A Comparison of Correspondence Analysis and Discriminant Analysis-Based Maps

John A. Fiedler

POPULUS, Inc.

AMA Advanced Research Techniques Forum

1996

Abstract

Perceptual mapping has become a widely-used technique in marketing research, and is often the method of choice for graphic display of market segmentation information. Two methods for constructing maps from product-by-attribute data are widely used: Discriminant Analysis (DA) and Correspondence Analysis (CA). This study compares the abilities of DA and CA to recover the true structure of data, using an easy-to-understand data set featuring 20 U.S. cities as "products," and 16 compass directions as "attributes." A data set was constructed using Monte Carlo methods, similar to what respondents might produce by fallibly rating each city of each direction. The simulated individual respondent data were subjected to DA and CA to attempt to recover the relations among cities, and relations between cities and directions. Although both techniques reproduced recognizable maps under all circumstances, DA always reproduced the true configuration much more accurately. We conclude that CA is indeed easier and less demanding, but that DA does a better job when adequate data are available.

Background

When first introduced to the field in the 1960s, perceptual mapping opened an exciting chapter in marketing research. Perceptual mapping was exciting methodologically because the earliest methods showed how information could be elevated from one level of measurement to another. The earliest methods dealt with data having only rank order properties, but provided results interpretable as distances, scaled at the ratio level. And perceptual mapping was exciting from a substantive point of view because it displayed relationships among products and attributes spatially, contributing dramatically to the generation of insights among manages and researchers alike.

Perceptual mapping market research has used a variety of methods. The earliest methods were based on perceived distances among pairs of products. Methods using distances were provided by several authors, including Kruskal's Nonmetric Multidimensional Scaling (1964), and Young's TORSCA (1968).

In 1970 Johnson proposed using multiple discriminant analysis for mapping, employing data consisting of ratings of products on attributes, rather than distances.

Discriminant analysis (DA) provided several benefits over methods based on perceived distances:

Tests of significance were available for dissimilarities among products.

Distances estimated among any two products did not depend upon other products included in the analysis.

The technique was robust and not vulnerable to local optima.

Also in the 1970s, French researchers developed a technique for graphical display of information in a table of frequencies which has come to be called Correspondence Analysis (CA). In 1984 Greenacre, as well as Lebart, Morineau, and Warwick, provided details in English-Language books, and in 1986 Haffman and Franke described the technique in JMR. CA can be viewed as a principal components analysis of a two-way table of frequencies, after double-centering and scaling by the reciprocal square roots of row and column sums. After this preprocessing, the data to be analyzed as square roots, f cell-by-cell contributions to the table's chi square.

Although we have not conducted a formal survey of methods currently used for perceptual mapping, usage of distance-based methods seems to have decline, despite their elegance. However, DA and CA are in wide use today, and each has advantages.

CA is much more convenient that DA. CA is usually done at the aggregate level, while DA requires data from individual respondents, and therefore often presents problems of missing data.

However, DA also has potential advantages:

1. Because it deals with individual rather than aggregate data, it may make fuller use of the data and more accurately reproduce structure inherent in the data.

DA maps may be less affected by the inclusion of redundant attributes than CA maps, since dimensions of DA maps measure ratios of between-product to within-product variation on linearly independent combinations of attributes. With CA maps, there is no consideration of "between" vs. "within" product variation. Redundant attributes would be expected to "attract" a CA dimension in the direction of those attributes, and to increase its size.

Also, distances between products in DA maps may be less affected by the inclusion of exclusion of other products than in CA maps. In DA maps, distances among products depend on only those products' means (as well as a common "error" covariance matrix). In CA maps, distances between a pair of products depend on the total frequencies for each attribute for all products. Adding a product with high values on some attributes and low values on others could have a large effect on the distances among other products.

2. In theory, DA maps should offer clearer interpretations of relationship between products and attributes than CA maps. Both methods provide information about relationships of products with one another, and of relationships of attributes with one another. Discriminant-based maps also permit interpretation of relationships between products and attributes, using projections of products of attribute vectors. However,

in the usual CA display, in which product and attribute points are each displayed with similar scalings, distances between product and attribute points are not meaningful and should not be interpreted, according to Greenacre (1989):

"The temptation is to interpret between-set (row-to-column) distances in the symmetric plot, but no such interpretation is in fact intended or valid."

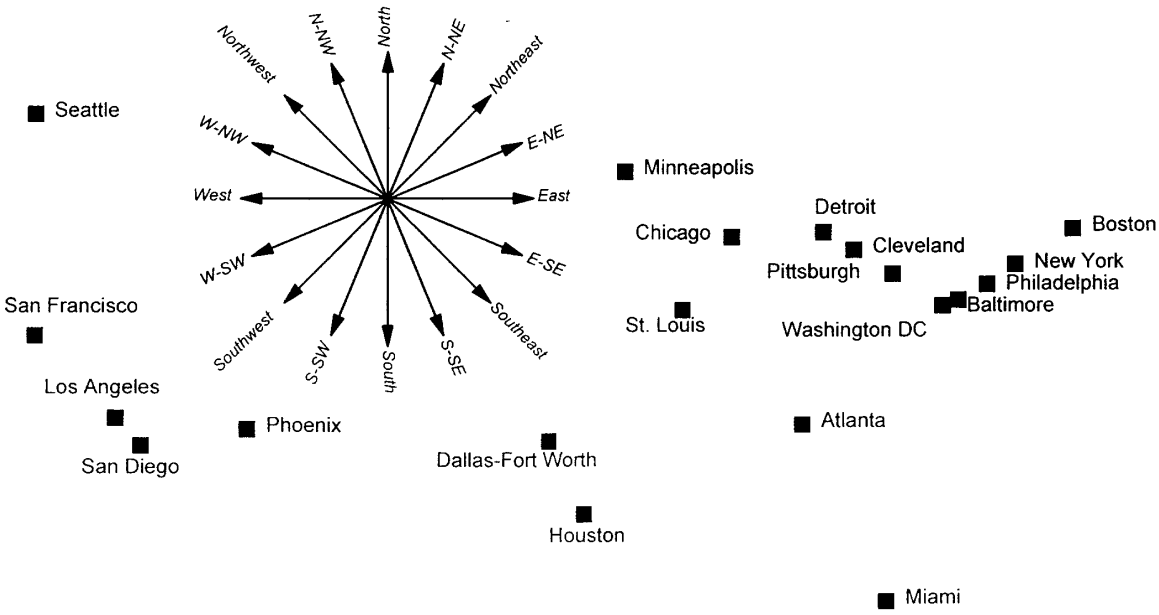
Our purpose is to compare DA and CA empirically, to see how their results compare in practice. We are interested in the question of which technique does the better job overall, and also when subsets of attributes or subsets of products are removed from the data set. There has been controversy in the literature about what conclusions can be drawn about relationships between products and attributes with CA (Greenacre, 1989). We do not intend to become involved in that argument, and out investigation is strictly empirical.

Our Approach

We created an artificial data set that we hoped would be familiar and intuitively meaningful, and which would also have precisely known properties. As "products" we used the 20 largest metropolitan areas in the U.S. As "attributes" we used directions of the compass rose. (Figure 1.)

FIGURE 1

"TRUE" DIRECTIONS AND LOCATIONS OF CITIES



Our physical locations were the principal airports serving these metropolitan areas, for which latitude and longitude coordinates are available. We wanted the precision provided by latitude and longitude coordinates, but that caused a problem in perspective. In the U.S. each degree of latitude accounts for an average of about 30% less distance in miles than each degree of longitude. We restored normal perspective by increasing north-south distances by 30%.

That gave us the coordinates of each city with respect to North, South, East and West. For those directions, we computed "attribute scores" for each city that were positive if the city was in the indicated direction from the centroid and negative if in the opposite direction, with a common unit of measurement. Next, we computed each city's projections on the remaining "attribute vectors" corresponding to the immediate compass directions, so as to obtain scores on a total of 16 attributes," corresponding to compass directions at intervals of 22.5 degree. We considered that configuration to be the "true map," which we would attempt to reproduce using DA and CA.

Next, we used a Monte Carlo approach to create a data base, analogous to what one might obtain by asking 100 respondents to rate each city with respect to each direction, using 5-point scales. The following steps were performed 100 times for each city:

1. Look up the city's "true score" on each attribute (which varied in the range of about plus or minus 25 units, with average standard deviations of about 10 units).

2. Add random normal error with a standard deviation of 10. (Note that this is a large amount of error, being approximately equal to the mound of true variation among cities.)

3. Discretize the resulting values to five categories by recoding:

Range	Code
SCORE < -22.5	1
-22.5 = SCORE < -7.5	2
-7.5 = SCORE < 7.5	3
7.5 = SCORE < 22.5	4
SCORE = 22.5	5

From this artificial data set we computed several maps, using both DA and CA, with various subsets of cities and compass directions. We used software provided by SPSS, although we have confirmed that identical results are provided by other software packages.

As a measure of "goodness of fit" or similarity between each map and the true map, we used the square of the Pearson product moment correlation coefficient between the intercity distances implied by that map and those of the true map. Neither DA nor CA is likely to produce a map with the familiar north/south and east/west orientation, so we performed orthogonal rotations to orient each map in approximately the right way. Distances are not affected by orthogonal rotations, so those rotations had no effect on our goodness of fit measure.

For diagnostic purpose, we also computed as "aspect ratio" of each map. Minneapolis and Houston have approximately the same east/west coordinates, but differ by about 20 units in the north/south direction. Washington and San Francisco have approximately the same north/south coordinates, but differ by about 45 units in the east/west direction. The ratio of the Minneapolis/Houston distance vs. the Washington/San Francisco distance can serve as a measure of the aspect ratio of each map. For the true map, the ratio of these two cities' distance is 20 / 45 = .44.

Results

All Cities and All Directions

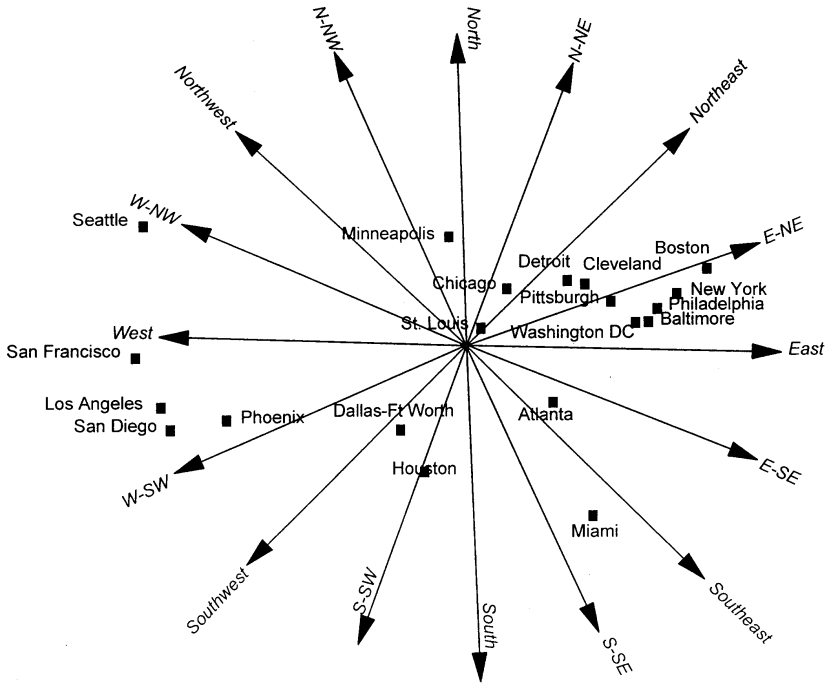
First we used DA and CA to compute a map from this data base, using all 20 cities and 16 attributes.

The DA map was based on the individual 5-point scores obtained by discretizing the data after adding random error. The discriminant dimensions are that that do the most effective job of distinguishing among cities, compared to variation among descriptions of the same cities. The two largest dimensions account for 99% of the total between-city variation. The city coordinates are values for the two discriminant functions, evaluated at the mean of each city. The attribute vectors are plotted using as coordinates their pooled within-group correlations with the dimensions.

The DA map (Figure 2) reproduces the true map quite faithfully.

FIGURE 2

DISCRIMINANT ANALYSIS: ALL CITIES – ALL DIRECTIONS



The r-squared value between its distances and the true distances is .99. R-squared can be interpreted as a percentage of variance accounted for, so its complement can be interpreted as a percentage of error, indicating a relative error level of 1%. This may be a surprisingly successful recovery of the true map, considering the large random component in the data.

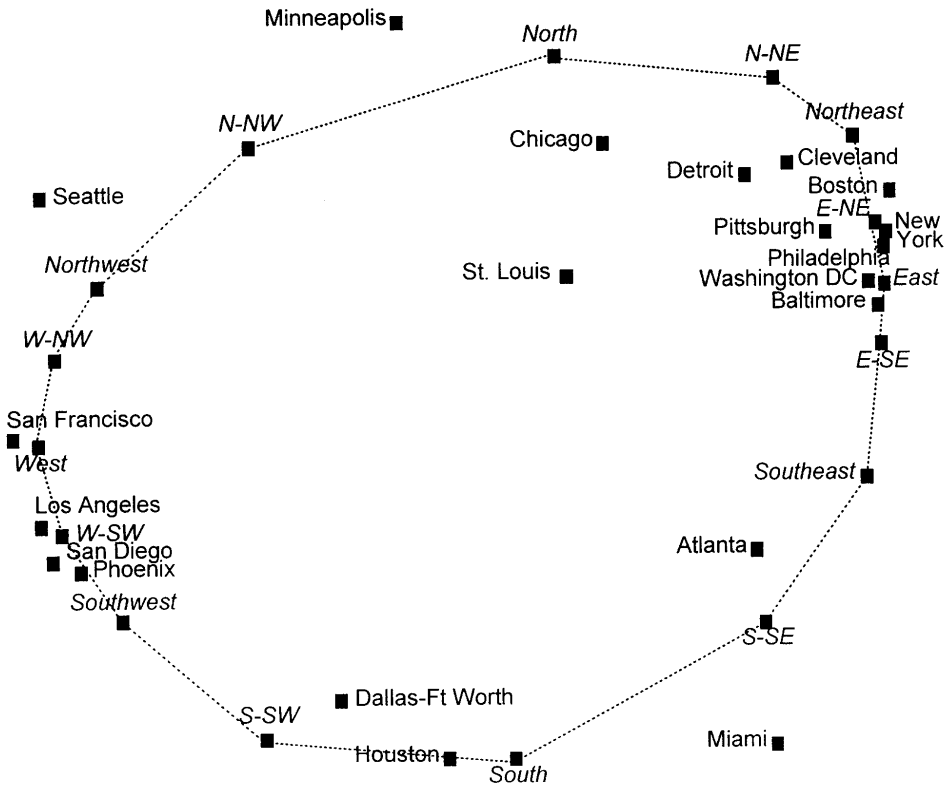
Also, it can be seen that the angles between adjacent compass directions are approximately equal, and he attribute vectors are approximately equal length, as they should be. The aspect ratio for this map is .47, reasonably close to the true aspect ratio of .44, though it does indicate a slight tendency for DA to exaggerate variation in the north/south direction.

With only 100 observations per city, our artificial data set simulates a study with only modest sample size. Yet the excellent recovery of the true map, even in the presence of a large amount of random error, confirms that DA-based perceptual mapping is a robust technique that can produce good results even with small sample sizes.

Next we used CA to produce a similar map (Figure 3), but from summarized data. The data used were "top two box scores," the percentages of times that each city scored in the top two categories for each attribute, a method of aggregation we believe to be frequently used with CA.

FIGURE 3

CORRESPONDENCE ANALYSIS: ALL CITIES – ALL DIRECTIONS



The CA map accounts for 96% of the between-city variation of aggregate data. However, the r-squared between its distances and the true distances is .83, indicating a relative error level of 17%; a much worse recovery of the true map than provided by DA.

CA displays the attributes as points rather than as vectors. We have connected adjacent attributes as a visual aid. The attributes are not equally spaced. Their outline is somewhat egg-shaped rather than circular, and they are closer together in regions of dense city points than in regions where there are few cities.

Finally, the aspect ratio of this map is .80, much larger than the values of the .44 for the true map and .47 for the discriminant map. In the true map there is much more variation in the east/west direction than in the north/south direction. However, CA indicates a tendency to exaggerate north/south variation among cities, tending to equalize the amount of variation in the two dimensions.

DA provides a substantially better fit to the true map in every way. DA does a better job a reproducing relationships among the cities, as evidenced by the r-squared value,

and also represents relationships among the attributes with a pattern more nearly circular in shape and equally spaced, as in the true map

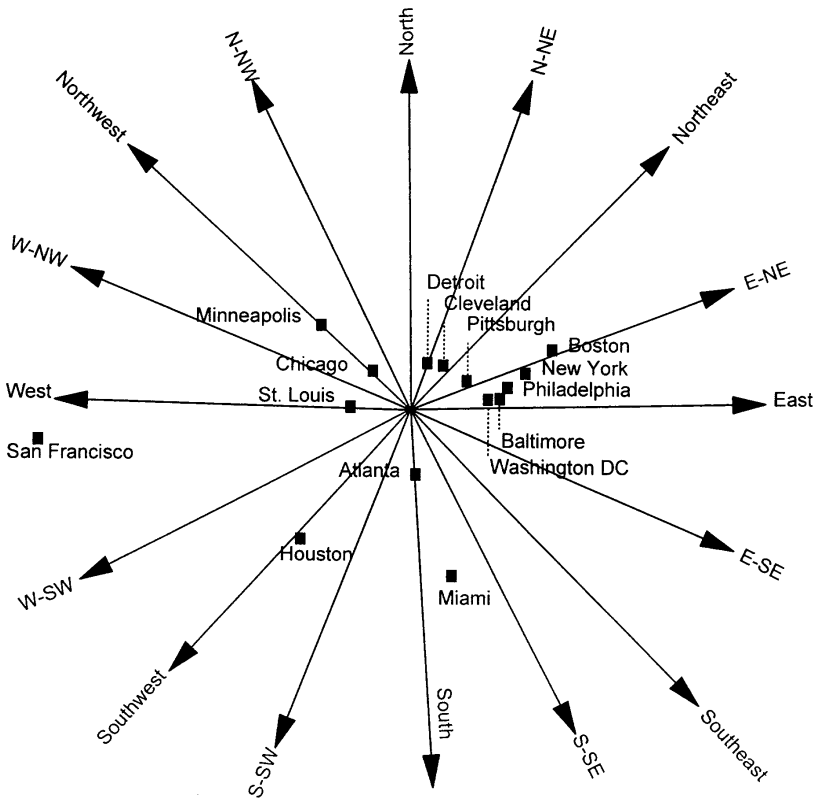
Subset of Cities, All Directions

U.S. cities are more densely concentrated in the eastern than the western half of the map. We experimented with increasing this imbalance by dropping five of the six western-most cities, although retaining San Francisco for use in comparing aspect ratios for the resulting maps. We anticipated that CA would be more upset than DA by the resulting imbalance from having cities much more densely concentrated in one half of the space than in the other half.

The centroids of both maps are shifted to the East, as would be expected. The r-squared value for the DA map (Figure 4) is .98, compared to .84 for the CA map. Neither value is much different from what was observed for all 20 cities, and DA again reproduces intercity distances much more successfully.

FIGURE 4

DISCRIMINANT ANALYSIS: FIVE WESTERN CITIES DROPPED – ALL DIRECTIONS

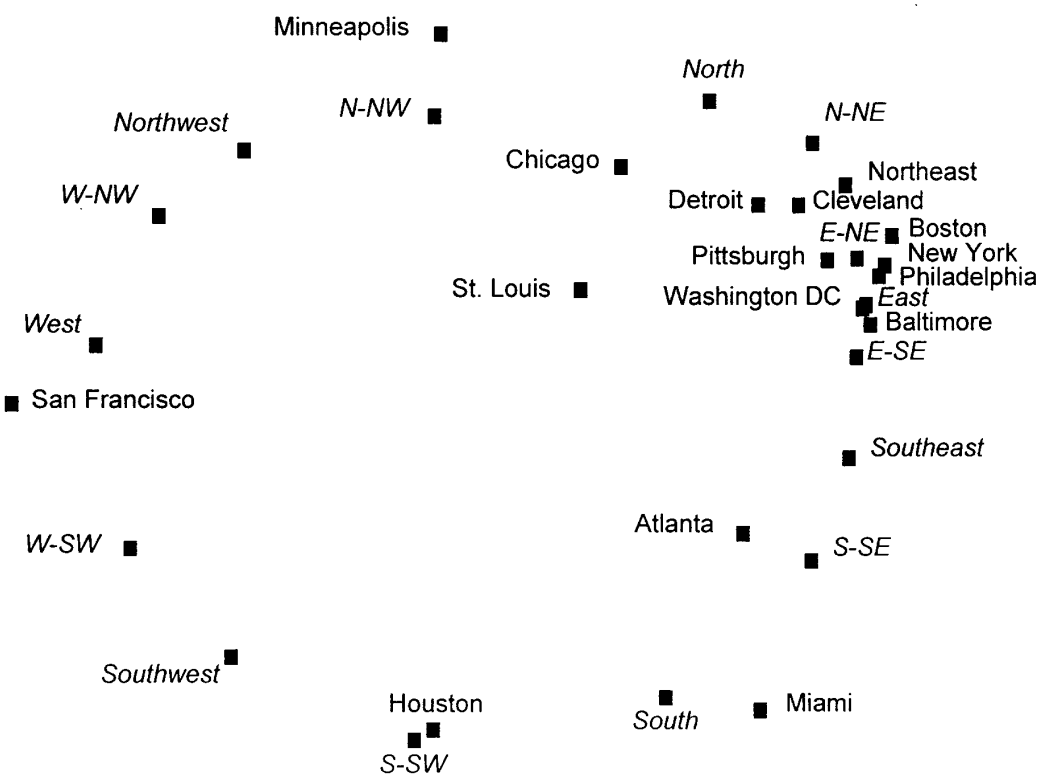


This DA map has an aspect ratio of .47 and this CA map has an aspect ratio of .84, neither much different from what was obtained with all 20 cities. The aspect ratio for DA is still much closer to the value of .44 for the true map.

One difference can be seen in the treatment of the attributes in the CA map (Figure 5). With 20 cities, attribute points were more densely concentrated in the southwest than in the north or south, corresponding to the fairly dense clustering of San Francisco, Los Angeles, San Diego, and Phoenix. With three of those cities dropped, cities are quite sparse throughout the western half of the map, and, apparently as a result, the western attributes are also more nearly equally spaced.

FIGURE 5

CORRESPONDENCE ANALYSIS: FIVE WESTERN CITIES DROPPED – ALL DIRECTIONS



All in all, though, it seems that making the density of cities less balanced by dropping most of the western cities has only a minor effect on either map's positioning of the cities with respect to one another. This is good news, particularly for CA, which we had conjectured might be more vulnerable to this data manipulation than DA.

Subset of Directions, All Cities

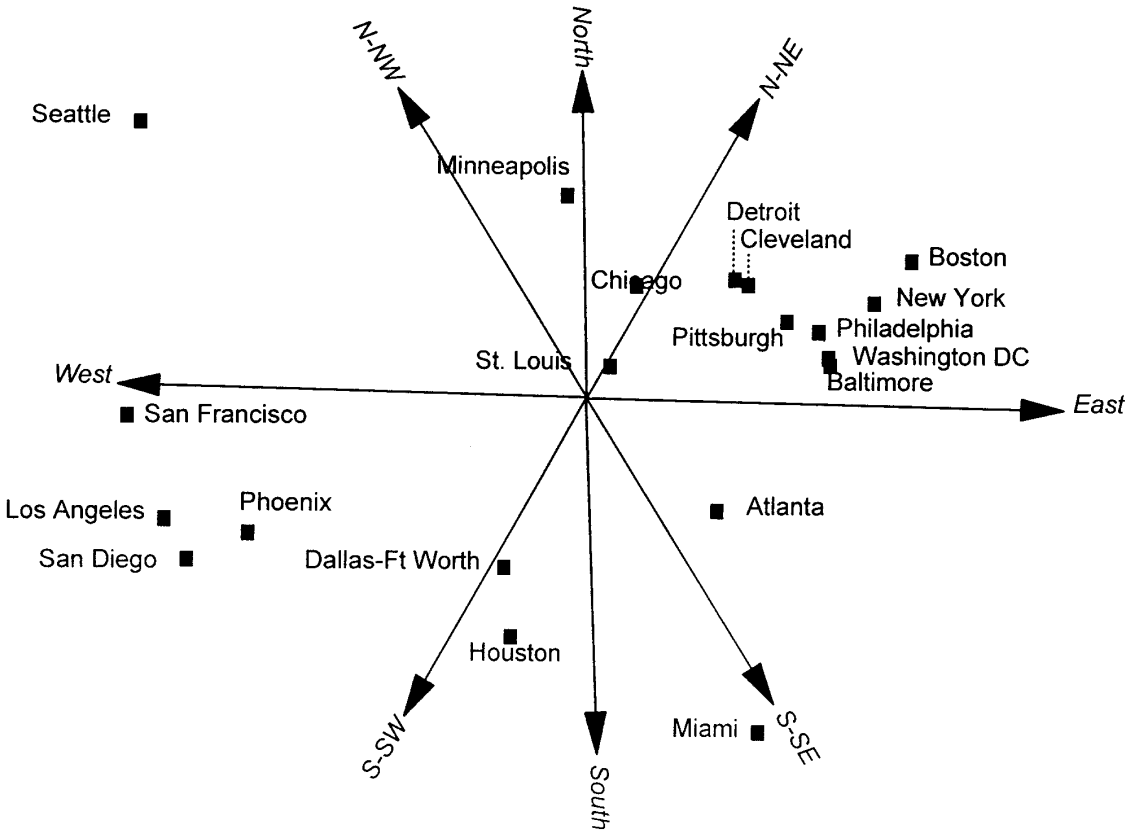
Our compass directions were chosen to be equally dense in all directions, with adjacent vectors having angular separations of 22.5 degrees. This ensured that there would be the same amount of information about city locations with respect to all directions. In the next comparison, we deliberately reduced the amount of information available in the easterly/westerly direction, by deleting eight of the sixteen directions. We retained the East and West attributed, but dropped NE, ENE, ESE, SE, SW, WSW, WNW, and NW. Among the attributes remaining, two defined the east/west direction, two defined the north/south direction, and four more were retained that were within 22.5 degrees of the north/south direction.

We did this to see how the two methods would deal with a data set that had much richer information about north/south differences among cities than east/west differences.

Both maps were notably less successful at reproducing the true distances among cities. The r-squared value for the DA map (Figure 6) drops to .93 for a relative error level of 7%, and r-squared for the CA map (Figure 7) drops to .73 for a relative error level of 27%. With the relatively greater amount of north/south information available, both methods exaggerate the variability of cities in the north/south direction, with aspect ratios of .63 for DA and .93 for CA, compared to .44 for the true map.

FIGURE 6

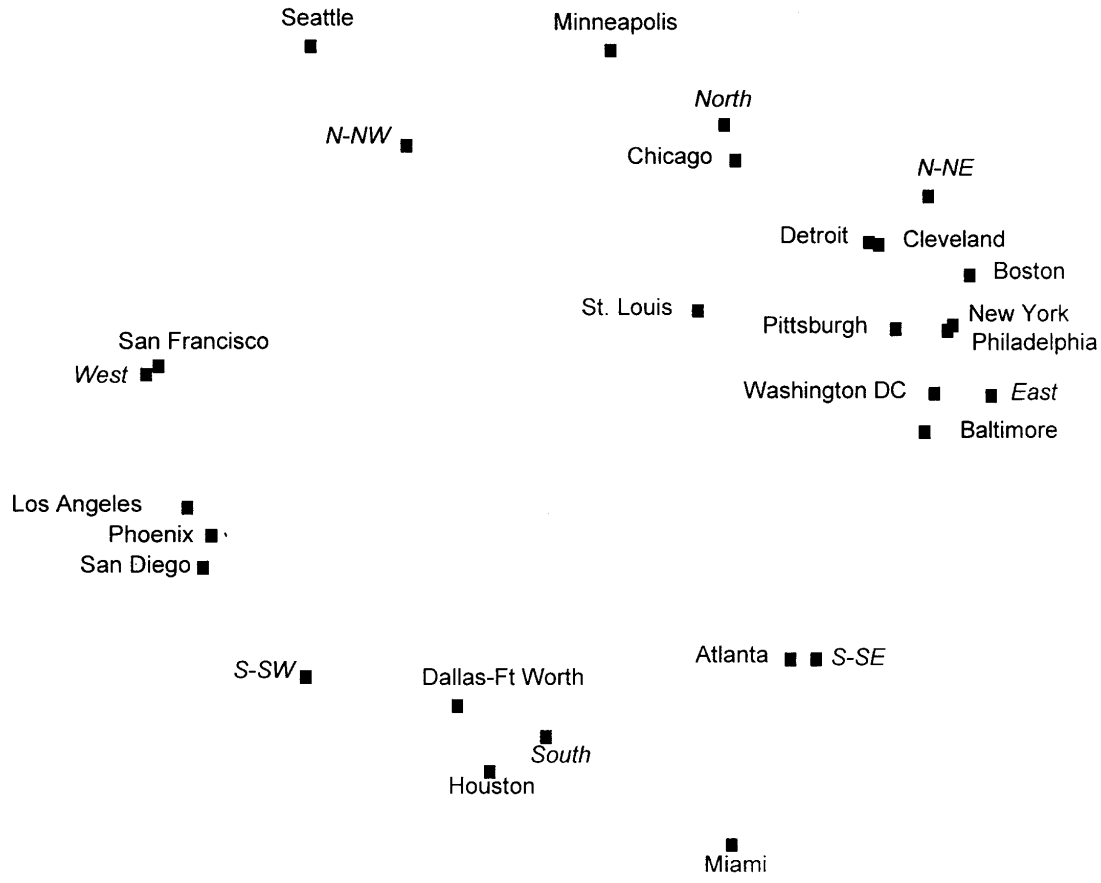
DISCRIMINNAT ANALYSIS: ALL CITIES – EIGHT DIRECTIONS DROPPED



Comparing relationships among attributes, both techniques have some tendency for "diagonal" directions to be represented as closer to east/west than they should be. However, the directions are much more faithfully reproduced by DA, which maintains apparent orthogonality between north/south and east/west. CA, by comparison, place N much closer to E than to W.

FIGURE 7

CORRESPONDENCE ANALYSIS: ALL CITIES – EIGHT DIRECTIONS DROPPED



Overall, DA again reproduces the underlying structure more successfully in every way than CA. Although both methods err by exaggerating differences among cities in the direction of corresponding to the richer representation of the attributes, this tendency is no greater for CA than DA.

Summary and Conclusions

This study compares the abilities of Discriminant Analysis (DA) and Correspondence Analysis (CA) to recover the true structure of a synthetic data set featuring 20 U.S. cities as "products," and 16 compass directions as "attributes."

The data set was constructed using Monte Carlo methods, similar to what respondents might produce when rating products on attributes: each city's true projection on each compass direction was first obtained, expressed in units of distance from the center of the map, and then a random component was added. The resulting values were discretized into 5 categories, as though they had been produced by respondents using

5-point rating scales. This was repeated 100 times, as though each of 100 respondents had fallibly rated each city in terms of each direction.

The simulated individual respondent data were subject to DA to attempt to recover the relations among cities, and relations between cities and directions. The individual data were also aggregated by counting the number of times each city was scored in the "top two boxes" for each attribute, and those aggregate frequencies were subjected to CA. Additional maps were also computed after systematic deletion of certain cities and directions.

We had expected that DA would be less affected than CA by deleting a group of cities, and by deleting clusters of attributes. However, both methods seemed to be little affected by deleing a group of cities, and both methods seemed about equally affected by deleting groups of attributes.

Although both techniques reproduced recognizable maps under all circumstances, DA always reproduced the true configuration much more accurately than CA. This was assessed both quantitatively in terms of correlations between true and reproduced intercity distances, and qualitatively by the visual integrity of the resulting maps.

We conclude that CA is easier and less demanding, but that DA does a better job. When data at the individual level are available, we believe it would generally be preferable to use DA.

NOTES

The author gratefully acknowledges the helpful suggestions of Rich Johnson, Chairman of Sawtooth Software, Sequim, Washington and thanks Marjorie Deninger, Director of Marketing Research, Home Box Office, New York, for suggesting the topic.

Carroll, J. Douglas, Paul E. Green, and Catherine M. Schaffer "Reply to Greenacre's Commentary on the Carroll-Green-Schaffer Scaling of Two-Way Correspondence Analysis Solutions," *Journal of Marketing Research* 26 366-8.

Greenacre, Michael J. (1984) *Theory and Applications of Correspondence Analysis*. London: Academic Press, Inc.

Greenacre, Michael J. (1989), The Carroll-Green-Schaffer Scaling in Correspondence Analysis: A Theoretical and Empirical Appraisal," *Journal of Marketing Research*, 26 (August), 358-365.

Hoffman, Donna L and George R. Franke (1986), "Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research," *Journal of Marketing Research*, 23 (August), 213-27.

Johnson, Richard M. "Multiple Discriminant Analysis," unpublished paper, "Workshop on Multivariate Methods in Marketing," University of Chicago, 1970.

Johnson, Richard M. "Market Segmentation: A Strategic Management Tool," Journal of Marketing Research, 8 (February, 1971), 13-8.

Kruskal, Joseph B. “Nonmetric Multidimensional Scaling: A Numerical Method,” Psychmetrika, 29 (June 1964), 115-29.

Lebart, Ludovic, Alain Morineau, and Kenneth M. Warwick (1985), *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques*. New York: John Wiley & Sons, Inc.

SPSS Categories, Release 5.0; *SPSS Advanced Statistics, Release 5.0*, SPSS, Inc. Chicago, 1993.

Young, F.W. “TORSCA, An IBM Program for Nonmetric Multidimensional Scaling,” *Journal of Marketing Research*, 5 (August 1968), 319-21.