Use of Cluster Analysis to Identify Size Structure in Irrigated Agriculture*

by

James C. Wade**

Abstract

Analysis of structure in American agriculture often overlooks both the subtle and complex differences among farms of varying sizes within a relatively small production area. The resource components of farms of various sizes significantly affect the size and crop mix structure of each group of farms. To examine this structure sampled farms have been grouped using cluster analysis. The multivariate data observations are grouped without preconceived hypotheses. Statistical analysis of the results reveal that farms group somewhat by size and crop mix. However, smaller farms do vary substantially in composition from larger farms and avoid significant production of risk-oriented crops.

INTRODUCTION

One of the hot items in agricultural policy analysis is the "structure" of agriculture. Like many other topics that have been of concern over the years the structure question is ill-defined. Structure questions seem to center on the question, "Who controls Agriculture?", at best a vague question. Among the many subcomponents of the structure question are the questions associated with acreage limitations in the Reclamation Projects of Western irrigated agriculture. The debate continues in the legislative centers as to what types of limits should be set on the size of farms that will receive federally subsidized water. No single, clearly superior size will ever emerge from the maze of complex regional, climatic, cultural and political variables.

---

* Paper presented as a selected paper at the annual meetings of the American Agricultural Economics Association, Urbana, Illinois, July 1980.

** Assistant Professor, Department of Agricultural Economics, University of Arizona, Tucson, Arizona

The issue of "proper" farm size is one that most likely will never be totally resolved. The continuous advance of technology has resulted in most farmers increasing the size of farm businesses to "spread" capital investment cost over a larger number of acres. Unenforced institutional restraints on farm size in Federal reclamation projects have allowed expansion of farm enterprises in these areas as investment cost and technology have required. The threat of enforcement of these restraints has caused much concern among irrigated farmers in existing and proposed reclamation projects (Wade, Selley and Baggs, 1978).

The purpose of this paper is to present a method that has been used to analyze farm size structure in Arizona, and to examine in the short space available some of the results of this analysis. The paper presents a brief summary of a farm size survey conducted in 1978, a methodology (cluster analysis) used to analyze the collected data and some of the implications for policy analysis of the analytical results.

FARM SIZE SURVEY

One of the major undertakings of the study reported here has been to design, administer and analyze a questionnaire on farm size structure in Arizona. This data collected in cooperation with the Arizona Cotton Growers Association is primarily on acreage harvested by crop, type of water used and percent of income derived from farming. All of the data collected are based on 1977 farm production information.

Table 1 summarizes the statistical characteristics of the returned questionnaire by county and subgroup. The total number of questionnaires mailed (including many duplicates) was 2126. The number of usable questionnaires

returned was 354. The estimated usable return percentage for Maricopa, Pinal and Yuma Counties are 17.3, 15.0 and 17.1, respectively. The total usable return is approximately 16.7 percent of the total number of questionnaires sent. Table 1 could be analyzed in depth to provide considerable information on farm structure in Arizona. However, this analysis is left in part to the reader, since the purpose of this paper is to examine the methodology used to analyze the data.

GROUPING FARMS FOR STRUCTURAL ANALYSIS

The farm sample described in the previous section comprises a significant portion of Arizona's cropland agriculture and a reasonable sample with which to analyze the structure of such agriculture in the state. To analyze the structure of farms in Arizona, cluster analysis, a mathematical technique, was used to group the sampled farms based on the values of the characteristics as reported by the farmers. These groups or clusters were then compared to determine if the clusters are indeed statistically different. This section briefly surveys the methodology used to group the data and to test to determine if the groups are indeed different.

Cluster Analysis

Cluster analysis is a set of procedures for analyzing the level or degree of association of multivariate data observations. The procedures analyze the data with minimum preconceived hypotheses to reduce the data to levels of commonality or association. Thus, cluster analysis utilizes the observed data and a series of mathematical measures of association to group a set of m observations into k groups. To utilize this procedure

three specific concepts of measurement of association are presented. These are distance, centroid computations and data normalization.

In the analysis used in this study distance between data units is measured using the squared distance metric.

$$D_2(x_j, x_k) = \sum_{i=1}^{n} |x_{ij}-x_{ik}|^2$$

where x_j and x_k are the jth and kth data units. Any data unit is specified by the vector of data values, $X_j^T = (x_{ij},\ldots, x_{nj})$, where x_{ij} is the value of the observed variable ($i = 1,\ldots,n$) for the jth data unit.

The centroid of a set of data units X_j^T, $j = 1,\ldots,m$, is computed as

$$X_c^T = (x_{1c},\ldots,x_{nc})$$

where

$$x_{ic} = \left(\sum_{j=1}^{m} x_{ij}\right)/m, \text{ for } i=1,..,n.$$

For a single variate data set the centroid is the mean of all observations. For a multivariate data set the centroid is the vector of means of all observed variables ($i=1,\ldots,n$) over all data points.

The squared distance metric has three characteristics that should be kept in mind: 1) Variables and the manner of representing them are taken as given. If one variable is expressed in feet and a second variable in pounds, the metric involves the sum of the squares of a difference in feet and the sum of the squares of a difference in pounds, 2) each variable is treated in a linear manner: that is, each variable appears as itself; more complex functional forms such as polynomials are not included explicitly, and 3) each variable is treated independently of all the others. The contribution of each variable is the squared difference in score

for two data units. This quantity depends in no way whatever on the scores achieved on other variables. If taken quite literally, these three characteristics would be very limiting. However, to overcome these limitations the data utilized in this study was normalized by generating a set of dummy variables and, thereby, revising the measurement scales.

The principal ideal of normalization is to remove the artifact of the measurement unit and anchor each variable to some common numerical property. Disposing of the measurement unit involves dividing all the scores for a variable by a suitable normalizing factor expressed in the same units.

For example, if X_{ij} is the score on the ith normalized variable for the jth data unit, then the quantity $| X_{ij}-X_{ik}|$ lies between 0 and 1. The "range" of the variable is the largest observed difference between any two data units so that the difference in normalized scores may be viewed as the fractional disagreement (relative to the maximum possible) between two data units. Since the observed range of a variable is the difference between its two most extreme scores, such scores and their associated data units are examined for errors of observation or indications that the data units really do not belong with the data set. Special scrutiny is indicated when an extreme score is grossly different from the next most extreme.

Nonhierarchical cluster methods divide m data units into k clusters or groups. The procedure normally starts with an initial partition of the m data units and then alters the partition to improve the level of association of the data units in the k clusters. Two primary questions are of utmost concern. What constitutes a better association of data units? And, how does one achieve improved sets of associations? Although several methods are available for developing such association, Folgy's Method which was used this study is simple and straightforward. (Forgy, 1966, and Anderberg, 1973)

Statistical Analysis

The division of a set of data units into clusters brings up the question of which clusters are statistically different from other clusters. To evaluate this question two statistical procedures were used. The observations were ranked without regard to assigned cluster using an observed variable as a bases for ranking. These rankings were analyzed using Kruskal-Wallis (K-W) test and the multiple comparisons test to determine which clusters were statistically different. These two nonparametric statistical tests are summarized in the following paragraphs.

The K-W test is a nonparametric statistical test on k independent samples utilizing one descriptive variable. The test is an extension of the Mann-Whitney test and is more powerful than the k-sample median test since it uses the rank value of each case not just its location relative to the median. The K-W test tests the hypothesis, H_0, that the median ranks of the k populations are equal. The alternative hypothesis, H_1, is that at least one of the populations has a median rank different from the other populations. (Siegel, 1956)

When testing using the K-W test leads to the rejection of the null hypothesis that not all of the sampled populations are identical the question follows as to which populations are different.

Testing all possible pairs of means in the usual way affects the probability of rejecting a true null hypothesis. One way to circumvent this problem is to use a multiple-comparison procedure that incorporates an adjustment for the problem regarding the level of significance. The multiple comparisons test describe in Daniel (1979) was used in this analysis to examine this question.

CLUSTERING OF ARIZONA IRRIGATED FARMS

The 354 farms described in previous sections were grouped into ten groups using cluster analysis. Ten was chosen arbitrarily.

The data used to form the clusters were the "ranged" estimated gross income of the farm (based on farmer reported acreage and county average yields and prices) by county and the percentage of the cropped acres on each farm that occurs in each of 10 crop groups shown in the lower part of Table 1. Based on these data the farms from all three counties were used as a single data set and divided into 10 clusters.

The clusters, arranged in ascending farm size (acreage) and numbered 1 to 10 for convenience, are described briefly in Table 2. Several summary statistics for each cluster are also given in the Table.

The largest farms are generally grouped in Cluster 10 which clearly dominates all other farms in size and estimated gross revenue. Twelve farms comprising approximately 51,096 acres in the three counties are included in this group. The acreage in this group is approximately 18.7% of the sampled acreage in the three counties.

Clusters 5, 9 and 10 dominate the grouping by combining to provide an explanatory group for 72% of the cropland in the sample. If Cluster 8 is added to these groups, 84.3% of the farm land will be in four clusters. These four clusters have some significant characteristics in common. Each is dominated by field crops, especially cotton. The groups have neither large nor small average estimated gross revenues. Clearly, tree crops (Cluster 1) and vegetable crops (Cluster 6) have higher estimated gross revenues per acre. The average acreage of these higher revenue farms is variable. The tree crop farms average only about 134 acres while the vegetable and grass farms of Cluster 6 average approximately 623 acres. Combined, however, the farms of Clusters 1 and 6 represent only 2.5% of the sampled acres.

Table 2 shows the farms of Pinal County are clearly more dependent on field crops, especially cotton, and occur in the three dominant clusters (5, 9, and 10). The sample shows that the farm structure of agriculture in Pinal County is different from either Yuma or Maricopa County.

The farms with less than 160 acres are heavily concentrated in Clusters 1, 2, 5 and 8 and 3 with 80.2% of these smaller farms occuring in these clusters. These clusters have no single characteristic that stands out. Clusters 1, 2 and 3 have the lowest average farm size apparently caused by the occurrence of numerous smaller farms in these groups. Cluster 2 picks up many of the smaller farms that concentrate on alfalfa production, Cluster 1 is obviously made up of small orchards; and Cluster 3 is made up of farmers using more than 50% of their acreage in field crops other than cotton. The average gross revenue per acre for Clusters 2 and 3 is relatively low while the average gross revenue per acre for Cluster 1 is among the highest of the groups.

Statistical Comparison of Groups

Utilizing the data from the groups, the statistical tests described in previous sections were used to determine if the clusters as derived mathematically were indeed different based on estimated total cropped acreage on each farm.

The results of these tests, shown in Table 2, indicate that several significant differences among clusters do exist. Cluster 10 which represents the very large farms is clearly different from all other clusters. However, other differences are less pronounced. Cluster 2 is clearly different from Clusters 4, 5, 6, 7, 8, 9 and 10. Cluster 2 farms are characterized as "100% Alfalfa" farms. Table 2 shows that 71.8% of the farms in Cluster 2 are 160 acres or less and almost two-thirds of the farms are in Maricopa County. Farms from Maricopa County also dominate Clusters 1 and 7, characterized as "100% tree crops" and "50% tree crops," respectively. However, substantially fewer of the farms are 160 acres or less. The mean farm size of Clusters 1 and 2 are smaller than for all other clusters. These farms do not tend to have the high revenue crops as previously suggested by some researchers.

Cluster 2, the predominantly alfalfa farms, is statistically different from Clusters 4, 5, 8 and 9 which could be labeled "the major cotton growing group." However, Cluster 3 which consists predominately of other field crop farms is statistically different only from Clusters 9 and 10. The "major cotton growing group" is also different from Cluster 1 (although Cluster 1 and 4 are not statistically different).

SUMMARY AND CONCLUSIONS

The approach in this analysis is to allow the current structure to speak for itself. The data of this study can never represent with complete accuracy the total population of farms in Arizona. This study is undoubtedly biased to some degree toward farmers who grow cotton on a major portion of the cropland available to them. In Arizona farm structure has been historically dominated by capital intensive selection of enterprises. The survey presented in this paper shows that the current farm structure is not oriented to "small" scale production units. However, neither is the concept of "giant", corporate farms a reality.

The clustering shows that several groups of farm types and sizes exist side-by-side. The largest farms representing 18.7% of the land sampled and 3.3% of the farm units sampled do not have the highest average gross revenue per acre. In fact, a group of somewhat smaller farms have substantially higher gross revenue per acre (Cluster 6).

Ignoring the small tree growers who have a substantial capital investment and a long-term outlook, the production of specialty crops (vegetables and melons, and grass seed) are concentrated in large farms (Clusters 6 and 10). Farms with less than 1,000 acres concentrate in the traditional field crops and cotton and, only, occasionally, and in small acreages, grow the higher risk crops. The agricultural structure examined in this study supports the hypothesis that larger farms can more easily offset the potential risk of certain speciality crops with other