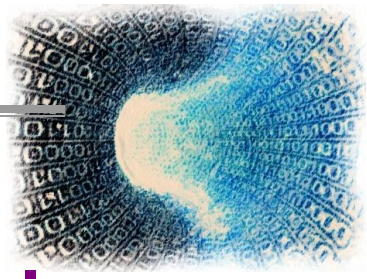
DATA USE



Multivariate and perceptual mapping with discriminant analysis

Editor's note: Steven Struhl is vice president, senior methodologist with Total Research Corporation.

Discriminant analysis can give you highly useful multivariate or "perceptual" maps of many types, including a few that should be better known. These maps often neatly summarize large amounts of information and can provide particularly strong insights about which of the variables you are studying best differentiate among groups, while showing how these variables relate to each other.

Discriminant analysis: a brief review

Discriminant analysis, as a reminder, was designed expressly to determine what best distinguishes among, or tells apart, groups. This makes it an excellent technique for use in research, where we often need to address such questions as how cluster groups differ, what differentiates those "extremely likely" to buy from others with less interest, and so on.

This procedure requires one dependent or "grouping" variable. This variable must be categorical, of course, and not continuous as it can be in regression. You must identify each respondent (or thing you are analyzing) as the member of one group.

You can become quite creative about what constitutes a "group," though, as long as no individual gets into more than one. We have encountered many original grouping variables - and sometimes invented a few. For instance, we came across a single grouping variable created from both age and income information. In this case, Group 1 was defined as those age 18-24 with incomes under $35,000; Group 2 as those age 18-24 with incomes over $35,000; Group 3 those age 25-35 with incomes under $35,000, and so on. Another original grouping variable divided respondents based on both their first and second favorite products in a category.

You can have a large number of groups, if your sample (and computer) will allow it. We have seen some analyses with over 50. With many groups, however, you should expect relatively low levels of correctly predicting who belongs in which group, although you can still get interesting maps.

Discriminant mapping basics

Discriminant analysis produces functions, or equations, that combine the independent variables. Each equation looks like a regression equation, in that it combines variables additively with a weight or coefficient given to each. However, discriminant analysis does not try to predict some specific value of y from a combination of x variables, as in the familiar regression form:

$$y = a_1x_1 + a_2x_2 + a_3x_3 \ldots + c.$$

Rather, discriminant analysis seeks to differentiate most strongly among the groups identified by the dependent variable. It develops coefficients for each independent variable ("x") that lead to total scores (a set of "y" values) for each respondent, with the goal of making the scores in each group as different as possible from those in all other groups.

To make matters more complicated, discriminant analysis can (and usually does) produce more than one function or equation. The number of functions is limited by the number of groups, or the number of independent variables—and must be at least one less than the number of variables or equal to the number of groups, whichever is smaller. So, a discriminant model looking at four groups and fifteen independent variables could have up to three dimensions. An analysis looking at ten groups and seven independent variables could have up to seven dimensions.

These dimensions are strictly independent of each other, and so fall at right angles to each other. A map of the first two discriminating dimensions would fall onto a standard x/y plane. Three dimensions becomes difficult to plot, and more difficult to interpret. You can get four dimensions (more or less) onto a single map, by overlaying different colors for values in the fourth dimension onto a three-dimensional surface. Trying to get a client to understand a map like this is another matter.

Fortunately, in many cases, the first two or three dimensions identified explain most of the variance, or patterns of differences, in the data. Plotting these often lets you see everything important. This ability to combine many variables into "dimensions" can give you the next best thing to "looking into high-dimensional space." That is, you can see the effects of many variables at the same time, perhaps more easily than with most other techniques.

Discriminant analysis also will provide you with other information that you can display graphically. For instance, it shows how well each group has been identified, gives detailed information about how much groups look alike, and even can provide each respondent's likelihood of belonging to each group.

Point-vector maps from discriminant analysis

Sometimes you will see a discriminant map in which groups being analyzed are shown as points and the significant independent variables are shown as vectors. The points represent the averages (or "centroids") of the groups on the dimensions shown. In the example following we will consider a map based on a survey questionnaire, with each variable being simply the responses to one question. The dependent or grouping variable identifies each respondent as a member of one group developed by a clustering analysis.

A. Putting respondents on the map

The group averages get plotted by using respondents' answers to each question contributing to a dimension. (Usually the "raw answers" get transformed into standardized form—t hat is, each variable is re-scaled to have a mean of zero and a standard deviation of 1). The contribution of each variable is shown by its coefficient. Let's suppose we have two variables that contribute to dimensions 1 and 2 as shown in the following table.

	Var1	Var2
Dimension 1	.90	.10
Dimension 2	.20	.80

Now suppose you have three respondents, whose standardized scores on variable 1 are 0.5, 0.6 and 0.7, and whose standardized scores on variable 2 are 0.2, 0.3, and 0.4. Their group average on dimension 1 would be:

.90 x ((0.5 + 0.6 + 0.7)/3) + (.10 x ((0.2 + 0.3 + 0.4)/3))), or .54 + 0.03, or 0.57

On dimension 2, their average would be:

.20 x ((0.5 + 0.6 + 0.7)/3) + (.80 x ((0.2 + 0.3 + 0.4)/3))), or 0.12 + 0.24, or 0.36

Therefore, the group average would get plotted at (.57, .36) on the map.

B. Putting variables on the map

Vectors representing independent variables can get drawn in any of several ways. Perhaps the simplest of these uses the variable's coefficient on each dimension to plot its location. For example, if a variable had coefficients of 0.90 on dimension 1 and 0.20 on dimension 2, its location would be (.90, .20), using the usual convention of showing dimension 1 on the x-axis.

Figure 1 shows a point-vector map, with the variables and groups labeled. You need to decide on the group labels by examining the position of each group, and the concerns most strongly related to it. The variable labels simply reflect the questions asked in the survey.

One problem with this approach is that the group "centroids," or averages, get larger –and so further from the center of the map –as more variables enter the discriminant functions. With a large number of variables (say, 50 or more), you often find the group centroids far beyond all the individual variable vectors on the map. Trying to plot everything together can lead to minuscule-seeming variable vectors that are hard to read.

Various solutions have been proposed for this problem, including multiplying the variable vectors by several types of constants. It is probably simpler to use a scale on the map that allows you to see the vectors easily, and then indicate where the group centroids would fall by drawing arrows to the edges of the map.

Interpreting the map is fairly simple, once you have determined what was done to produce it. If the variable vectors have been left unchanged, then the length of each reflects its effect in discriminating on each dimension. Longer vectors pointing more closely toward a given group average, or centroid, represent variables most strongly associated with that group. Vectors pointing in the opposite direction represent concerns associated less with members of that group than other groups. You can often label the axes in discriminant analysis, just as you would in factor analysis. Variables with long vectors in a given dimension, and particularly those with long vectors closest to the axis, have the most to do with that dimension. Looking at the variables that influence each dimension most strongly can often give you a name for that dimension.

However, you will sometimes find two or more "ideas" in a single discriminant dimension–unlike factor analytical dimensions, which tend to capture single ideas. This happens because discriminant analysis does not intend to group similar variables, but rather to find the combination of variables where responses best distinguish between groups of respondents.

Some authors have suggested plotting the correlations between variables and each discriminant function. This is possible, but creates some difficulties. Correlations measure similarities and not effects. Because variables usually are correlated to each other, as well as to the discriminant function, one variable will sometimes largely explain the effect of another. When this happens, a variable having a fairly high correlation with the function can have a small coefficient–its effects are largely redundant. Since the variable in this case would do little to distinguish between groups, it would be misleading to show it with a long vector. Most audiences expect variables

shown with long vectors to have strong effects. This is not always true when you plot correlations.

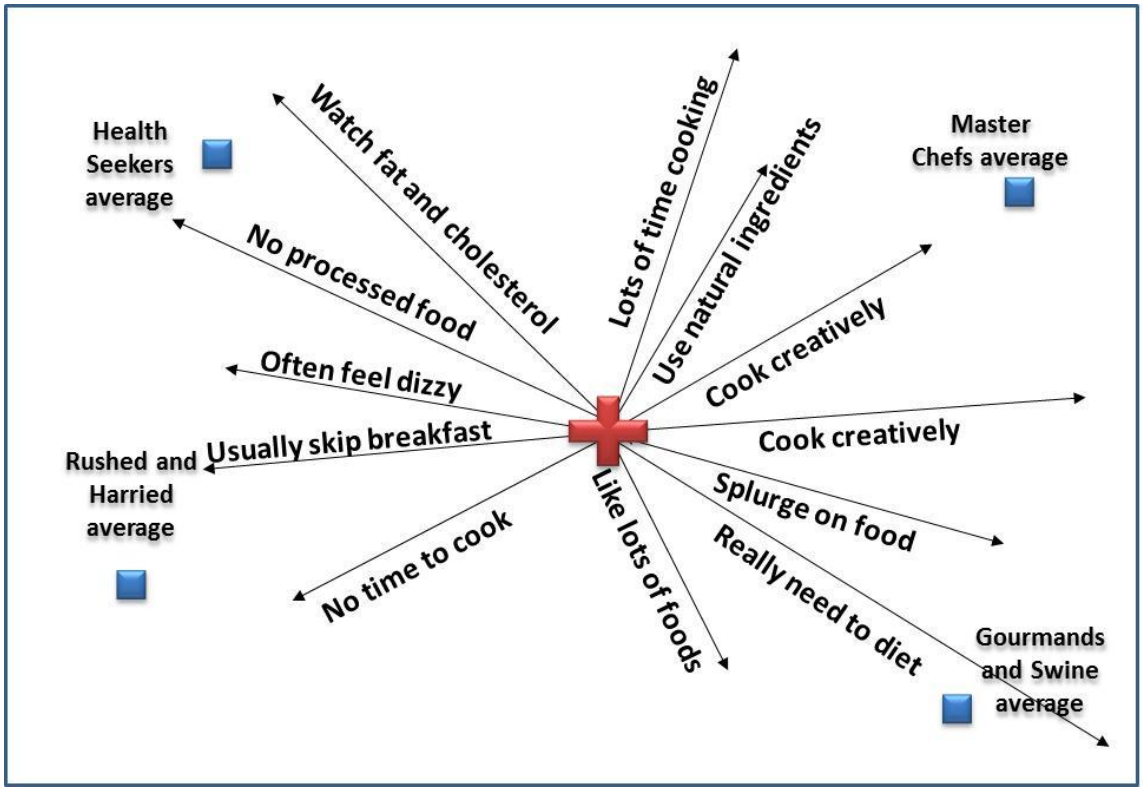


Figure 1

I prefer using both the correlations and coefficients for information, but plotting the coefficients. If a variable has a high coefficient and a low correlation (the opposite of the situation above), we would need to reconsider the model or drop the variable entirely. Sometimes you find that variables with high coefficients and low correlations with the axis are mostly helping to classify (or identify group membership) of a few respondents with strange response patterns. Discriminant analysis, like regression, can suffer from incorrectly specified models—too many variables or key variables missing. This is signaled in particular by the coefficient of the variable running in the opposite direction from what you expect. If this happens, see if you really need all the variables and try the model without those that are not critical—until effects fall into place.

The all-group scatter-plot

This type of plot shows how members of various groups fall on a pair of discriminant axes or dimensions. Sometimes, now that plotting software is quickly improving, you sometimes will see 3-dimensional scatter-plots. In our opinion, these can be highly impressive, but usually end up looking too complex to help simplify data.

All-group scatter-plots obviously provide a lot more detail about respondents than a simple group centroid or average, and can help you see how groups tend to overlap, or where they can most easily get confused.

In the example (Figure 2), the analysis produced 5 dimensions, so you would not necessarily expect to find clear structures in just the first two dimensions. What you see in this plot is something like the "shadows" of the six groups projected onto two dimensions from five dimensional space. Since this plot does not show three of the five dimensions, the groups could possibly be separated in a way we cannot see. However, if the first two dimensions explain all or

nearly all of the variance in the discriminant solution, you would expect to see clearly-defined groups in the scatter-plot.

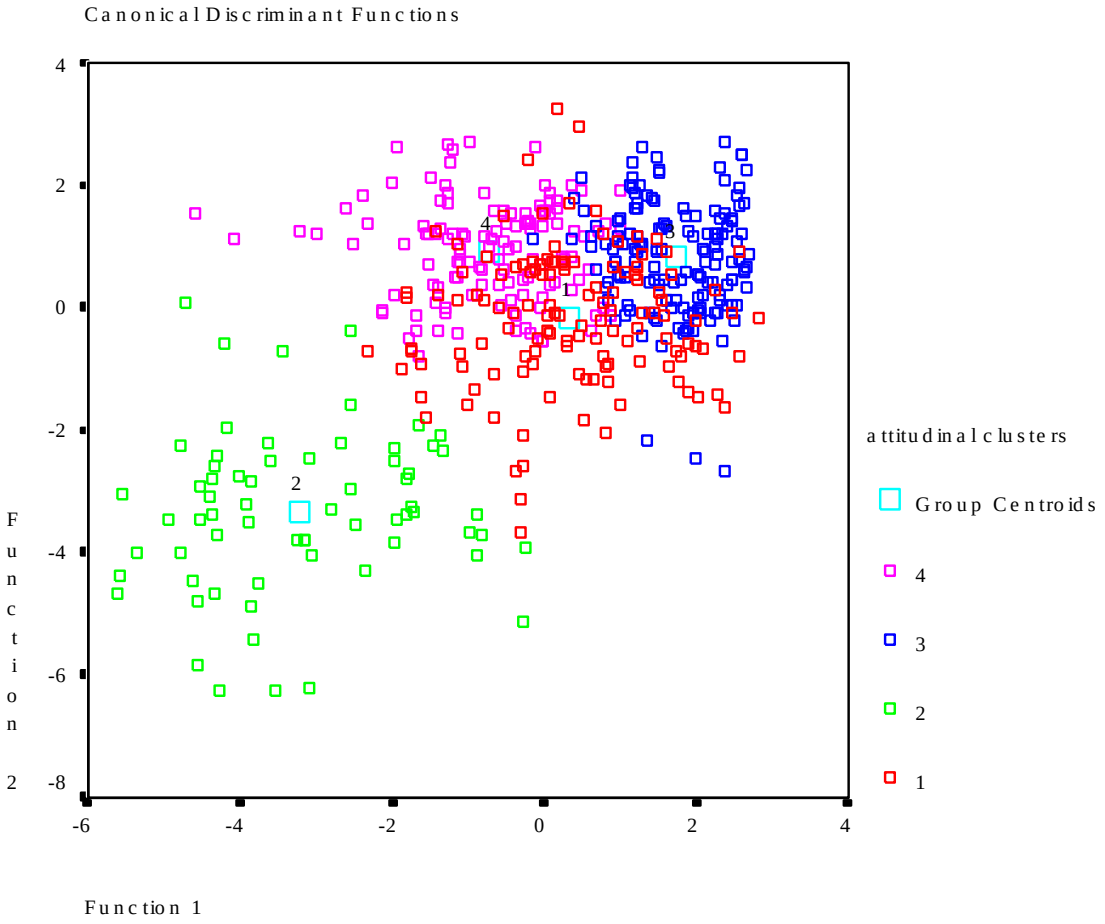


Figure 2

The discriminant "territorial map"

A. What are territories?

This map provides a concise summary, but more information than a point-vector map about how groups differ on the discriminating dimensions. Based on the discriminant scores of the groups' members, this map shows regions, or territories, most strongly associated with each group.

Each territory defines a "place" where you are most likely to find members of one group. Recalling that each dimension in discriminant analysis represents a set of variables that distinguish between groups, you can see how study participants will get into a territory.

For instance, looking at Figure 3 (following), respondents with high scores in both dimensions are highly likely to fall into Group 4. However, respondents who have middling scores on the first dimension can have a very high score on the second dimension and still fall into Group 3. If respondents have very low scores in dimension 2, they are likely to be in Group 2 regardless of their scores in dimension 1. (We can see this last fact represented by the way the territory for Group 2 extends across the entire first dimension.) Finally, respondents with low scores in the first dimension and high scores in the second are very likely to fall in Group 3.

Crossing a line in the chart below brings you from an area where respondents are most likely to belong to one group into an area where they are most likely to belong to another. This type of map can tell you much more about how groups are similar and different than one simply showing group averages (or centroids). You see what combinations of discriminating variables (such as opinions, behaviors, perceptions) most strongly characterize each group and how extreme respondents' opinions must be for them to belong in a group.

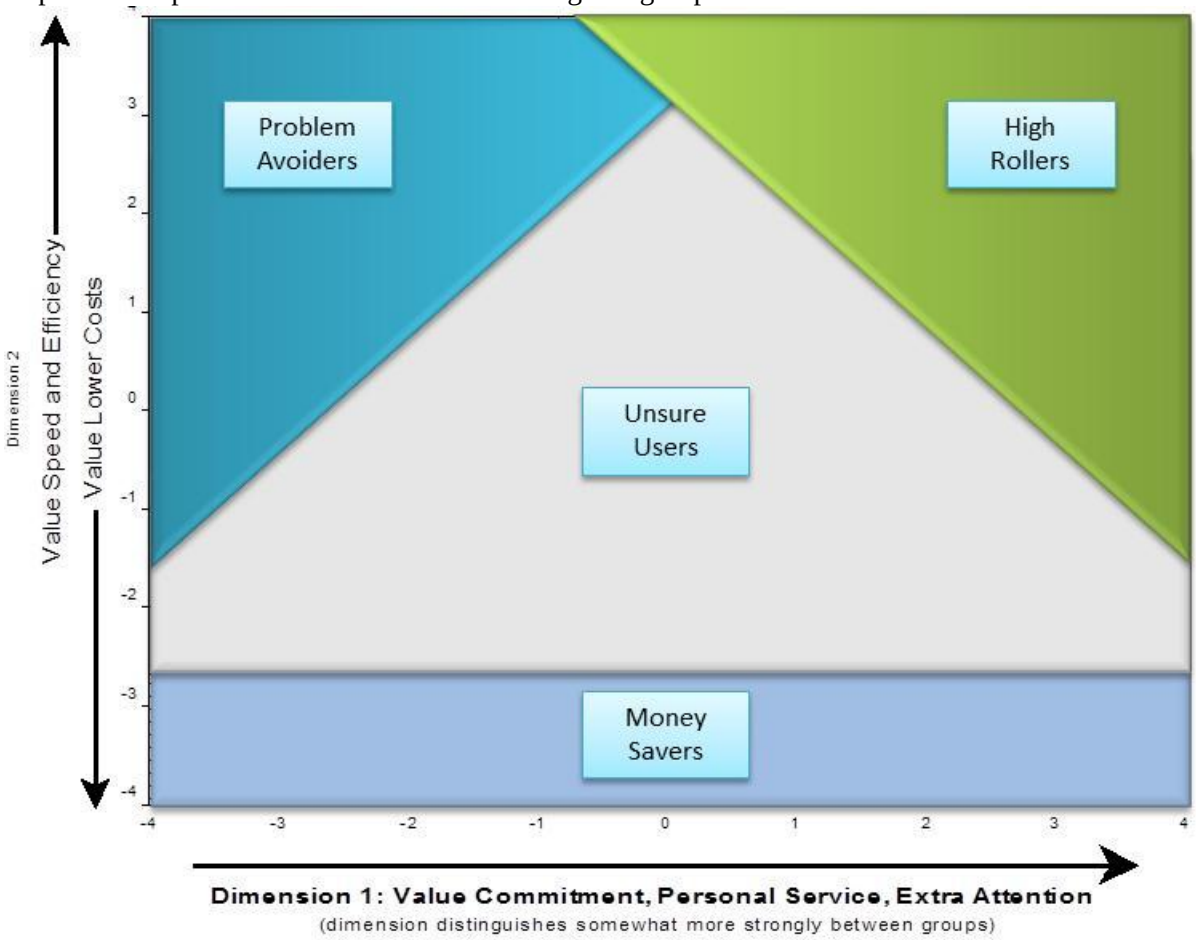


Figure 3

B. Some cautions

These characteristics of the map may require some explanation to audiences unfamiliar with this form of data presentation. For instance, the map does not show the relative size of the groups. Rather, the width of and height of each group represents its range of opinions. Width represents the range of opinions in each group on the first axis or dimension. Groups that vary more on the first dimension will appear wider. Similarly, the height of each group represents its range of opinions on the second dimension. The map does not show how much the groups overlap. Rather, it accentuates the differences between the groups.

Unfortunately, the map, as produced by available statistical programs, needs work before it becomes fully informative. What you get from the statistics program still looks like an old typewriter or a line printer from an old mainframe. You will need to label the dimensions by finding which variables contribute strongly to each. (As a quick rule, look for the variables that have both a relatively high standardized coefficient within, and some correlation with, the dimension.) You will also need to demarcate the territories clearly. Spending some time cleaning up the map and redrawing it (as in Figure 3) can make this a highly informative and visually compelling summary of what differentiates the groups.

Putting techniques together: the vector-territorial map

Overlaying vector for variables and territories for groups can provide an interesting summary of the data, and give more information than the simple centroid (or dot) that defines each group in a point-vector map. So far, while your author has produced a number of these, we do not know of these appearing elsewhere. This type of map could well deserve more use. The example following (Figure 4) uses the same data as in Figure 1.

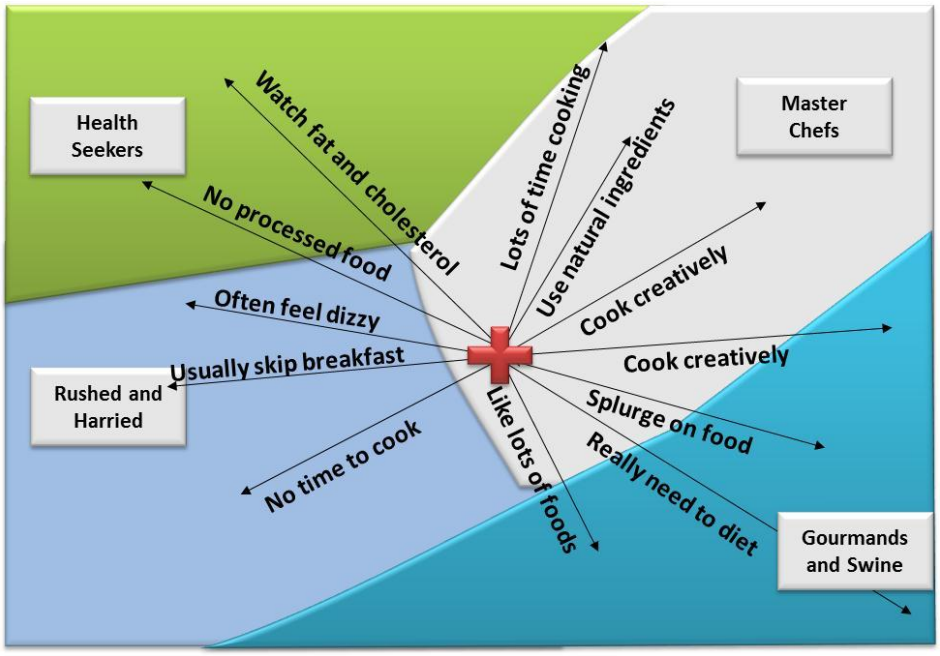


Figure 4

Practical aspects of mapping

These maps are not yet as simple to create as a bar chart, but with advances in charting software they have become relatively painless. The basic data can come from any statistical program that performs complete discriminant analysis. SPSS, Systat and SAS will give you nearly everything you need to do these maps, although (for now) SPSS is the only one that easily and directly produces a territorial map like the character-based (typed) figure we mentioned.

One problem with these programs is that they do not always produce the chart output you need for presentations. The character-based scatter-plot and territorial map would hardly be counted aesthetic wonders by your clients. While the Windows-based versions of SPSS and Systat, at the least, will get you close to finished appearance for some types of charts, all charting/plotting programs we have seen so far require you to do work by hand on most charts. No program will put labels along vectors, for instance, and few even will place labels next to (or near) points in scatter-plots.

We prefer to use programs that have powerful on-screen editing features as well as charting capabilities. Many notable competitors have emerged under Windows, and some in the Macintosh environment, so you may want to do some serious shopping. New programs are appearing all the time. Some will even allow you to trace over the character-based output that many statistics programs produce.

Whatever your choice of software, with a little experimentation and patience, you can create maps based on discriminant analysis that summarize masses of information clearly and usefully.