Market Figmentation:
Clustering on Factor Scores versus Individual Variables

John Fiedler
Principal
POPULUS, Inc.

John J. McDonald
Senior Manager, Research & Planning
Donnelley Directory
A Company of the Dun & Bradstreet Corporation

Paper Presented to the AMA Advanced Research Techniques Forum, 1993

Background

In the last twenty-five years, hundreds and perhaps thousands of market segmentation studies have been conducted. Typically, a collection of variables is subjected to cluster analysis, and the groups of respondents thus identified are regarded as prototypes of "market segments."

Aaldenderfer and Blashfield (1984), in a particularly useful Sage monograph on cluster analysis, comment:

> *The strategy of cluster analysis is structure-seeking although its operation is structure-imposing. That is, clustering methods are used to discover structure in data that is not readily apparent... [but a] clustering method will always place objects into groups, and these groups may be radically different in composition when differing cluster methods are used.*

It is widely recognized that with cluster analyses you can get different answers if you use different techniques with the same data, the same technique with different samples of data, or even the same technique and the same data, but after a trivial operation such as reversing the order of two respondents in the same data file. Although it may be hoped that users of cluster analysis are becoming more sophisticated, it seems a fair bet that many market segmentations of the past and present may be more akin to figments of the imagination than segments of the market.

Having explained our title, we shall describe what we have tried to do about this problem.

There has been much discussion regarding choice of variables to be used in clustering. Many practitioners simply take an entire battery of individual items and "throw them all into the pot." Others, perhaps more thoughtfully, first transform the data to principal component or factor scores. Still others use subsets of individual items, one subset chosen to represent each factor or principal component.

Our purpose is to examine the relative effectiveness of those three approaches to the selection of variables for clustering. We have measured effectiveness in two ways:

- In terms of *reproducibility* of the resulting cluster solutions; and
- In terms of discrimination between clusters on variables considered to be critical for the product category but not used in the clustering.

Our study involved the reanalysis of four data sets from commercial research projects originally designed to produce data for cluster analysis.

Rationales for Choosing Variables

Alderderfer and Blashfield comment:

> *The temptation to succumb to a naive empiricism in the use of cluster analysis is very strong, since the technique is ostensibly designed to produce "objective" groupings of entities.*

However, the "throw them all in the pot" approach has serious shortcomings. Milligan (1980) found that inclusion of even one irrelevant variable could seriously reduce the extent of cluster recovery. When using this approach, the analyst also relinquishes control over the weighting of his variables. If some factor is represented by several correlated variables, the effect of including them all is similar to giving that factor increased weight.

In theory, there would seem to be considerable benefit in first doing a principal component analysis, and then clustering on the resulting component scores. This avoids problems of unequal weighting. Also, since factor scores are weighted combinations of correlated variables, they are likely to be more reliable, and generally of higher quality than the individual variables. However, this approach also presents problems.

Aldenderfer and Blashfield comment, "Factor analysis tends to blur the relationship between clusters. . . ."

Johnson (1988) comments, "The central limit theorem assures that when variables are grouped into factors a lot of 'smoothing' will occur. Cluster analysis can take advantage of the 'lumpiness' of data, and will be impeded by any smoothing. . . ."

And Millgan (1987) states that if clusters exist in the original variable space, then factor analysis can distort or hide the true structure, as shown in a study by Sneath (1980).

The third strategy, and one that we have favored in the past, is that of first conducting a factor analysis, and then representing each resulting factor with a subset of, say, two or three variables. This approach given the analyst control over weighting and avoids the smoothing that might be encountered with factor scores.

The Data Sets

Each of the four studies contained variables used in the clustering, as well as others thought to be significant of the product category of concern, but not used in the clustering.

Table 1

CHARACTERISTICS OF THE DATA SETS

Study	Method	Sample	Variables	PC Scores	Variance Accounted for	Item Subsets	Criterion Variable
A	Telephone	1620	31	7	60.6%	14	Interval
B	Personal / Self-Administered	1000	42	13	54.7	26	Interval
C	Telephone	1473	16	4	42.1	8	Interval
D	Telephone	1060	18	4	59.0	8	Categorical

Our results are based on 648 cluster analysis solutions:

- For each data set, we used three difference methods of choosing variables (4 x 3 = 12)
- Sawtooth Software's CCA$^{\text{TM}}$ System, a κ-means clustering technique, was used to develop three-through-eight-cluster solutions for each combination of data set and preparation method (12 x 6 = 72)
- Each solution was replicated nine times from different starting points (72 x 9 = 648)

Data Preparation

Date set B was obtained by computer-assisted personal interviewing. The other three data were obtained in telephone interviews. There were very few instances of missing data; each was recoded with the modal value for that variable.

Since the data in each set consisted of rating scale variable using common scales, the "All Variables" analyses were none using the data "as is," with no standardization.

The "Principal Component Scores" analyses were down after converting the data to orthogonally rotated component. We retained those principal components with eigenvalues great than unity, performed Varimax rotations, and computed component scores for the rotated components. These scores were automatically standardized to unit variance.

The "Item Subsets" analyses were done by choosing the two items most highly loaded on each rotated component. Those data were not standardized.

The Clustering Method

In an excellent review of clustering methods, Milligan and Cooper (1987) conclude:

> *In summary, the convergent* K*-means method tended to give the best recovery of cluster structure.*

In a more recent comparison of eighteen clustering methods used in marketing research, Neal (1989) concluded that "optimizing" methods are likely to outperform hierarchical methods in the marketing research environment. Among optimizing methods, he included K-means methods, the Howard Harris method, and Ward's method.

Our clustering was done using Sawtooth Software's CCA System for Convergent Cluster Analysis. CCA uses a K-means method, and has two additional capabilities that facilitated our analysis.

The first of these capabilities involved the starting points for each solution. If k clusters are to be determined, K-means methods begin by choosing a set of k starting points, each usually consisting of data for one respondent. Each iteration consists of two steps:

- ➢ Every respondent is classified into the group with which he is "most similar."
- ➢ After all respondents are classified, group means are computed.

These two steps are repeatedly iteratively until the process converges and no respondents are reclassified.

It is widely recognized that the quality of K-means solutions is dependent on the quality of the starting points. Starting points that are close to the centers of final clusters are much more successful than randomly chosen ones. CCA provides three methods for choosing starting points:

- *Distance-based* starting points, relatively distant from one another, but not on the "outer fringes" of the configuration of points;
- *Hierarchical-based* points, chosen by a hierarchical clustering of a random sample of 50 respondents; and
- *Density-based* points, chosen to be near the centers of relatively dense portions of the configurations of points.

CCA replicates clusterings automatically using different starting solutions. We used nine replications of each solution, making use of all starting points.

A second feature of CCA that was useful to us is its automatic measurement of reproducibility. Every replicate is compared with every other, to see if respondents who are clustered together in one replicate are also clustered in the other. A pairwise "percent reproducibility" is reported, equal to the percentage of respondents for whom this is true. These pairwise reproducibilities are averaged across replicates to obtain an average percentage for each replicate. The replicate with highest average percent reproducibility is automatically chosen.

Results

We do not propose to identify the "true" number of clusters for any of these data sets, for two reasons.

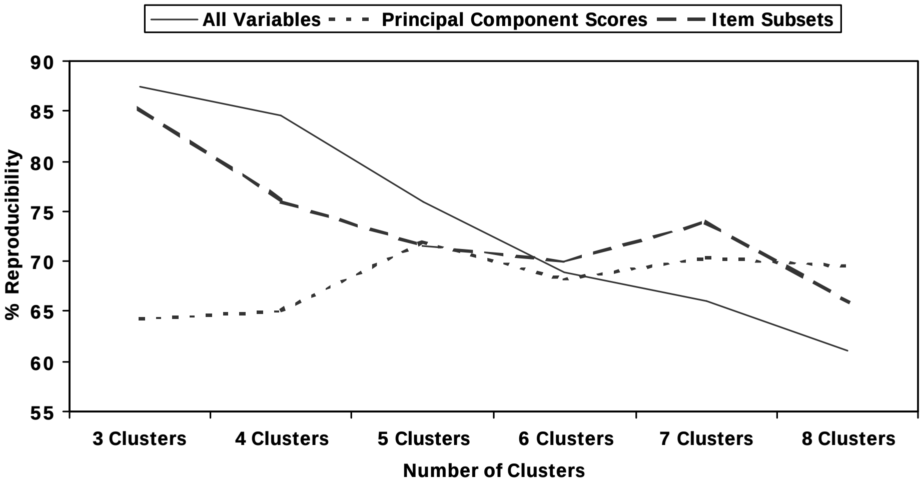
- In this paper we cannot go into the marketing background for each data set. We have tried to remain entirely neutral in this respect, even to the extent of coding the variables as "A, B, C. . ." rather than with descriptive labels.
- We regard the very idea of the "true number of clusters" as a fiction. There are doubtless alternative solutions that would be nearly as useful. Rather we want to examine how one *might* choose the number of clusters, and to see if one would be led to the same decisions by different methods of data preparation. Also, we hope to make general conclusions about the overall quality of *all* possible solutions for each method of data preparation.

For the All Variables method, three solutions contained degeneracies of clusters with only single respondents; for the Item Subsets method there was one such cluster. In all other cases the minimum number of respondents per cluster was greater than 80. This suggests that the Principal Component Scores method is least vulnerable to degeneracies.

Figure 1 show a typical plot of CCA's reproducibility measure as the number of clusters increases from three to eight. There is no information in this figure about how many clusters may be the best choice for any data set, because it shows the average for all four data sets. For All Variables and Item Subsets the curves tend to decrease. For Principal Component Scores the curve is much flatter and less regular.

FIGURE 1

REPRODUCIBILITY MEASURES AVERAGING ACROSS DATA SETS



Reproducibility measures for all solutions are shown in Table 2. Every method produced solutions with reproducibility in the 70s or better except for Principal Component Scores, whose best reproducibility for data set B is 49%.

Comparisons among different columns are not strictly appropriate because reproducibility tends to be lower when there are more variables. Comparing columns gives an unfair advantage to Principal Component Scores, and to a lesser extent to Item Subsets. However, we have highlighted the largest number of each row, and the count of “bests” is nearly the same for each method. The near equality of these three counts is evidence that the Principal Component Scores method produced the least reproducible results. Also, the All Variables method produced reproducibilities in the high nineties for two data sets; by comparison, the Principal Component Scores method produced no values in the nineties.

TABLE 2

REPRODUCIBILITY MEASURES: ALL SOLUTIONS

Data Set	All Variables	Principal Component Scores	Item Subsets
Study A:			
3 Clusters	**99.47%**	59.85%	83.28%
4 Clusters	**96.33**	57.99	83.88
5 Clusters	**80.12**	77.37	74.38
6 Clusters	74.26	72.54	**75.37**
7 Clusters	70.39	77.86	**81.91**
8 Clusters	61.87	**84.12**	66.43
Study B:			
3 Clusters	72.74%	48.68%	**85.03%**
4 Clusters	**76.33**	44.08	71.78
5 Clusters	**76.25**	43.90	65.14
6 Clusters	**64.23**	41.30	59.68
7 Clusters	59.51	44.10	**59.93**
8 Clusters	55.51	43.66	**58.23**
Study C:			
3 Clusters	77.75%	80.48%	**99.23%**
4 Clusters	**86.79**	74.88	76.65
5 Clusters	73.51	**80.28**	69.03
6 Clusters	63.66	70.41	**72.87**
7 Clusters	60.57	73.34	**77.18**
8 Clusters	56.61	71.67	**72.58**
Study D:			
3 Clusters	**98.71%**	66.99%	74.61%
4 Clusters	77.23	**81.80**	76.30
5 Clusters	78.43	**86.92**	78.76
6 Clusters	72.76	**88.75**	74.23
7 Clusters	73.09	**85.29**	77.94
8 Clusters	67.29	**75.74**	67.85
Number of “Bests”	8	7	9

If we were to select the solutions with the highest reproducibility, we would choose different numbers of clusters for different methods. The solutions chosen by this criterion are summarized in Table 3. By this criterion, all three methods disagree for data sets A and D; for the All Variables method suggests three clusters for both data sets, and the other methods we would choose more clusters. For data sets B and C, the All Variables method suggest four cluster, whereas the other methods agree on three.

TABLE 3

NUMBER OF CLUSTERS IN SOLUTION WITH GREATEST REPRODUCIBILITY

Data Set	All Variables	Principal Component Scores	Item Subsets
Study A	3	8	4
Study B	4	3	3
Study C	4	3	3
Study D	3	6	5
Average	3.50	5.00	3.75

In looking for an "elbow" in a curve, it is useful to examine differences between each point and the previous one. Table 4 presents those differences (the first solution is arbitrarily given a difference of zero) with the largest difference highlighted in each section.

TABLE 4

REPRODUCIBILITY MEASURES: DIFFERENCES BETWEEN SOLUTIONS

Data Set	All Variables	Principal Component Scores	Item Subsets
Study A:			
3 Clusters	**0.00%**	0.00%	0.00%
4 Clusters	-3.14	-1.86	0.60
5 Clusters	-16.21	**19.38**	-9.50
6 Clusters	-5.86	-4.83	0.99
7 Clusters	-3.87	5.32	**6.54**
8 Clusters	-8.52	6.26	-15.48
Study B:			
3 Clusters	0.00%	0.00%	0.00%
4 Clusters	**3.59**	-4.60	-13.25
5 Clusters	-0.08	-0.18	-6.64
6 Clusters	-12.02	-2.60	-5.46
7 Clusters	-4.72	**2.80**	**0.25**
8 Clusters	-4.00	-0.44	-1.70
Study C:			
3 Clusters	0.00%	0.00%	0.00%
4 Clusters	**9.04**	-5.60	-22.58
5 Clusters	-13.28	**5.40**	-7.62
6 Clusters	-9.85	-9.87	3.84
7 Clusters	-3.09	5.93	**4.31**
8 Clusters	-3.96	-4.67	-4.60
Study D:			
3 Clusters	0.00%	0.00%	0.00%
4 Clusters	-21.48	**14.81**	1.69
5 Clusters	**1.20**	**5.12**	2.46
6 Clusters	-5.67	1.83	-4.53
7 Clusters	0.33	-3.46	**3.71**
8 Clusters	-5.80	-10.05	-10.09

Table 5 summarizes the solutions that would be chosen by this criterion. Naturally, this criterion tends to choose solutions with larger numbers of clusters. All three methods again disagree for data sets A and D. For data sets B and C the All Variables method suggests four clusters, whereas the other two methods agree, this time on seven clusters.

TABLE 5

NUMBER OF CLUSTERS IN SOLUTION WITH GREATEST FIRST DIFFERENCE

Data Set	All Variables	Principal Component Scores	Item Subsets
Study A	3	5	7
Study B	4	7	7
Study C	2	7	7
Study D	5	4	7
Average	3.50	5.75	7.00

We would be likely to choose difference numbers of clusters for each method of data preparation if we were to rely entirely on reproducibility statistics, and that choice would be still difference whether we were looking for maximum reproducibility or elbows in the curve.

We turn now to tabulations of cluster members with other variables that were regarded as being of critical importance in the studies from which these data sets were drawn. We have chosen a single (but different) criterion variable from each study. In each case it is so fundamental to the product category that, unless the clusters were to differ meaningfully on this variable, the segmentation would not have been accepted by management. For data sets A through C the variable was intervally scaled, so we have used a one-way F statistic. For data set D the variable was categorical so we have used Chi Square statistics.

Table 6 presents these statistics. Here it is appropriate to compare statistics in different columns, and we have highlighted the best value in each row. Note that the All Variables method achieves seventeen “bests,” whereas each other method achieves only three or four. This, in itself, appears to be strong evidence in favor of the All Variables method, more than compensating for previously noted tendency toward degeneracy.

TABLE 6

MEASURES OF DISCRIMINATION*

Data Set	All Variables	Principal Component Scores	Item Subsets
Study A:			
3 Clusters	**1035**	206	629
4 Clusters	**1013**	492	581
5 Clusters	**852**	379	381
6 Clusters	**722**	236	299
7 Clusters	**662**	104	252
8 Clusters	**524**	173	261
Study B:			
3 Clusters	25.4	19.0	**32.2**
4 Clusters	**25.2**	3.1	20.2
5 Clusters	20.2	9.7	**20.8**
6 Clusters	**17.1**	4.3	13.8
7 Clusters	**14.6**	9.7	14.1
8 Clusters	**18.5**	16.0	12.9
Study C:			
3 Clusters	34.1	**61.1**	22.1
4 Clusters	**46.2**	41.2	46.0
5 Clusters	22.2	28.1	**30.0**
6 Clusters	27.0	**31.3**	26.3
7 Clusters	19.7	**25.1**	22.7
8 Clusters	18.2	**22.1**	18.1
Study D:			
3 Clusters	**510**	347	416
4 Clusters	**527**	359	361
5 Clusters	**559**	365	418
6 Clusters	**555**	362	395
7 Clusters	**592**	393	423
8 Clusters	**580**	403	416

*Studies A, B, and C: F ratio; Study D: Chi Square

Informal comparison of discrimination with reproducibility shows that there is not a strong relationship between the measures in all cases. We have summarized that relationship by computing the correlation between reproducibility and discrimination measures in each of the twelve cells. As the number of clusters increases, both reproducibility statistics and F ratios are expected to decrease, leading to a positive bias in the correlations; and as the number of clusters increases, the expected Chi Square value tends to decrease, leading to a negative bias in the correlations. However, despite these biases, we can see whether reproducibility and discrimination are more highly related with one data preparation method than another.

TABLE 7

CORRELATIONS BETWEEN REPRODUCIBILITY AND DISCRIMINATION

Data Set	All Variables	Principal Component Scores	Item Subsets
Study A	**.989**	-.519	.648
Study B	.341	.722	**.973**
Study C	**.894**	.502	.188
Study D	**.802**	.147	.067

Table 7 provides further evidence favorable to the All Variables method. Although we do not report means, the values in the first column obviously tend to be higher. Three of the four largest values in the table are found in column one, indicating that for the All Variables method, there is the closest relationship between reproducibility and discrimination.

The Component Score column has the least favorable correlations; not only does it include a large negative value, but it is dominated by some other method for each of the four data sets.

Finally, although we have seen evidence that the results differ from method to method, we have not yet confronted the question of how different the methods really are in terms of classifying respondents. We have tried to answer that question by counting the percentage of respondents who are classified identically by each pair of methods, averaged over all solution and data sets. These percentages are shown in Table 8.

TABLE 8

AVERAGE PAIRWISE SIMILARITY OF SOLUTIONS

Data Set	All Variables	Principal Component Scores	Item Subsets
All Variables		39.70%	43.59%
Component Scores			42.58
Item Subsets			

These percentages are computed in the same way as the reproducibility statistics, and may be evaluated using the same frame of reference. For the three-cluster solution the change level is 1/3 = 33.3%, whereas for the eight-cluster solution the chance level is 1/8 = 12.5%. The average expected similarity over all clusterings is approximately 20%.

While better than chance, the similarity between methods is disappointingly low. As can be seen, All Variables and Item Subsets are most similar in the way they classify respondents. All three sets of solutions are more similar than one would expect due to chance, but they are far from identical, all three pairwise percentages being substantially lower than the reproducibility values.

Discussion

When we undertook this study we anticipated quite different results. We had expected the various data preparation techniques to show considerable agreement with one another about how many clusters to use, both internally in terms of reproducibility, and externally in terms of discrimination with respect to important criterion variables.

However, we also anticipated differences: we expected the Item Subsets approach to be most effective and the Principal Components and All Variables approaches to be far less effective. Thus, our preconceptions could scarcely have been more at odds with the results.

We had also hoped to lay this issue to rest. Instead, we have results that do not provide much comfort for the cluster analyst.

Our main findings are as follows:

- It makes a difference how you prepare the data. The three data preparation methods we examined lead to quite different solutions.
- Each method produces what appear to be stable and reproducible solutions for most data sets; in most cases the different methods do so with different numbers of clusters.
- Using the reproducibility criterion, the All Variables method tends to favor fewer clusters than the other methods.

- The All Variables method appears to be superior in terms of reproducibility, and the Component Score method appears to produce the least reproducible solutions.
- The All Variables method is clearly superior in terms of providing sharp among-cluster differences on important external variables.
- The All Variables method also shows the strongest relationship between internal reproducibility and external discrimination. Although generalization is precarious, it appears that the odds are maximized that clusters will differ on important external variables if the All Variables method is used, and a highly reproducible solution is chosen.

The All Variables method appears to be the winner. In pondering this, it has occurred to us that we believe our variable sets were well constructed, and may have been inherently weighted to reflect the concerns of management. To whatever extent this is true, this result would not generalize to variable sets that were thrown together with less care.

Although we prefer to end with answers than with cautions, it remains clear to us that cluster analysis is more art than science. It is a good idea to try things several ways. Although we haven't dealt with "meaningfulness" of clusters, that is surely the most critical test of any cluster analysis. If anything, we have demonstrated that the naive approach of throwing variables into the pot, turning the crank, and accepting whatever comes out, it likely to result in Market Figmentation.

References

Aldenderfer, Mark S. and Roger K. Blashfield (1984), *Cluster Analysis*, Beverly Hills, CA: Sage Publications

Johnson, Richard M. (1980), "Convergent Cluster Analysis," Working Paper, Sawtooth Software.

Milligan, Glenn W. (1980), "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms," *Psychometrika*, 45, 325-342.

Sawtooth Software (1988), *CCA System for Convergent Cluster Analysis*, Ketchum, ID: Sawtooth Software.

Sneath, P.H.A. (1980), "The Risk of Not Recognizing from Ordinations that Clusters Are Distinct," *Classification Society Bulletin*, 4, 22-43.