MULTIPLE CLASSIFICATION ANALYSIS WITHOUT ASSUMPTION OF INTERVAL MEASUREMENT, LINEARITY, OR ADDITIVITY: A COMPARISON OF TECHNIQUES

James M. Carman, University of California (Berkeley)

The interest of this paper is data analysis, not inference. In survey research, one is commonly faced with the problem of analysis, often exploratory analysis, of data from a relatively large number of subjects on which values of a number of variables have been collected. In social science research, the measures we have of these variables often do not meet the standards that statisticians would like. We are faced with mixed interval, ordinal and nominal data, nonlinearity, nonorthogonality, and interactions. Thus, the restrictions of the common variations of the general linear regression model are not often met. Aided by the computer, the rate of development of operational, often heuristic, schemes for analysis of these kinds of data has increased in recent years. This paper will review some of these newer techniques and empirically compare their relative efficiencies and shortcomings.

In attempting to see the relationship between variables, the analyst is inevitably faced with the problem of having more data than can be comprehended by the human mind at one time. The particular problem discussed here is one where the task is to relate a large set of predictors to some specific dependent variable in such a way as to isolate intervening conditions and discard spurious and irrelevant variables. In this process it is necessary to reduce the quantity of data to a level of rapid comprehension.

Classification Techniques

It should be noted that in some problems of this general type the practice is to perform a data reduction operation prior to analysis of the effects of the predictors on a particular dependent variable. There are two related approaches to data reduction, each of which has developed a variety of models based on whether the data are normally distributed or simple classifications.

The first of these approaches is the taxonometric approach--that is, reduce the number of subjects by placing them into subcategories so that the nature of everyone in the subcategory is more like each of the other n-1 persons than he is like any other person in any other subcategory. For continuous variables, the models of Tryon [11] and Cattell [3] are well known. For nominal and ordinal data, McQuitty has made substantial contributions [8].

The second approach is the factor analysis approach--that is, reducing the number of predictors by collapsing them into construct factors and then constructing measures for the construct based on weighted factor scores.

While a few of the variables used in the example which follows are factor scores, in the main, this approach to data reduction has been avoided for two reasons. First, there is serious question as to whether the measures of our data (or for that matter, most psychological data) meet the requirements of the factor analysis model. Second, the factor analysis model is most appropriate when one has a number of measures of one or several closely-related variables or constructs. In the survey research problem, as opposed to the psychological test problem, one has a number of variables which are intercorrelated, but which relate to characteristics and attitudes stemming from very different question formats and which refer to different time periods in the subject's life. Thus, a priori, it is difficult to judge which variables should be proxies for a single construct. Consequently, both statistical and behavioral theory would suggest that factor analysis solely for the purpose of data reduction should be avoided.

We turn, then, to a search for analysis techniques which will give insight into the effects of a number of predictors on a dependent variable and, at the same time, provide some amount of data reduction. One recent and fresh approach to this problem has been made by James Coleman [4]. Unfortunately, Coleman's technique requires that the dependent variable be dichotomous. In addition, he has not, to the best of my knowledge, solved the interaction problem. Since Coleman's problem can be solved more efficiently by a dummy regression or, more precisely, a two-group discriminant model, it was not included in the empirical comparison to follow.

The one approach which does seem to have more merit than a long series of exploratory dummy regressions is a branching technique. (It is important to note, however, that neither branching or dummy regression will handle the problem of nominal predictors with a nominal criterion variable of more than two levels.) A binary branching schema does provide a kind of data reduction which the dummy regression model does not. In dummy regression, all levels of all nominal predictors must be established as potential predictors. In a branching technique, the algorithm searches for that split in the classification which maximizes the distance between the mean value of the dependent variable in the two subcategories. While other branching techniques are to be found in the literature, the one which has been developed most completely is the Sonquist-Morgan Automatic Interaction Detector Algorithm [10]. This technique is a center of interest in the empirical comparison.

An Example

As a vehicle for comparing the effectiveness and efficiency of some models for

social science data analysis, we have chosen a problem from the study of consumer behavior. The data came from the Berkeley Food Panel, a study in which the food shopping habits of panel members were studied over a period of fifteen weeks [1].

The particular problem of interest here was whether characteristics of the respondents would predict the stability through time of their buying patterns with respect to the food chains they patronized. After some collapsing of small, independent stores into groups, it was possible for a respondent to have shopped in twenty-three different chains or independent stores. Each respondent was classified as having stable or unstable buying patterns during the period based on whether her pattern rejected a null hypothesis of temporal symmetry in a test involving the Kruskal-Wallis H-statistic [2]. Thus, the dependent variable was dichotomous, taking a value of 1 for unstable patterns and 0 for stable patterns.

The predictors were social, environmental, economic, demographic, psychological, attitudinal, and behavioral characteristics of the respondents collected during the course of the panel study. These were typical social science data in that a few were true interval measures, some were rank measures, and many were simple classifications. There was considerable correlation between predictors and, for the continuous measures, linearity was not a good assumption. In all, we had about 95 predictors: 25 continuous, 36 ordinal, 7 dichotomous, and 27 nominal, with an average of 6 levels each. There were 235 observations.

How might one approach analysis of these data? Cross-classification analysis is probably the most obvious approach, but consider what is required. First, the interval and ordinal scale would have to be treated as classifications and, as a start, 95 two-way tables produced. Even if one could cope with this many tables, the analysis would be void of any investigation of joint effects. If interactions and intercorrelations were considered, the problem gets completely out of hand. Even with the computer, cross classification requires a great deal of setup for very little data reduction.

Another approach might be to analyze the data as a dummy regression problem. The chief advantages of this approach are the very large amount of data reduction it achieves plus the availability of a variety of convenient computer programs. Unfortunately, there are a number of serious disadvantages. Initially, additivity would be assumed and, for the continuous variables, one would probably assume linearity also. The ordinal variables would have to be converted to dummies. Most serious is the fact that, in this example, there are insufficient degrees of freedom to analyze the data initially as a dummy regression.

Automatic Interaction Detector

A more fruitful approach proved to be analysis of the data with the binary branching techniques of the Morgan-Sonquist Automatic Interaction Detector (A.I.D.) schema. Using the amount of data reduction as a criterion, A.I.D. falls in between the cross-classification approach and the dummy regression approach. While the number of pages of output, amount of brute-force study, and number of reruns necessary to get meaningful results are significantly less than in cross classification, do not expect to get instant answers. The A.I.D. trees presented in Exhibits 1, 2, 3, and 4 each required at least two computer runs, each produced approximately 100 pages of computer printout, and each required this investigator about one-half hour to digest. With the present version of the A.I.D. algorithm, data analysis is still an investigator activity rather than a computer activity.

Using the amount of flexibility and generality as criteria, A.I.D. comes off significantly better than either cross classification or dummy regression. The continuous variables must be treated as categories, but any predictor can be restricted to have a monotonic relationship with the criterion variable. Thus, A.I.D. can discover nonlinear relationships with continuous or ordinal predictors without reporting spurious and meaningless minima and maxima. A.I.D. is well suited to analyze classification data and, of course, it is ideally suited to handle interactions between predictors [9].

The A.I.D. algorithm solves the degrees of freedom problem by calculating the deviation of every observation from its branch mean and making the deviation available as the dependent variable in a subsequent analysis. On any one run it is important not to introduce so many predictors that the degrees of freedom become used up before some important predictors have had a chance to enter the analysis. We have found it useful to follow the practice of cross-classification analysis and enter predictors in time order of occurrence for the respondent. For example, in the first tree the predictors relate to the respondent's childhood experience and her environment; the second tree predictors relate to general personality characteristics which are, in part, a function of background; the third tree predictors are attitudes related specifically to homemaking; the fourth tree predictors are characteristics of shopping behavior which are, themselves, a function of the predictors in the earlier trees.

The results of the A.I.D. analysis are presented in Exhibits 1 through 4. It should be emphasized in passing that the best method for summarizing and presenting A.I.D. results is not obvious or well established.

Dummy Regression

It is useful, for comparison purposes, to see how the A.I.D. results would compare with results from a dummy regression. Regression analysis is possible now because the results of the A.I.D. analysis can be used to collapse some categories and to eliminate variables which the tree analysis showed to be poor predictors. We introduced 44 predictor variables and dummies into a standard stepwise linear regression program. Thirty-two of those

entered with alpha risks of less than .30. The results are summarized in Exhibit 5.

Following are eleven hypotheses which might be advanced, based on the regression results:

Proneness toward unstable food store shopping patterns:

1. Increases with income.
2. Decreases with asset accumulation.
3. Increases with cultural status, i.e. education and occupation status.
4. Is greatest among the unmarried under 45 years of age.
5. Is least if shopper's Mother lives nearby.
6. Is inversely related to the degree of training as a child on the value of money and to dissatisfaction with present economic situation.
7. Is greatest among those with high religious commitment.
8. Is greatest among those who are interested homemakers and mothers, but not devoted cooks or shoppers.
9. Those with unstable patterns are liberal in their economic thinking, don't make a special effort to please others, and "have a complete, realistic, practical respect for the facts."
10. Increases with weekly food expenditures.
11. Is least among those with the greatest amount of store choice.

The eleven generalizations leave out some rather disturbing inconsistencies within the regression findings. Only a part of these inconsistencies can be traced to multicolinearity, which was clearly evident. One is also struck by the low fraction of variance explained.

What is even more disturbing is that when we analyze the trees, we find that three of these eleven generalizations do not appear to be correct interpretations of the data.

1. The income and asset factors do not show up in the trees at all and one wonders if the regression results are not related in some way to the social class effect which shows up significantly in both the regression and the trees.
2. The trees show that it is not being unmarried which is related to unstable buying practice. Rather, the relationship is with family structure. The least stable are families with four or more children living at home; second are young families with older children at home; the most stable are older families with no children at home. Again, the regression model is confusing because of a failure to cope with an interaction between life cycle and family structure.
3. While the Yeasay and Personality types agree between the regression and tree analyses, the relationship of the Economic Conservative scale is not as clear. The tree analysis shows this scale to be interacting with the Yeasay scale in a fashion which suggests, on balance, an effect the reverse of that shown by the regression results.

Another way to compare the results is to compare the statistics in Exhibit 5. The R^2 and β^2 statistics for the discriminant (regression) analysis have the usual interpretation. The proportion of variance explained by the trees is simply the between group sum of squares over the total sum of squares. There is no adjustment for loss of degrees of freedom; yet clearly, this statistic is a function of the number of observations and the number of groups.

The reduction in unexplained variance from any one split can be calculated from the program by:

$$D = \frac{TSS_i}{TSS_T} - \left(\frac{TSS_j}{TSS_T} + \frac{TSS_k}{TSS_T}\right)$$

where i is the parent group and j and k the resultant groups. There are other statistics which can be calculated from the A.I.D. output which have intuitive appeal because of their parallel to analysis of variance. However, the critical distinction between them is that the A.I.D. model involves sequential solution with the statistics generated at each branch, while the ANOVA model involves a simultaneous solution. In general, one would expect that in Exhibit 5 the A.I.D. reductions in unexplained variance would overstate β^2. This is not true in many cases, leading to the conclusion that the results given by the two models are different.

To summarize, the regression analysis, even after some initial doctoring of the data based on the tree analysis, explained only 18 percent of the total variance, passed over seven predictors which the tree analysis indicated were important, and yielded results which in many instances mislead the analyst in understanding the information contained in the data.

The A.I.D. analysis, on the other hand, leads to a much better understanding of the data, but can give misleading results when the number of observations in a branch gets small. It is important not to introduce too many predictors in one run. For example, one final A.I.D. run introduced 30 predictors which were shown in earlier runs to be important. Only 13 of the most powerful of these entered the analysis before the degrees of freedom had been exhausted.

Holmes' Substrata Analysis

Another branching scheme which appeared to offer some usefulness to the analysis problem at this point was Holmes' Substrata Analysis [6]. This technique was developed by the late Jack A. Holmes in a project which was trying to identify the factors and mechanism which leads some children to read at an earlier age than others. The technique did help Holmes to gain new insight into the reading process. In this scheme a set of first-level predictors are regressed on the criterion variable. Then a set of second-level predictors are regressed on

each significant predictor in the first-level analysis. If desired, a set of third-level predictors may be regressed on each second-level predictor. In this way a tree of regressions is constructed. Each regression is the standard, stepwise, linear, additive algorithm. The user may allow all predictors to be eligible to enter the analysis at any level or may specify the level at which they are to be considered. The user may also specify "fundamental" variables which are not permitted to be criteria in subsequent levels.

In some ways the Substrata Algorithm appears to be similar to A. I. D. Predictors may have a direct influence on the criterion or may only work through a first-level predictor. In many key respects, however, the two techniques are quite different. For one thing, at each level the Substrata Algorithm makes all of the usual linearity, additivity, independence assumptions of the general linear regression model. Therefore, even though it is a branching model, it is not a very general model. On the contrary, it is quite specific and requires the analyst to start with a theory which will justify the substrata model. In our problem, the model looked reasonable, i.e. stability is a function of shopping habits which, in turn, are functions of personal characteristics, personality, and early training. In practice, however, the results from this model did not match up with theory. Predictors entered at wrong levels and individual regressions did not make as much sense as the single equation regression model. The total amount of output was just as great as with A. I. D., but supplied much less information.

Data Reduction and Real Time Analysis

This problem of a large volume of output is a serious one. If the data will not permit the luxury of reduction to a single, simple correlation matrix, then any analysis scheme will not yield the amount of data reduction common in regression analysis. Further, since our problem is one of heuristic data analysis and not inference, the analyst learns more about how to proceed as he goes along. These two characteristics--large volumes of data and a heuristic process--make real time computer analysis the next logical step in the development of branching processes. F. H. Westerfelt developed at the University of Michigan a stepwise, polynomial, regression procedure which maximizes predictability with a minimum number of terms. David Evans developed, at Berkeley, a way to display this and alternative models on an oscilloscope, while the analyst interacts with the computer in real time. The day is not far off when the A. I. D. trees presented here can be generated in real time with visual display output in such a way that a large variety of alternative orders of entry and reentry into the analysis can be accomplished in the time required to study the output from one run in a batch processing system. Thus, it should soon be possible to teach the logic of data analysis developed by Hyman [7] over twelve years ago without having the student and instructor feel the frustration of having no analytical technique for making this logic operational.

REFERENCES

[1] Bucklin, L. P., and Carman, J. M. The Design of Consumer Research Panels: Conception and Administration of the Berkeley Food Panel. Berkeley: Institute of Business and Economic Research, University of California, 1967.

[2] Carman, J. M., and Stromberg, J. L. "A Comparison of Some Measures of Brand Loyalty," Working Paper No. 26. Berkeley: Research Program in Marketing, Institute of Business and Economic Research, University of California, 1967.

[3] Cattell, R. B. Handbook of Multivariate Experimental Psychology. Chicago: Rand McNally and Co., 1966.

[4] Coleman, James S. Introduction to Mathematical Sociology. New York: Free Press, 1964, Ch. 6.

[5] Harmon, H. H. Modern Factor Analysis. Chicago: University of Chicago Press, 1960.

[6] Holmes, J. A. Substrata-Factor Differences Underlying Reading Ability in Known Groups. Report to the U. S. Office of Education under Contract No. 538-8176 (1961).

[7] Hyman, H. Survey Design and Analysis. Glencoe: Free Press, 1955.

[8] McQuitty, L. I. "Typal Analysis," Educational Psychological Measurement, 21 (1961), 677-696.

[9] Morgan, J. N., and Sonquist, J. A. "Problems in the Analysis of Survey Data, and a Proposal," Journal of the American Statistical Association, 58 (1963), 415-435.

[10] Sonquist, J. A., and Morgan, J. N. The Detection of Interaction Effects. Ann Arbor: Survey Research Center, University of Michigan, 1964.

[11] Tryon, R. C., and Bailey, D. E. "Cluster and Factor Analysis." Berkeley: Computer Center, University of California, 1965.

EXHIBIT 1

FIRST A.I.D. TREE
DISCRIMINANT FUNCTION FOR SHIFT IN SHOPPING PATTERN

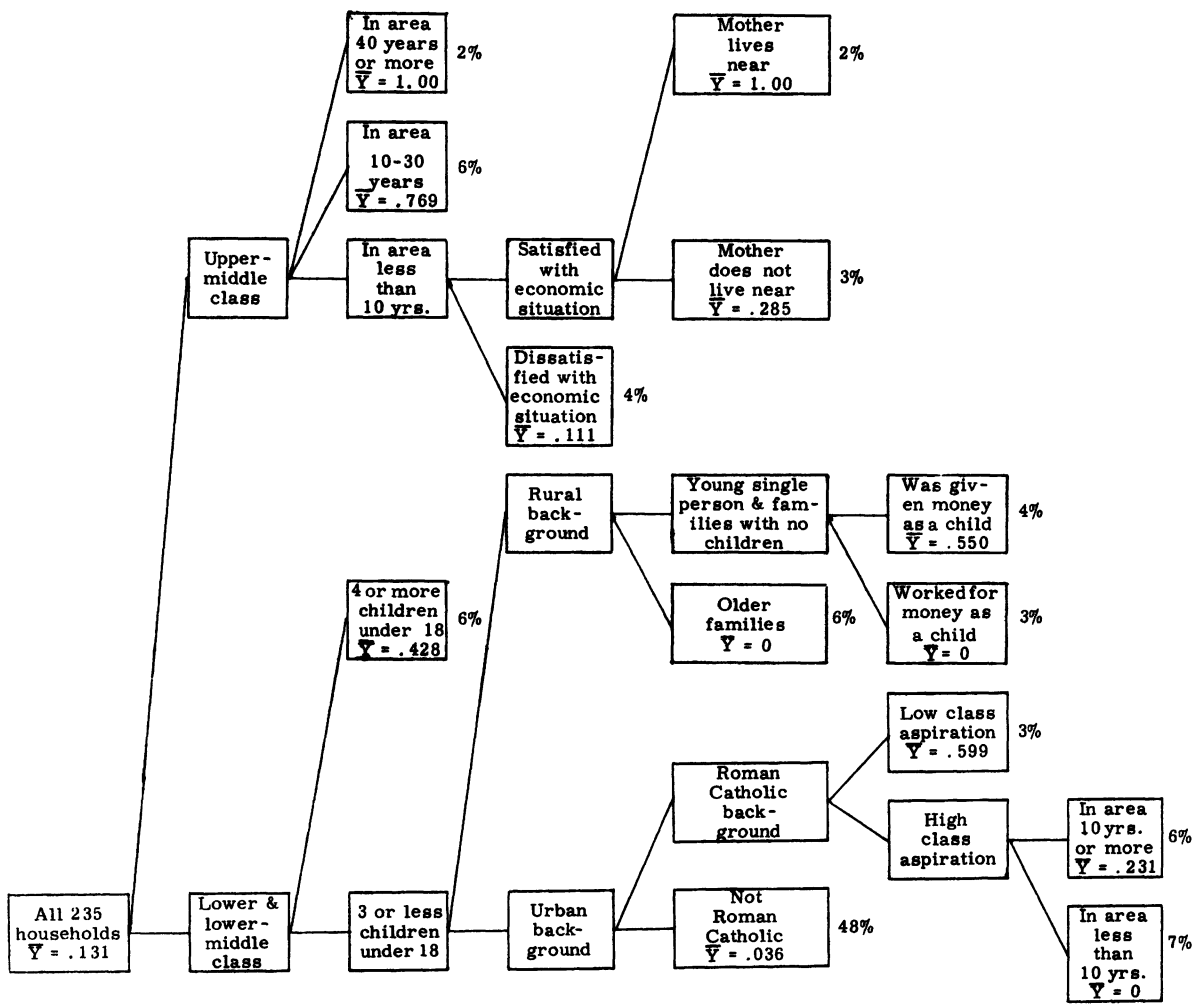

EXHIBIT 2

SECOND A. I. D. TREE
DISCRIMINANT FUNCTION FOR SHIFT IN SHOPPING PATTERN

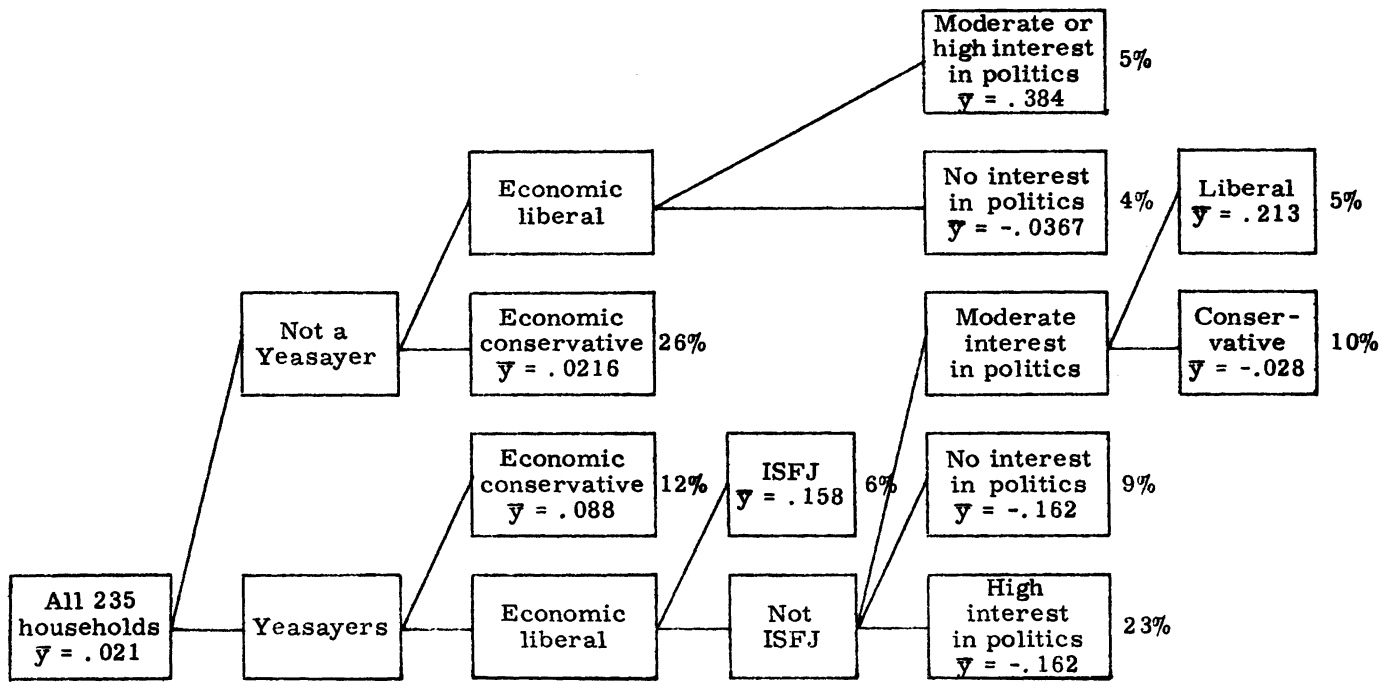

EXHIBIT 3

THIRD A. I. D. TREE
DISCRIMINANT FUNCTION FOR SHIFT IN SHOPPING PATTERN

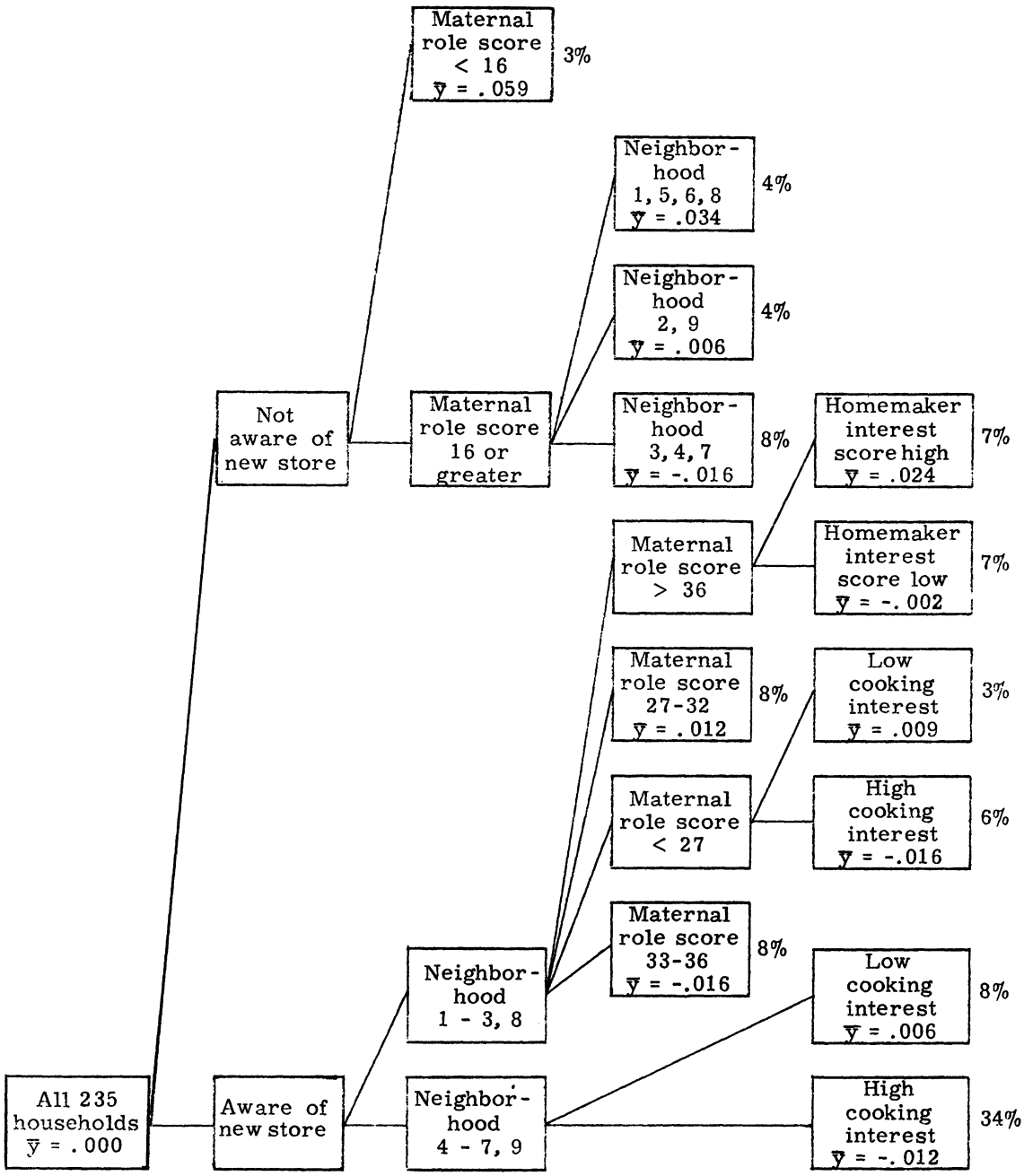

EXHIBIT 4

FOURTH A. I. D. TREE
DISCRIMINANT FUNCTION FOR SHIFT IN SHOPPING PATTERN

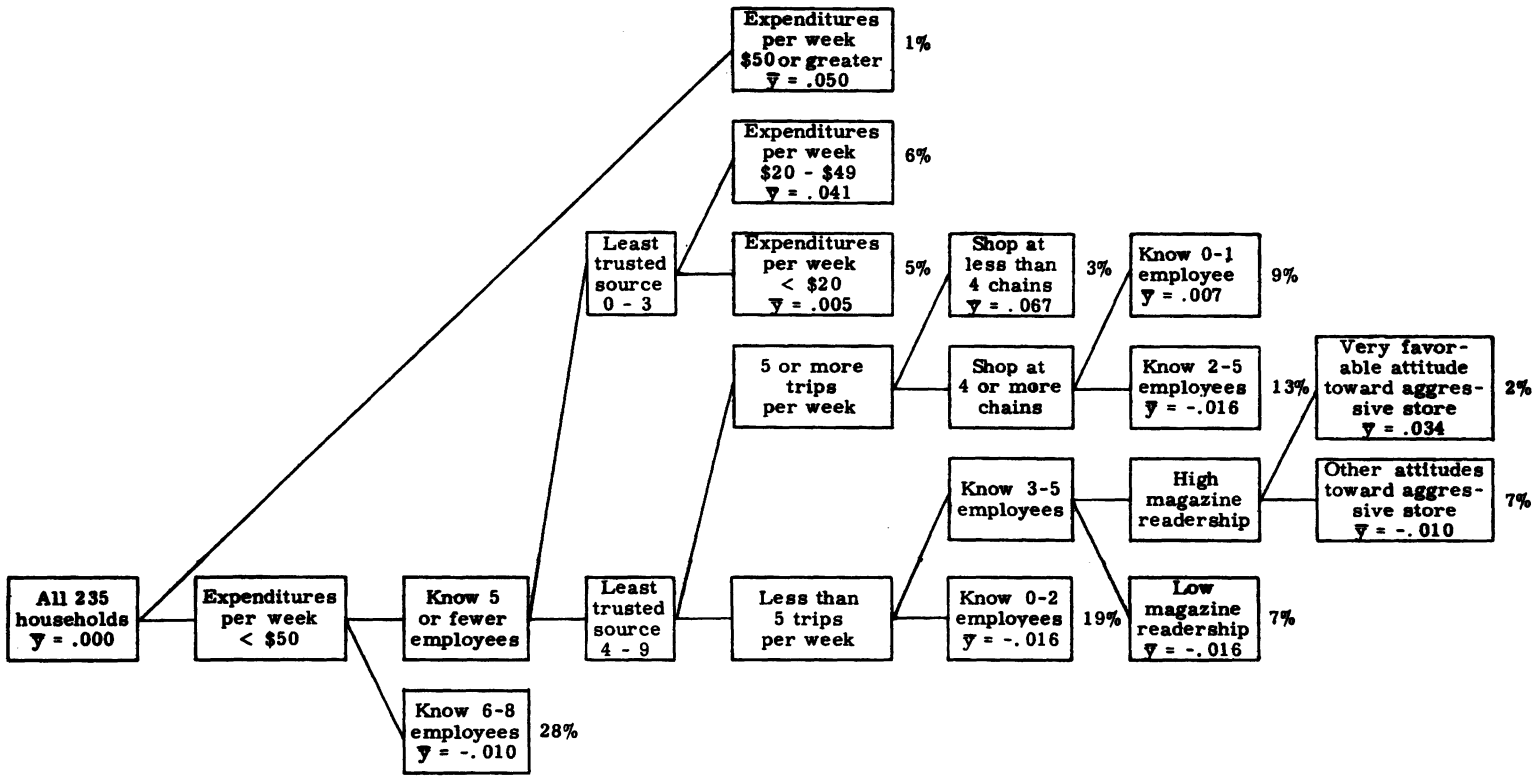

EXHIBIT 5

COMPARISON OF LINEAR DISCRIMINANT AND A. I. D. RESULTS

Proportion of Variance Explained By:	
Discriminant function, adjusted	.18
First tree, no adjustment	.41
Second tree	.17
Third tree	.30
Fourth tree	.30
$1 - (1-R_1^2)(1-R_2^2)(1-R_3^2)(1-R_4^2)$	.72

Predictor	Discriminant Function		A. I. D. Split Reduction in Unexplained Variance	Comments
	Significance	β^2		
Effects of Background, Social, Demographic, and Economic Environment:				
Social class	.01	.0428	.0365	
Income	.05	.0454	-	
Property value	.05	(-).0428	-	Sig. corr. with income
Many investments	.25	(-).0094	-	Sig. corr. with income
Life cycle	.05	.0299	.0270	Splits are not the same
Number of children under 18	ns	-	.0419	
Mother lives near	.15	(-).112	.0468	
Early independence training	.20	(-).083	.0311	
Rural background	ns	-	.0281	
Tenure in area	ns	-	.0120 .0468	
Roman Catholic index	.05	.0279	.0162	
Religious involvement score	.10	.0310	-	
Dissatisfied with economic situation	.15	.0161	.0360	
Class aspiration	ns	-	.0330	

EXHIBIT 5--Continued

Predictor	Discriminant Function: Significance	Discriminant Function: β^2	A.I.D. Split Reduction in Unexplained Variance	Comments
Effects of General Personality Characteristics:				
Economic conservative	.10	(-).0172	.0103 .0165	
Yeasaying score	.28	.0061	.0160	
Personality type ISFJ	.20	.0071	.0174	
Politically active	ns	-	.0064 .0973	
General conservative	ns	-	.0095	
Effects of Attitudes Related to Homemaking:				
Maternal role dominant	.05	.0454	.0712 .0578	Significant correlation
Homemaker role dominant	.10	.0182	.0442	between these
Cooking interest score	.15	(-).0174	.0157 .0094	three predictors
Aware of new supermarket	.10	(-).0142	.0283	
Magazine readership	.01	.0502	.0209	
Frequency of entertaining at home	.05	(-).0437	-	
Frequency of entertaining neighbors	.20	.0166	-	Significant correlation with home entertainment
Effects of Shopping Behavior:				
Number of market employees known	.10	(-).0146	.0248 .0173 .0150	
Don't trust home economists	.10	.0372	.0361	
Don't trust friends for food information	.10	.0552		
Don't trust store clerks	.20	(-).0071		
Husband influential in setting food budget	.20	(-).0079	-	Significant correlation with life cycle
Favorable attitude toward aggressive store	ns	-	.0192	

EXHIBIT 5--Continued

Predictor	Discriminant Function		A.I.D. Split Reduction in Unexplained Variance	Comments
	Significance	β^2		
Live in Neighborhood 8	.05	.0279	.0508	
Live in Neighborhood 7	.10	(-).0135	.0230	
Weekly food expenditures	.01	.1239	.0244 .0333	
Number of stores visited per week	.01	(-).0835	-	
Number of different stores visited in 15 weeks	.05	.0339	.0633	Significant correlation with stores per week
Number of shopping trips per week	-	-	.0491	
Mean interval between shopping trips	.20	(-).0204	-	
Mean expenditures per trip	.05	(-).0061	-	Significant correlation with expenditures and interval between trips