A MORE OBJECTIVE PROCEDURE FOR DETERMINING ECONOMIC SUBREGIONS: CLUSTER ANALYSIS

P. Thomas Cox, Bernard Siskin, and Allan Miller*

In his 1967 presidential address to the AAEA, Charles E. Bishop raises the question: "Why have agricultural economists not devoted more resources to the study of structural changes in rural communities and to public policies relating to the location of economic activity and of population?" [1, p.1001]. Later in the address, he partially provides the answer through the comment that "We must reorient our thinking in terms of location and scale of organizations and interrelations among firms and among communities . . . it will be necessary for us to make basic changes in our philosophical approaches to problems, our analytical tools, . . ." [1, p. 1007].

One of the newer analytical tools which will allow us to give attention to economic activities associated with area and/or community problems is cluster analysis. The determination of economic subregions to facilitate the study of some predetermined area has presented a thorny problem for social scientists for some time [10, p. 293]. Isard states:

> *This problem is present in most regional investigations and is rarely fully solved. This situation obtains not only because of different philosophical approaches and welfare values connected with regional studies,. . . but also because an analyst typically finds reasonable alternative interpretations of the same objective data for delineating regions.*

Thus, the use of concentric circles around population centers, measured distances between cities, and mere observations of aerial maps are examples of subjective decision criteria which have heretofore been utilized to delineate subregions. These visual measures, together with a scattering of economic variables, result in a subjective determination of economic subregions. Such pitfalls can be avoided if we do not hold to our present simplified conception and follow Kelso's plea for the use of more complex decision criteria in our models [11, p. 857].

THE PROBLEM

The authors, engaged in a study of the James River Basin in Virginia, were concerned with the delineation of economic subregions of the Basin for analytical purposes. Selection of economic subregions exhibiting different characteristics would allow additional refinement in projections and assessment of proposed development of these local economies. Knowing the James River Basin to be predominately agricultural with a large rural population and a scattering of industrial centers, the task was to group the Basin's counties into economic subregions with similar resources. In this specific study, it was important that these counties be fairly close to one another in order to properly assess the effect of water resource development upon the subregions. After the selection of several socio-economic variables to represent the agricultural and industrial activities of the region, counties were originally grouped according to visual comparison of the similarity of variables by several researchers. It was found that no two grouping patterns were the same, yet the same set of variables was considered by all.

It was soon evident that several subjective considerations were required to account for different grouping patterns. Most happily, the solution to this problem was found through the use of an objective, repeatable procedure called cluster analysis.[1]

* P. Thomas Cox is an agricultural economist and leader, Northeastern Resource Group, Natural Resource Economics Division, U.S. Department of Agriculture, Upper Darby, Pennsylvania. Mr. Siskin is a faculty member of the Statistics Department, Temple University, Philadelphia, Pennsylvania. Mr. Miller is currently manager of Industrial Relations and Personnel, Milprint, Incorporated, San Francisco, California.

[1] This is not King's [12] model which clusters on correlations, but is similar to Green's [8] which was developed almost simultaneously with the model reported herein.

THE CLUSTER MODEL

Having k different entities, each of which can be described by a certain set of n descriptive variables, it may be desirable to group the k entities into a number of subsets or clusters such that the entities within each subset are highly similar and not so similar to any other subset of entities. One measure of similarity between two entities is the squared common Cartesian distance. The similarity between the two entities i and j is computed as: [8]

$$d_{ij} = (X_{1j} - X_{1i})^2 + (X_{2j} - X_{2i})^2 + \ldots + (X_{nj} - X_{ni})^2$$

where $X_{ij} \equiv$ the ith descriptive variable of the jth entity.

It is clear that the smaller the d_{ij} the more similar are the two entities i and j.

The basic concept of cluster analysis can easily be seen in the following example. Consider the simplified case where one has four counties each described economically by percentage of acreage in farmland and average yield per acre. If we plotted the two variables for each of the four counties we might find the conditions as shown in Figure 1. It is obvious that if we want two clusters, we should group entities one and two together and entities three and four together. This type of graphic analysis, thanks to present day high-speed computers, can be easily adapted to the case where there are n variables of description and k entities. However, since we are dealing with n dimensional space, it becomes impossible to visualize.

The cluster program developed by Nigil Howard and Frank Carmone at the University of Pennsylvania first standardizes[2] all variables in order to eliminate the problem of scale [9]. Sixteen variables may be used with this specific program. It then calculates a distance or similarity matrix whose elements are the d_{ij}'s referred to above. Next, it clusters all the entities into the two groups which yield the smallest within group distance. The within group distance is defined as the sum of all the distances between the entities in each cluster. That is, if the first cluster contained entities 1, 3, and 5, the within group distance for that cluster would be:

$$d_{13} + d_{15} + d_{35} = \tilde{d}_1$$

or, generally, the within group distance for the kth group would be:

$\tilde{d}_k = \sum\sum_{i<j} d_{ij}$ for all i, j ε cluster k. Then the total within group distance for k clusters would be:

$$\sum_{i=1}^{k} \tilde{d}_i .$$

This procedure is equivalent to maximizing the distance between clusters since the distance between two clusters ℓ and m would be:

$$d^*_{lm} = \sum\sum_{ij} d_{ij}$$

over all i ε cluster l and over all j ε cluster m. Thus, the total difference between all clusters would be:

$$\sum\sum_{l<m} d^*_{lm} \qquad \text{for } \ell = 1, \ldots, k-1.$$

Since the total distance, $\sum\sum_{i<j} d_{ij}$, between all k entities is constant regardless of what clusters are formed, and since

$$\sum\sum_{i<j} d_{ij} = \sum_{i=1}^{k} \tilde{d}_i + \sum\sum_{l<m} d^*_{lm}$$

it is obvious that to minimize the sum of the within group distances is equivalent to maximizing the sum of the between group distances.

The program will allow the variables to be weighted unequally in order to give more importance to selected variables. The weights are placed on each variable in

[2] To standardize the jth variable of the ith entity, X_{ij}, the following values are calculated:

$$\bar{X}_j = \sum_{i=1}^{k} \frac{X_{ij}}{k} \qquad S_j = \sqrt{\frac{\sum_{i=1}^{k} (X_{ij} - \bar{X}_j)^2}{k-1}} .$$

The standardized value is then

$$\frac{X_{ij} - \bar{X}_j}{S_j} .$$

the squared distance function; that is, the distance measure would be:

$$d_{ij} = [w_1(X_{1i}-X_{1j})]^2 + [w_2(X_{2i}-X_{2j})]^2 + \dots + [w_n(X_{ni}-X_{nj})]^2$$

where w_i = weight on the ith descriptive variable. Therefore, the importance of each variable in effecting the distance function (thus, the clustering process) is in relationship to its importance as specified by the weights.

After clustering the entities into two groups, this specific program clusters them into three groups, four groups, up to a maximum of twenty groups, on the same basis as before, minimum distance within groups. The proper number of clusters to use must be determined by the user since the problem of the optimal number of clusters to form has not yet been solved. However, examination of the differences in the total sum of squares (within groups) will reveal the magnitude of difference between groups and aid in the selection.

STATISTICAL CONSIDERATIONS

The question of whether the clusters are statistically significantly different is one often raised, sometimes justifiably and sometimes not justifiably. If the entities in the study are the population, and the measures are parameters, not sample statistics (i.e. census information), then the question of statistical significance has no meaning. To ask if county A is different from B can invoke only an answer of yes or no. However, if we consider each cluster as a sample from its respective population, then the question of statistical significance has meaning. Moreover, if we are willing to assume tha' the set of descriptive variables has the multivariate normal distribution with a common variance-covariance matrix, the question can be answered. In this case, we could find the mean vector for each cluster and use the Hottelings T^2 statistic to test the hypothesis (the population mean vectors are the same) against the alternative hypothesis (the mean vectors are different).[3] The clustering technique being basically a data analysis method, rather than an inference technique, will not assure that the clusters will prove to be statistically significantly different.

The weights must be exogenously determined by the researcher to reflect the importance he wishes to place on each variable. The weights used by the researcher, as well as the variables selected, are clearly visible and open for evaluation. Hopefully, the choices can be justified by the situation. The selection of the variables, of course, is influenced by the type of socioeconomic activity found in the region under analysis. The same considerations must be given to the weighting process. One of our constraints was that the counties be contiguous, which required the selection of a distance variable. This was accomplished by drawing a grid on a map of the James River Basin and placing coordinates in the center of each county. The distance variable was weighted more heavily than the others (Table 1) and resulted in forced clustering of contiguous counties.

The major advantage of cluster analysis is that it gives us a quick and systematic method, based on an intuitively appealing concept, of how to separate a large number of entities into various subsets. Moreover, if the variables and weights are agreed upon, the results would be repeatable regardless of who is the investigator. Thus, the criteria for grouping is based upon theoretical and economic considerations.

THE PROGRAM OUTPUT

The main output of this specific program is a matrix of entities 1 through a maximum of 160 for the rows and the number of clusters 1 through a maximum of 20 as the columns. The interpretation of the matrix provides additional knowledge in that patterns may be observed by the movement of individual entities from one cluster to another (Table 2). The within sum of squares are printed for clusters 1 through 20 and their differences may be considered in selecting a cluster (Table 3). An option is available which allows the plotting of selected clusters for a visual "feel" of the clustered data.

Nine economic subregions, derived through the above clustering procedure, are being utilized in the James River Basin of Virginia investigations and studies as a method of differentiating between areas of varying growth potential. In the assessment of the need for water resource development and the effects on the local economies of such development, differential rates of growth may be expected between clusters since it has been determined that the differences between clusters is maximized.

A TOOL FOR SAMPLING

The ability to differentiate between economic subregions provides the researcher with a very useful analytical tool. It may also be utilized as a sampling

[3] Less restrictive non-parametric tests developed by Mardia in the *Journal of the Royal Statistical Society*, 1967, and Tamura, *Annals of Mathematical Statistics*, 1966, could be used.

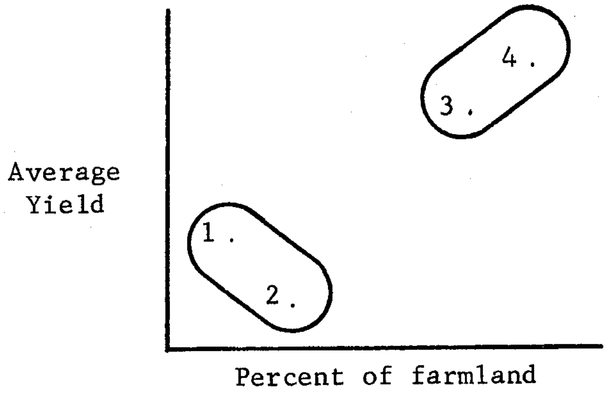

FIGURE 1. CLUSTER OF YIELD DATA BY PERCENT OF FARMLAND FOR FOUR COUNTIES

TABLE 1. ECONOMIC VARIABLES AND THEIR WEIGHTS AS USED IN THE DETERMINATION OF SUBREGIONS OF THE JAMES RIVER BASIN

Variable	Weight
Intercounty Distance	7
Agriculture:	
Value of farm products	4
Percent of income from farming	4
Percent decline in number of farms	4
Percent decline in farm acreage	4
Percent of work force in farming	3
Population change over past ten years	3
Income:	
Per capita income	2
Percent less than $3,000	2
Labor:	
Size of the work force	2
Potential labor supply	2
Interindustrial transfer of labor[a]:	
County only	3
Twenty-mile radius	3
Industry:	
Number of firms	2
Average annual wage/worker	1
Percent employed in nonagriculture	1

[a] Including farm labor

TABLE 2. MATRIX OF COUNTIES AND CLUSTERS FOR JAMES RIVER BASIN IN VIRGINIA

County	Clusters																	
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	1	4	4	4	4	4	4	4	4	4	4	4	4	4	4	18	18	18
2	1	1	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
3	1	1	4	4	4	4	4	4	4	11	11	11	11	11	11	11	11	11	11
4	1	1	4	5	5	5	8	8	8	8	8	8	8	8	8	8	8	8	8
5	1	1	1	1	1	1	1	9	9	9	9	13	13	13	13	13	13	13	13
6	1	1	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	20
7	1	1	1	1	1	1	1	9	9	9	9	13	14	14	14	14	14	14	14
8	1	1	4	4	4	4	5	1	1	1	1	1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1	1	1	1	9	9	9	9	9	9	9	9
10	1	1	1	1	1	1	1	9	9	9	9	13	14	14	14	17	17	17	17
11	1	1	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
12	1	1	4	5	5	5	5	5	5	5	5	1	1	1	1	1	1	19	19
13	1	1	1	1	1	1	1	9	9	9	9	13	14	14	14	17	17	17	17
14	1	1	1	1	1	1	1	1	1	1	1	9	9	9	9	9	9	9	9
15	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
16	2	3	3	3	3	3	3	3	3	3	12	12	12	12	12	12	12	12	12
17	2	2	2	2	6	6	6	6	10	10	10	10	10	15	15	15	15	15	15
18	2	2	2	2	2	6	6	6	10	10	10	10	10	10	10	10	10	10	10
19	2	2	2	2	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
20	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
21	2	2	2	2	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7
22	2	2	2	2	6	7	7	7	7	7	7	7	7	7	16	16	16	16	16

TABLE 3. TOTAL SUM OF SQUARES (WITHIN GROUPS) FOR THE TWENTY CLUSTERS, JAMES RIVER BASIN

Cluster	Total Sum of Squares
1	6335.999
2	3821.449
3	3024.302
4	2285.729
5	1854.368
6	1535.327
7	1279.735
8	1075.349
9	936.346
10	757.855
11	674.431
12	604.491
13	498.641
14	398.984
15	343.446
16	274.238
17	204.541
18	145.755
19	89.849
20	52.225

tool. Quite often, it is the goal of the researcher to discover differences in order to better understand the phenomenon under analysis. The use of cluster analysis can enable the researcher to group the several entities into those most alike, choose the most typical entity of the cluster, survey the entity, and then blow up the results to represent the entire cluster. Duplicating this procedure for each cluster, the information can then be compared.

For each cluster, the most typical entity can be selected, utilizing a mathematical procedure of locating the distance each entity value is from the group centroid value.[4] The square of this difference is deter-

[4] This procedure was developed by Bernard Siskin for use with the cluster program. The group centroid is printed out by the program cited in [9].

mined and summed. The lowest value of total distance is selected as the most typical entity. This standardized value can be denoted as:

$$C_{jk} = \sum_{i=1}^{n} w_i (S_{ijk} - X_{ik})^2$$

where,

S_{ijk} = standardized value of the ith variate of the jth entity in the kth group

X_{ik} = centroid value of ith variate in the kth group

C_{jk} = distance measure for jth entity of kth group

n = number of variates.

CONCLUSION

In summary, cluster analysis may be utilized to analyze a set of complex factors that may be too complicated to visualize. The interaction and interrelations of communities are known to be very complex. We must, therefore, look to tools, such as cluster analysis and other multivariate analyses [2, 3, 4], which will enable us to analyze complicated phenomena such as growth and development. The use of cluster analysis in the James River Basin study proved beneficial by providing economic subregions exhibiting different growth rates, allowing more specific projections and analyses to be applied.

Additional experience [5, 6, 7] in the use of these techniques by the authors has demonstrated the capability to pay more attention to economic problems, that are much more important to the majority of the rural population, as requested by Charles Bishop [1, p. 999].

The advantages of clearly delineating criteria and replication are foremost. The observer may analyze the algorithm, model, variables and weights and make a scientific evaluation of the selected clusters.

REFERENCES

1. Bishop, C.E., "The Urbanization of Rural America: Implications for Agricultural Economics," *Journal of Farm Economics,* Vol. 49, No. 5, December 1967, pp. 999-1007.

2. Cox, P. Thomas, "Factor Analysis of Upstream Watershed Development Projects," *Proceedings of the Western Farm Economics Association,* August 1966.

3. Cox, P. Thomas and Dainel D. Badger, "Factors Contributing to the success of Upstream Watershed Development in Oklahoma ," Oklahoma Agricultural Experiment Station and U.S. Department of Agriculture, Process Series 578, November 1967.

4. Cox, P. Thomas, "Toward Including Ethnological Parameters in River Basin Models," *Water Resources and Economic Development of the West,* Report No. 15 of the Western Agricultural Economics Research Council, December 1966, pp. 37-39.

5. Cox, P. Thomas, "Land Use Projections by Towns, Charles River Basin, Massachusetts," Unpublished Mimeograph Paper, 1968.

6. Cox, P. Thomas, "Urban Encroachment on Agricultural Lands, North Atlantic Region," Unpublished Mimeograph Paper, 1968.

7. Cox, P. Thomas, "Analysis and Projections of Industrial and Institutional Lawn Irrigation for the North Atlantic Region," Unpublished Mimeograph Paper, 1968.

8. Green, Paul, Ronald Frank, and Patrick Robinson, "Cluster Analysis In Test Market Selection," *Management Science,* Series B, Vol. 13, April 1967, pp. 387-399.

9. Howard, Nigil and Frank Carmone, Jr., "Howard Type Clustering Program," Marketing Science Institute, University of Pennsylvania, Philadelphia, Pennsylvania, 1967.

10. Isard, Walter, *Methods of Regional Analysis: An Introduction to Regional Science,* The M.I.T. Press, Cambridge, Massachusetts, 1960.

11. Kelso, M.M., "The Author's Last Word - For Now!" *Journal of Farm Economics,* Vol. 47, No. 3, August 1965, pp. 856-857.

12. King, Benjamin, "Step-Wise Clustering Procedures," *Journal of the American Statistical Association,* Vol. 62, 1967, pp. 86-101.