#1940

January, 1989

COMPARING NEAREST CENTROID SORTING CLUSTER ANALYSIS TO LITTLE JIFFY FACTOR ANALYSIS FOR CREATING A TYPOLOGY OF SMALL-SCALE FARMERS[1]

Michael D. Schulman
Department of Sociology, Anthropology, and Social Work
North Carolina State University
Raleigh, N.C. 27695-8107

Charles H. Proctor
Department of Statistics
North Carolina State University
Raleigh, N.C. 27695-8203

[1]This research received support from the North Carolina State University Title XII Strengthening Grant (AID/DSAN-XII-G-103) funded jointly by the U.S. Agency for International Development and the North Carolina Agricultural Research Service and conducted in collaboration with the North Carolina Agricultural Extension Service and the North Carolina A & T State University (Greensboro). Data analysis and interpretation were also supported by the North Carolina Agricultural Research Service (Project No. 5641). The opinions expressed are those of the authors.

Introduction

Renewed academic interest in small-scale farmers has coincided with increased public awareness of the changing structure of American agriculture. Social scientists, notably sociologists and agricultural economists at Land Grant universities, have devoted increased attention to analyzing the social and economic characteristics of small-scale farmers.

According to Madden and Tischbein (11), the exact nature of the small farm problem cannot be determined because existing data do not clearly or accurately describe the conditions or trends concerning small-scale operations. The lack of consensus on the definition of a small-scale farm contributes to this basic problem. The standard Census of Agriculture definition (under $20,000 in gross farm sales) has been criticized as being insufficient to describe the diversity that exists within the small farm segment (5;9).

Many researchers have attempted to deal with these definitional and operationalization problems by generating typologies of small-scale farmers. Representative studies include in Munoz (2), Buttel (4), Tweeten et al. (16), Trank and Brinkman (15), Barlett (4), and Carlin and Crecink (7). Table One summarizes several aspects of relevant studies. Despite differences in populations researched, definitions operationalized, data collection procedures utilized, and statistical analyses performed, these studies find considerable overlap in the relevant characteristics of small-scale farms. The studies basically agree that scale of the farm operation, the age of the farm operator, and the extent of off-farm work are the

TABLE 1

Aspects of Small Farm Studies

	STUDY							
	Munoz (1983)	Carlin and Crecink (1979)	Carlin (1979)	Thompson and Hepp (1976)	Buttel (1981)	Tweeten et al. (1988)	Trant and Brinkman (1979)	Barlett (1984)
Geographical Area of Sample								
National		X	X			X		
Regional	X			X	X		X[a]	X
Small Farm Definition								
≤ $20,000 Gross Farm	X	X	X	X	X	X[b]	X[c]	[d]
≤ $40,000 Gross Farm		X			X			
≤ Median Non-Metro Inc of Region	X	X	X					
Variables in Typology								
Age				X	X	X		
Disability								X
Off-Farm Labor				X	X	X	X	X
Data Analysis								
Cross-Tabulation	X	X	X	X	X	X	X	X
Factor Analysis					X			
Qualitative Methods								X

[a] Canada
[b] Between $2,500 and $20,000
[c] Under $15,000 in 1971; under $25,000 in 1975
[d] Random sample of all farms in one county

major variables differentiating types of small-scale farmers.

However, the aforementioned studies also share a methodological shortcoming. Their typologies were generated on an a priori basis. That is, the typologies were defined prior to any data analysis. Nominal definitions were operationalized, and typologies were generated. Empirical data were then used to illustrate the appropriateness of the typology via cross-tabulation techniques. While a priori assumptions are a necessary part of theory and research, appropriate multivariate techniques, particularly cluster analysis, can support theoretical inquiry, improving analysis and ensuring that typologies are accurate descriptions of data. The utility of cluster analysis will be demonstrated using data from a sample of small-scale farmers in three North Carolina Piedmont counties.

Methodology

Samples of small-scale farmers in three North Carolina Piedmont Counties were selected via an area sampling with screening procedure.[2] A total of 107 smallholders fell into the sample. Ninety operators were interviewed in person with a questionnaire that covered crop production, income, employment, and background characteristics. All data pertain to the 1981 agricultural year. The sample is predominantly black, low income, and engaged in flue-cured tobacco production.

Cluster analysis is a general term for a set of procedures that can be utilized to create a classification schema or typology. A clustering method is a multivariate statistical procedure that empirically groups homogeneous cases or entities. It places cases into clusters suggested by the data, such that

objects in a given cluster tend to be similar to each other (1).

As a data analytic strategy, cluster analysis offers advantages for both data reduction and interpretation. Although the study is a survey and used a standard personal/questionnaire method of data collection, the resulting mass of data also resembles a series of case studies. Detailed information on the household, farm operations, and farming systems was collected, since the project team represented different disciplines with different research agendas. As a result, we are faced with data on numerous variables for relatively few cases.

Several variables (e.g., age, education, size of family, debt, etc.) jointly play the role of control variables. If one were to use a regression analysis to explain a dependent variable (e.g., an adoption index), the large number of control variables could easily swamp the results since the sample size is none too large. The damage that could be done by too many control variables is difficult to foresee exactly. Basically, it makes multicolinearity more likely. In the present study there could be, among the control variables, some that are themselves confounded with the geographical basis of the sample. Our study design was limited to three counties, each of which was slightly different in terms of industrial and agricultural structure, and in terms of the types of agricultural extension programs. Thus entering all control variables separately into a regression is almost guaranteed to cause collinearity.

The purpose of the cluster analysis is to replace the numerous control variables by a single categorical, typological variable. Our aim is similar to that underlying the method of

preliminary principal axis factor analysis for the independent variables in regression analysis (7). The cluster analysis typology, we feel, is more faithful to the realities of small-scale farmers, but we will also verify that factor analysis would give essentially the same results.

The variables are a mixture of exogenous and endogenous or of causes and effects. For instance, age is set relentlessly, number of children is a fairly abiding characteristic, and income responds to some conditions and affects other conditions of the family. We see these variables changing over the life course of the farm family in response to changes in underlying latent characteristics and as affected by conditions of the economic and social environment. It may happen that farm families rearrange the levels of their control variables in a limited number of typical ways. Some may embrace off-farm employment, others may rent land out, others rent land in, and so forth. If there are a limited number of adaptations, then one might find in the multivariate distribution of these control variables points of concentration. These will characterize the combinations of levels of all variables that represent typical responses to external conditions.

This viewpoint of condensation points in high dimensional variable space is consistent with clustering methods. They attempt to locate the centers of gravity of the sub-swarms and then to classify the cases by membership in one swarm or the other. Actually, this characterization is an oversimplification of the many approaches of clustering as found, for example, in Everitt (8).

If clusters are not well defined in this multidimensional space, then the distribution may be more correctly described as multivariate normal with patterns of correlations causing the contours of equal density to be severely elongated ellipsoids. That is, the distribution may not be multimodal, but only unimodal. Such a distribution is better fit by a factor analysis solution which identifies a few underlying latent variables as responsible for the elongations. Since it is rather difficult to distinguish between condensation points and elongations if one has only relatively few cases in many dimensions, we cross-classified the results of the cluster analysis with the results of a factor analysis. That is, just to be on the safe side, we supposed we had both.

Data Analysis and Results

The particular method of cluster analysis selected, the SAS procedure Fastclus, is a nearest centroid sorting method based upon Euclidean distances. Fastclus finds disjoint clusters of cases using a K-means method in which each case is placed in one and only one cluster. This is an iterative partitioning method that minimizes the sum of the squared distances from cluster means (13).

A set of variables, all with reference to the 1981 agriculture year, is utilized in the cluster analysis. Principal variables and their indicators are:

Income: gross farm, total family, and total off-farm family income;

Debt: money borrowed during agricultural year, total farm debt;

Land tenure: total acres owned, total acres leased and rented;

Tobacco allotment ownership: yes or no (a dummy variable);

Tobacco production: total acres grown;

Household's allocation of labor: days of on-farm and off-farm labor by operator and by spouse; days of on-farm labor by other household members;

Demographic characteristics: total household size; education, years farming, and age of farm operator.

All variables were standardized to a mean of 0 and unit variance prior to cluster analysis. Two observations were found to be outliers (i.e., extreme cases not easily grouped into any cluster) and were eliminated from the data set. The Fastclus procedure, like all iterative cluster methods, requires the specification of the number of clusters prior to the creation of these groups. Analyses were run on two through eight cluster solutions.

A major problem with cluster analysis is that there is no consensus on how to determine the number of clusters that exist in a data set (1). Few formal tests have been widely accepted, and many of the heuristic procedures are not applicable to iterative clustering methods. For the North Carolina smallholder data set, a three cluster solution was found to be the most appropriate. This was based upon an examination of a) the number of cases within each cluster; b) the distances between cluster means (group centroids); c) the differences in the means of the standardized variables between clusters; and d) the face validity of the solution.

The three clusters identified by the Fastclus procedure contain 38, 19, and 31 cases respectively (see Table Two). The first and largest cluster (N=38) consists of older operators who own tobacco allotments. This cluster is distinguished by high total farm debt, large number of acres owned, ownership of tobacco allotments, advanced age and experience in farming. Cluster One can be labeled senior agriculturalist.

The second cluster contains the fewest observations (N=19), and contains full-time, farm operators who do not own tobacco allotments. This cluster has the highest gross farm income, the highest number of acres rented, the highest number of tobacco acres, the highest number of days on-farm labor by operator, spouse, and family members, and the largest household size. Cluster Two also has the lowest rank on total family income, total off-farm family income, acres owned, education, and days of off-farm family labor by both operator and spouse. Cluster Two can be labeled full-time farmers.

The third cluster consists of part-time, younger farm operators. This cluster has the highest total family income, highest total off-farm income, and the highest amounts of off-farm labor by spouse and operator. Cluster Three also has the lowest age of farm operator, gross farm income, debt and money borrowed, the fewest tobacco acres, and days on-farm labor by operator, spouse, and family. Cluster Three can be labeled part-time farmers.

In order to verify the results of the cluster analysis, a factor analysis was performed. Factor analysis is a technique that represents the distinct patterns of relationships among a

TABLE 2

Cluster Means on Standardized Variables

Cluster	Gross Farm Inc	Total Off-Farm Family Inc	Total Family Inc	Total Farm Debt	Money Borrowed	Total Acres Owned	Total Acres Rented	Tobacco Allot Owner-ship (1=Non-owner)	Total Tobac-co Acres	Days On-Farm Work Oper	Days On-Farm Work Spouse	Days On-Farm Work Other Family	Days Off-Farm Work Oper	Days Off-Farm Work Spouse	House-hold Size	Age of Farm Oper	Educ of Farm Oper	Years Farming
	VARIABLES																	
One: Senior Agriculturalist (N=38)	0.34	-0.11	0.31	0.54	0.44	0.66	-0.22	-0.69	0.15	0.36	0.05	-0.07	-0.23	0.03	-0.09	0.27	0.03	0.27
Two: Full-Time (N=19)	0.56	-0.94	-1.32	-0.31	0.40	-0.64	0.47	1.06	0.69	0.43	0.13	0.73	-0.61	-0.64	0.53	0.04	-0.21	-0.46
Three: Part-Time (N=31)	-0.76	0.72	0.42	-0.48	-0.79	-0.42	-0.02	0.19	-0.61	-0.71	-0.15	-0.36	0.66	0.35	-0.22	-0.35	0.09	-0.05

set of variables as due to the fact that the variables belong to groups, each group being indicative of an underlying factor. Both methods (cluster and factor analysis) capitalize on departures in the multivariate joint distribution of the data from the spherically even case of independence.

The eighteen variables specified earlier were subjected to the usual "Little Jiffy" approach (10) which includes a varimax (orthogonal) rotation. Three factors were identified under the rule that their eigenvalues were above one. The results show the following constellation of variables:

Factor I: gross farm income (pos.); total farm debt (pos.); money borrowed (pos.); total tobacco acres (pos.); total acres owned (pos.); days on-farm work by operator (pos.).

Factor II: tobacco allotment ownership (pos.--i.e., does not own); years farming (neg.); household size (pos.); days on-farm work by other family members (pos.); total rented acres (pos.) days on-farm work by spouse (pos.).

Factor III: total family income (pos.); total off-farm family income (pos.); education of operator (pos.); days worked off-farm by spouse (pos.); and age of operator (neg.).

A high score on the first factor would reflect relatively high gross farm income, sizable debt, many acres in tobacco, considerable acreage owned, and the operator working much time on-farm. On Factor II, a high score reflects a household with few years farming, non-ownership of tobacco allotment, many household members, renting acreage, and with many days of on-farm

work by the spouse and by other family members. The archetypical case with a high score on Factor III would be a household with both adults young and working off-farm, and with relatively high levels of education, off-farm income and total family income.

In identifying the three factors, the variables were classified into groups on the basis of their factor loadings. Sub-scale scores for each respondent were next constructed by averaging over the variables within each group. We weighted inversely proportional to standard deviations if there was wide variety in the ranges of the variables in a group. We then standardized, to zero mean and unit variance, these sub-scale scores. There were three factors so there were three scores for each case. We located the maximum of these three and assigned the farm to one factor "archetype" according to its highest factor score (absolute value). Cluster assignments were then cross-tabulated with the factor archetype assignments (see Table Three). The correspondence is clear: Cluster One is Factor Archetype I, Cluster Two is Factor Archetype II, and Cluster Three is Factor Archetype III. While twenty of the 88 total cases (23%) received different assignments, the factor analysis lends validity to the cluster analysis solution.[3]

Discussion and Conclusion

Cluster analysis demonstrated that three groups could be identified among sample smallholders: full-time, part-time, and older full-time farmers. Variables included in the analysis reflected both national and regionally specific aspects of agricultural production. In the North Carolina Piedmont, land tenure, especially the ownership of tobacco allotments, was an

TABLE 3

Cross-Tabulation of Cluster Assignment and Factor Archtype Assignments (N's)

Cluster	Factor Archtype		
	I	II	III
One	29	1	8
Two	2	17	0
Three	0	9	22

important variable in distinguishing among the clusters. Scale, age of operator, off-farm income, and labor allocation on-farm and off-farm were other important variables for cluster differentiation.

These results are generally consistent with earlier cited studies of small-scale farm operators in the United States. All research, however, has a specificity which derives from the time when the work was conducted and the nature of the population analyzed. This study is no exception. The clusters identified are specific to contemporary smallholders in the Piedmont of North Carolina. Nevertheless, the research design utilized in this study is generalizable. Cluster analysis seems to be a useful technique to identify typologies from survey research data and its judicious use can make contributions to research.

Notes

2. In Caswell County, the sample of small-scale farmers is based upon those small-scale farmers who were participating in an Extension paraprofessional program during 1981. These were farmers who were under 65 and who had under $20,000 in gross farm sales. In Person and Granville Counties, samples of small-scale farmers within each county were drawn. First, census enumeration districts within each county were selected at random. Second, all farmers within each district were administered a short screening questionnaire. A farmer was eligible for the sample if the data from the screening survey revealed that he/she met the following characteristics: 1) gross farm sales of $20,000 or under in 1981; 2) farm operator 65 years of age or less; 3) agriculture a significant part (20%) of total family income. A fourth criteria, working less than 100 days off-farm, was dropped after the screening data revealed that the farmers meeting the other three criteria were bimodal with regard to off-farm work: one group had less than 100 days, but another group had more than 200 days. A total of 107 small-scale farmers fell into the sample: 27 in Caswell, 41 in Person, and 39 in Granville County. Ninety interviews were completed (21 in Caswell; 37 in Person; and 32 in Granville).
3. Within each cluster, the means of the cases that were "incorrectly" assigned to a Factor Archetype were compared with the means of the "correctly" assigned cases. The eight Cluster One/Factor II error cases were, on average, younger

and had lower gross farm incomes, higher total family and off-farm family incomes, fewer tobacco acres, and more off-farm labor by the spouse and operator than the Cluster One/Factor I cases. The single Cluster One/Factor II error was younger and had lower gross farm and total family income, and more on-farm labor than the average of the Cluster One/Factor I cases. The two Cluster Two/Factor I errors were older, had more tobacco acreage, and had fewer household members and less family on-farm labor than their Cluster Two/Factor II counterparts. The Cluster Three/Factor II error cases had higher gross farm incomes, lower off-farm family and total family incomes, less ownership of tobacco allotments, less off-farm labor by spouse and operator, and more on-farm labor by spouse and operator than the Cluster Three/Factor III cases. Generally, the error cases of the Cluster/Factor cross-classification appear to be the outlying (farthest from group centroids) cases within each cluster.

References

1. Aldenderfer, Mark S., and Roger K. Blashfield. Cluster Analysis. Beverly Hills: Sage, 1984.

2. Barlett, Peggy F. "Microdynamics of Debt, Drought, and Default in South Georgia." American J. Agr. Econ. 66 (1984):836-843.

3. Bibb, Robert, and Dennis W. Roncek. "Investigating Group Differences: An Explication of the Sociological Potential of Discriminant Analysis." Sociological Methods and Research 4(1976):349-379.

4. Buttel, Frederick H. Toward A Typology of Small Farms: A Preliminary Empirical Analysis. Ithaca, N.Y.: Cornell University, Rural Sociology Bulletin No. 116, 1981.

5. Carlin, Thomas. "Small-Farm Component of U.S. Farm Structure," pp. 274-277 in Structure Issues in American Agriculture. Washington, D.C.: USDA, ESCS Agr. Econ. Rep. 438, Nov. 1979.

6. Carlin, Thomas A., and John Crecink. "Small Farm Definition and Public Policy." Amer. J. Agr. Econ. 61(1979):933-939.

7. Draper, N. R., and Smith, H. Applied Repression Analysis. New York: Wiley, 1981.

8. Everitt, B. Cluster Analysis. London: Social Science Research Council, Heinemann Educational Books, 1974.

9. Heffernan, William D., Gary Green, Paul Lasley, and Michael F. Noland. "Small Farms: A Heterogeneous Category." The Rural Sociologist 2(1982):62-71

10. Kaiser, H. F. "A Second-Generation Little Jiffy." Psychometrika 35(1970):401-415.

11. Madden, J. Patrick, and Heather Tischbein. "Toward an Agenda for Small Farm Research." Amer. J. Agr. Econ. 61 (1979):940-945.

12. Munoz, Robert D. Small Family Farms in Mississippi and Tennessee: A Comparison of Small Farm Definitions. Mississippi State, Miss.: Mississippi State University, Ag. Economics Research Report No. 141, 1983.

13. SAS Institute, SAS User's Guide: Statistics. Cary, N.C.: Statistical Analysis System, 1982.

14. Thompson, Ronald, and Ralph Hepp. Description and Analysis of Michigan Small Farm. East Lansing: Michigan State University, Michigan Cooperative Extension Research Report No. 296, 1976.

15. Trank, M. J., and G. L. Brinkman. "A Classification of Limited Resource Farmers." Canadian Farm Economics 14(1979):21-29.

16. Tweeten, Luther, G. Bradley Cilley, and Isaac Popoola. "Typology and Policy for Small Farms." Southern Journal of Ag. Econ. 12(1980):77-85.