes any number of optional columns used for supervised clustering. These columns are identified by a leading "@" in the variable name in the header row. Supervised clustering is a way of indicating which cases fall in clusters already defined

(either a priori or through an unsupervised cluster analysis). Values in this column should either be a cluster number already specified or "-9999", indicating "no data" (not used to determine a cluster). Any case with a cluster number of -9999 will be classified in the most similar cluster by the supervised clustering algorithm. For instance, it seems likely that the supervised cluster analysis would place the Bribri in cluster 2, while the Lengua would be put in cluster 1.

The remaining columns are the variables describing the societies. These are either integer or decimal values. A value of –9999 indicates that no data was available for that society in that variable (for instance, if the measurement was not, or cannot be, obtained).

Data sets can be easily created in spreadsheet programs. When saving a data set, either choose to save the data as comma-delimited or tab-delimited.

B. Uploading the Data

The interface to LOICZView is divided into five tabs – Data, MDL, Eigen, Cluster, and View. Tabs are accessed by clicking on them. On the Data tab, there is a button labeled "Upload Data." Clicking this button will open a page on the bottom of the screen. This screen contains useful information about formatting and uploading data sets. To upload a data set, click the "browse" button and select your file. Then click the "Upload File" button. This file will remain on the LOICZView server for analyses until the user deliberately deletes it. If the data set is incorrectly formatted, the program will attempt to provide helpful error messages. If necessary, the user should go back to the original spreadsheet file and fix the error. For instance, the error message may show two commas (used as dividers in the program) next to a number; in this case, the number in the original spreadsheet should be erased and reentered.

C. Selecting Variables

After uploading a data set, the user can click the button labeled "Select Variables." This will bring up an options screen where the user can indicate which variables are to be used in the current analysis. Also, the user can adjust the relative weighting of the variables. The weighting determines the relative importance of a variable when LOICZView determines the similarities between societies.

D. Determining the Optimal Number of Clusters

The "MDL" tab (Minimum Description Length) contains a tool useful for determining the optimal number of clusters in a data set. Clicking on the button labeled "Do MDL" will perform the analysis. Once the analysis has completed, a name for the analysis (based on the name of the data set) will appear in the list of "MDL Files." Selecting the analysis and clicking "View" will display the results of the analysis. This page will indicate a range of cluster numbers that were determined to be suitable by a mathematical optimization technique called

"Minimum Description Length." This provides the user with a good starting point for the number of clusters to use in further analysis. However, the user should not feel confined to only performing clusterings with the number of clusters indicated by the MDL analysis. The "mathematically optimal" number of clusters is not suitable for all purposes.

E. Calculating an Unsupervised Cluster Analysis

The "Cluster" tab contains the settings for performing cluster analyses. The main clustering algorithm provided by LOICZView is k-means clustering. The k-means is an iterative optimization method, meaning that it starts in one state and refines that state by some update rule. Each refinement is called an "iteration." Iterations are performed until the algorithm reaches convergence (no further improvement can be made via the update rule) or until it has updated the state some set maximum number of times. As is true of most iterative optimization methods, k-means is subject to finding a "locally" good solution rather than the "globally" best solution. Thus, the algorithm should be run several times with different initializations.

The clustering tab provides fields for setting the number of iterations and the number of runs. The fields are labeled "Maximum Number of Iterations" and "Number of Clustering Runs," respectively. In order to be assured of a "good" solution, the number of runs can be increased to as many as 100 and the number of iterations to as many as 200. The program will return the results from the most successful run. Also on the clustering tab is a field for the number of clusters. The k-means technique attempts to find the best way to partition the data into the specified number of clusters.

To perform the cluster analysis, click the button labeled "Cluster Data."

F. Viewing the Clustering Results

On the "View" tab is a list of all the cluster analyses that have been performed. Selecting an analysis will display some information about the analysis, including the time and date the analysis was performed (this is useful in assuring that the selected analysis is the most recent analysis). Clicking the "Visualize" button will display the results of the selected analysis. The results are displayed as a map where a dot represents each society, and the color of the dot indicates the cluster membership of the society. Clicking on the colored cluster labels at the bottom of the page will display the properties of the cluster as well as a list of the societies that belong to the selected cluster. Clicking "View Clustering Info" will display a summary of the characteristics of all the clusters. Selecting "View Cluster Distance Matrix" will display a matrix indicating the relative dissimilarities of each cluster to every other cluster.
In order to retrieve the clustering information to place in a spreadsheet, click the source button in the "View" tab. Clicking the "tag" button will display a list of all the societies plus a "tag" column which indicates the cluster number for that point resulting from the classification. There is also an "archetype" column. This column indicates one point for each

cluster that is most representative of that cluster.

G. Supervised Clustering

If the data set has a variable that begins with "@" (which will be called a "supervision variable" in this context), a supervised clustering can be performed. Values of the supervision variable are either cluster numbers or –9999. If a society has a cluster number, say 2, in the supervision variable, then that society will be part of cluster 2 when the clustering is completed, and the values of the other variables for that society will help determine the properties of cluster 2. If a society has a cluster value of –9999 in the supervision variable, then that society is not yet classified, and it will be classified to the cluster that it most closely resembles in its variable values. For supervised clustering, the number of iterations, number of runs, and random seed do not affect the results. To perform supervised clustering, click the button labeled "Supervised Cluster" on the data tab.

The user is given the choice of k-nearest neighbors or archetype averaging. These options are different ways of dealing with multiple examples for each cluster, for instance if there were two societies in the @ column with the value 3. Archetype averaging averages the values of the representative points to determine a cluster mean. In this example the values in the two points for cluster 3 would be averaged, and the result would be the center of cluster 3. K-nearest neighbors does not average the values. Instead, it records the distance between a point and the k closest representative points of each cluster, where k is some positive integer. Then, the point is classified based on the minimum value of the sum of the distance to the k nearest representative points for each cluster. Archetype Averaging and k-nearest neighbors reduce to the same definition when only one representative point is provided for each cluster: the values of the variables in the provided point are the average values of the cluster.

The results, which are obtained from the tag file in the source button on the "View" tab, provide the cluster number for the entire sample. The program has a mild peculiarity in that, in some cases, the cluster numbers are changed (so that the former cluster 2 becomes cluster 4, etc.) but this can easily be taken into account.

In the example in this paper, the cluster number of the economic system for foraging societies from the original data set was used as a supervision variable. Therefore, the cluster number of the economic system was known for societies in the original clustering task but not for the set of more agricultural societies added to the data set. For each society for which the economic system was not known the supervised clustering algorithm classified it into the economic system that had members with variable values most similar to that society.