

Intertechnique Cross-Validation in Cluster Analysis

A. Marvin Roscoe, Jr., Jagdish N. Sheth, and Welling Howell *

Clustering methods are often used in marketing research to define homogeneous market segments. It should be determined in these studies that the derived clusters represent actual clusters. However, replication or external validation is not always practical. An alternative procedure, cross-validation using intertechnique comparisons, is described in a study of geographical market heterogeneity for the telephone industry.

Cross-validation among techniques seems essential in cluster analysis because most clustering methods tend to be heuristic algorithms instead of analytically optimal solutions.¹ As heuristic algorithms, they have no sampling theory for statistical inferences about the size and the number of clusters. Also, there are no external validation procedures to ensure that the clusters derived from a specific cluster analysis are, in reality, the true invariant clusters. The potential statistical problem of obtaining artifacts as clusters is further compounded in some procedures which require *a priori* assumptions about the size and the number of clusters. Although a number of clustering methods perform statistical tests such as the F ratio or Wilks' Lambda based on analysis of variance principles to guard against obtaining random solutions, no procedure exists which will increase the assurance that a nonrandom cluster solution is in fact the true cluster solution.

Because clustering methods are used in marketing research to identify homogeneous market segments for selective marketing efforts, it is critical that the clusters derived from a heuristic algorithm are the true clusters. One procedure to ensure cluster invariance is replication which, however, is not always practical. Another procedure is the common practice in psychometrics of cross-validating the results by external validation. Surprisingly, there are very few studies in which cross-validation has been utilized to insure that the derived clusters are indeed invariant.

* A. Marvin Roscoe, Jr. and Welling Howell are Marketing Supervisors in the Market Research Section of the Marketing Department of the AT&T Company. Jagdish N. Sheth is I.B.A. Distinguished Professor and Research Professor at the University of Illinois, Urbana-Champaign.

¹ See Joyce and Channon (6) and Frank and Green (2) for a review of the numerous clustering methods available today.

Although several studies have pointed out the dramatic changes in the cluster structures as a function of data input [4, 8], there seems to be only one published study to our knowledge which has examined the question of intertechnique validation of clusters [3].

The objective of this paper is to describe a cross-validation procedure which utilizes intertechnique comparisons of the clustering results. Although the actual study entailed applications of five different clustering techniques, our discussion is limited to two techniques in this paper due to space limitations. A brief description of the large scale research project is provided in which the clustering results were essential to formulating an experimental design for a field experiment.

DESCRIPTION OF THE STUDY

The major research study consisted of a three factorial-64 cell experimentation on survey research methods. The three factors were: (1) two different lengths of the questionnaire; (2) four different follow-up procedures; and, (3) the market heterogeneity of geographical areas of the United States with respect to consumer telephone behavior and socioeconomic-demographic characteristics [9]. The levels of the first two factors were predetermined based on theory, prior research and practical implications for the ongoing research on a longitudinal national panel of telephone customers. For the third factor, it was necessary to determine the heterogeneity of the markets by empirical research which utilized clustering methods.

To define the market heterogeneity, profile data on 30,000 residential telephone customers were used for clustering. These customers are part of a longitudinal

consumer panel called the Marketing Research Information System which is maintained for the Bell System by AT&T. The panel members are selected based on a multistaged stratified sample in which the first stage of the sampling procedure consists of 100 Revenue Accounting Offices (RAOs) representing the entire Bell System. The profile consists of essentially three types of information about each panel member: (a) his socioeconomic-demographic status and housing characteristics determined by a survey conducted in early 1970 and matched with the 1970 Census, (b) his monthly telephone behavior broken down into several categories as determined by the industry practice, and (c) an inventory of his telephone equipment including number and types of telephones, and additional services.

Since it was required to investigate empirically the geographical heterogeneity of the markets, an average profile of the residential telephone customers was determined for each of the 86 RAOs for which detailed and complete information was available.

Table 1
LIST OF VARIABLES

<p>Housing</p> <ol style="list-style-type: none"> 1. Own-rent home 2. Type of residence 3. Number of rooms <p>Mobility</p> <ol style="list-style-type: none"> 4. Length of residence <p>Head of Household</p> <ol style="list-style-type: none"> 5. Sex 6. Age 7. Education 8. Occupation 	<p>Family</p> <ol style="list-style-type: none"> 9. Income 10. Number in family 11. Average Age 12. Life cycle 13. SES status <p>Telephone Service and Equipment</p> <ol style="list-style-type: none"> 14. Class of service 15. Grade of service 16. Number of telephones 17. Number of vertical services
<p>Billing Items 12 months</p>	
<p>18-29 Local service</p> <p>30-41 Local message</p> <p>42-53 Intrastate long distance</p> <p>54-65 Interstate long distance</p>	

A total of 65 customer descriptors were used to represent the total profile of customers. A list of the variables is shown in Table 1. A factor analysis (principal components) solution with orthogonal Varimax rotation was performed on the data for the following reasons: (a) to reduce the multicollinearity among variables so that the profile consisted of orthogonal factor scores which are geometrically essential to calculate Euclidian distances, (b) to equalize the relative weights of each of the underlying dimensions which could otherwise be easily changed by arbitrary dropping or adding of profile variables, and (c) to standardize the diverse scales of measurement common across the socio-economic, demographic and telephone information [7]. Ten significant factors were extracted from the analysis which summarized 92 percent of the total variance. A brief description of the factors is provided in Table 2.

The number of significant factors was determined using several criteria, both statistical and judgmental, following the recommendations of Rummel [10]. In addition, the stability of the factor structure also was determined by comparing the

Table 2
FACTOR DIMENSION LABELS

FACTOR DIMENSION LABELS	
1. Local service billing	6. Life cycle
2. Local message billing	7. Service and equipment
3. Intrastate long distance	8. Interstate long distance 1 *
4. Family - housing	9. Interstate long distance 2 *
5. Interstate long distance	10. Socioeconomic characteristics

* The two factors for interstate long distance represent different seasonal patterns of calling across geographical areas.

results with other data analyses to ensure the invariance of the fundamental dimensionality and structure of the profile data.

The standardized rotated factor scores for each RAO were then utilized to compute Euclidian distances between all combinations of RAOs. The resultant 86 X 86 distance matrix became the input to the clustering procedures.

Due to the following distinct advantages, Johnson's Hierarchical Clustering method [5] was chosen as the primary clustering technique for determining the market heterogeneity. First, it is strictly empirical; second, no prior assumptions are required on the part of the researcher; and third, a hierarchical display is provided of the clusters being formed based on a function minimizing the pairwise distances among entities. While the size of the distance matrix is a limitation of the technique, it was not a problem in our case because of the relatively small number of RAOs to be clustered. Due to the structure of the distance matrix and the presumption of the "ultrametric inequality", (5, p. 248-9) the diameter method was chosen instead of the connectedness method in the Be-HICLUST solutions. The results are diagrammed in Figure 1.

While the hierarchical clusters from HICLUST were meaningful and had strong face validity, it was necessary to cross-validate the results by at least one other technique which was essentially similar in its input requirements, analytic strategies and the output format. For this we chose the cluster analysis program developed as part of the BMDP Series which is also a hierarchical clustering routine based on sum of squares distances and the amalgamation principle [1]. In short, BMDP2M amalgamates entities based on the criterion of the smallest distance. Once a cluster is formed, consisting of at least two entities, it calculates the average profile of the cluster and treats it as if it were a new entity which is then clustered with other entities or clusters based on the principle of smallest distances. The process continues until all entities and clusters are hierarchically linked at different levels of distances. The results of the BMDP2M analysis are diagrammed in Figure 2.

As can be seen, the two hierarchical clusters are similar in their structure and hierarchy suggesting that there is a good cross-validation between the two analyses. In order to quantitatively assess the degree of congruence between the two hierarchical clusters, two distinct statistical procedures were utilized. The first procedure consisted of calculating the correlation

Table 3 - Continued

LISTING OF CLUSTERS

Western		Metropolitan		Unique	
BE-HICLUST	BMDP2M	BE-HICLUST	BMDP2M	BE-HICLUST	BMDP2M
M1 East Houston	Orange	M1 North Central	Brooklyn	Nevada	Eastern Pa
San Antonio	Fort Worth	Brooklyn	Queens	Eastern Pa	West Houston
Fort Worth	San Antonio			West Houston	Nevada
Dallas	East Houston			Arlington Heights	Lubbock
Tulsa	Austin			East Manhattan	Philadelphia
Austin	Dallas			South Boston	Metropolitan
	Tulsa	M2 Westchester	Westchester	Marquette Park	St. Louis
		Harvey	Harvey	Lubbock	East Manhattan
M2 Sacramento	Sacramento	Joliet	Joliet	Metropolitan	South Manhattan
San Diego	San Diego			St. Louis	Arlington Heights
San Jose	Los Angeles			Bronx	Marquette Park
Orange	Vermont	M3 Indianapolis	Indianapolis	South Manhattan	South Boston
	Charleston	Central Pa	Central Pa	Nassau Suffolk	Bronx
	North Central	Philadelphia		Phoenix	Phoenix
					Nassau Suffolk
M3 East Bay	Foothill				
Van Nuys	East Bay				
Foothill	Van Nuys				
San Francisco	San Jose				
	San Francisco				

Table 4
CROSS-TABULATION OF CLUSTERS

BE-HICLUST	EASTERN			SOUTHERN			CENTRAL				WESTERN			METROPOLITAN			UNIQUE RAGS	
	1	2	3	1	2	3	1	2	3	4	1	2	3	1	2	3		
EASTERN	1	6																
	2	1	3															
	3		2	3								2						
												1						
SOUTHERN	1			4	1	2												
	2				2													
	3					0												
CENTRAL	1						3	2										
	2							4										
	3	1				1		5	3									
	4								3									
WESTERN	1											6						
	2											1	2	1				
	3												4					
METROPOLITAN	1													1				
	2														2			
	3														3			
UNIQUE RAGS																2		1
																		14

are differences in the example worth noting. The BE-HICLUST algorithm appears to provide a more logical structure to the clusters which are grouped by region as indicated in Figure 2. In addition, the BE-HICLUST method seems to work better where large distances are involved, associating 8 of the 14 unique entities with meaningful clusters. Such differences reinforce the need to use several techniques and to understand the advantages of each especially where the researcher's judgement plays

such an important role.

SUMMARY AND CONCLUSIONS

We have pointed out the need for intertechnique cross-validation in cluster analysis due to the heuristic nature of most clustering procedures and the subjective judgements required to interpret the results. In this paper, we have also presented a

BE-HICLUST

Table 5
CROSS-TABULATION OF REGIONAL,
METROPOLITAN AND UNIQUE CLUSTERS
BMDP2M

	EASTERN	SOUTHERN	CENTRAL	WESTERN	METROPOLITAN			UNIQUE RACS
					1	2	3	
EASTERN	11	3	1	3				
SOUTHERN		9						
CENTRAL	1	1	20					
WESTERN				14				
METROPOLITAN 1				1	2			
METROPOLITAN 2						3		
METROPOLITAN 3							2	1
UNIQUE RACS								14

concrete application of two statistical procedures which enable the researcher to quantitatively measure the congruence of structure and content of clusters across techniques. The first consists of a correlation coefficient index calculated on the distributions of distances at which sequential linkages are made among entities or clusters, or both. The second consists of a cross-tabulation of specific clusters derived across two different solutions.

In this paper, the intertechnique cross-validation procedures have been applied with respect to two hierarchical clustering procedures in which the problem was the determination of geographical heterogeneity of markets for the telephone industry. This application considered the general housing and population characteristics along with a complete profile of telephone behavior. Other uses of the intertechnique cross-validation procedure also have been made by the authors for a variety of telephone behavior and markets.

REFERENCES

1 Dixon, W. J. "BMD P Series Documentation," Health Sciences Computing Facility. Los Angeles: University of California, 1971.

2 Frank, Ronald B. and Paul E. Green. "Numerical Taxonomy in Marketing Analysis: A Review Article," *Journal of Marketing Research*, 5 (February 1968), 83-98.

3 Golob, Thomas F., Eugene T. Canty, and Richard L. Gustafson. "Classification of Metropolitan Areas for the Study of New Systems of Arterial Transportation." Paper presented at the 1972 Annual Meeting of the Transportation Research Forum, Denver, Colorado, November 8-10, 1972.

4 Green, Paul E., Ronald E. Frank, and Patrick J. Robinson. "Cluster Analysis in Test Market Selection," *Management Science*, 13 (April 1967), 387-400.

5 Johnson, Stephen C. "Hierarchical Clustering Schemes," *Psychometrika*, 32 (September 1967), 241-54.

6 Joyce, Timothy and C. Channon. "Classifying Market Segment Respondents," *Applied Statistics*, 15 (November 1966), 191-215.

7 Morrison, Donald G. "Measurement Problems in Cluster Analysis" *Management Science*, 13 (August 1967), B775-80.

8 Neidell, Lester. "Comments on Typology and Cluster Analysis," Paper presented at the AMA Workshop on Multivariate Methods in Marketing, Chicago, Illinois, January 1970.

9 Roscoe, A. Marvin, Dorothy Lang, and Jagdish N. Sheth. "Experimental Effects of Follow-up Methods, Questionnaire Length, and Market Heterogeneity in Mail Surveys," Manuscript submitted for publication, 1974.

10 Rummel, R. J. *Applied Factor Analysis*. Evanston: Northwestern University Press, 1970, Chapter 15.