Michael Greenacre
Professor of statistics in the Department of Economics and Business at the Pompeu Fabra University, Barcelona.

Constructing maps of data using correspondence analysis

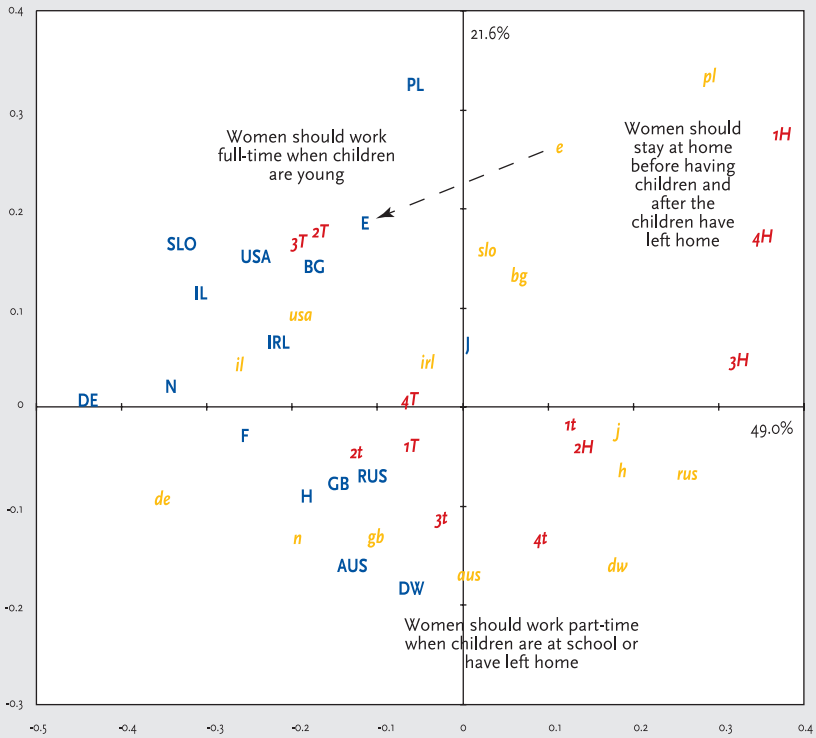
Recently the Spanish stock market suffered a downturn compared to other international markets when the shares in the Spanish construction sector fell dramatically in value. This event could be understood by even the least technically minded person, since the behavior of the Spanish stock market is summarized by an index, the IBEX 35, which serves as an indicator of the overall value of the market. The statement that "in the last two weeks of April the IBEX 35 has dropped by 4.1% while the DOW JONES has increased by 3.2%" is not difficult to comprehend, especially when complemented by the information about the spectacular drop in Spanish construction shares that form an important part of the IBEX 35.

Turning our attention from an economic to a sociological phenomenon, let us try to understand changes in people's attitudes about whether women should work or not during different stages of their married lives; for example, should a woman work when she has a child before school-going age, or should she stay at home? And how do Spaniards feel about this issue in general? The way sociologists would gather information on this issue is typified by a series of questions in the Survey on the Family and Changing Gender Roles, conducted by the International Social Survey Program (ISSP). In this survey, four questions relevant to this issue were asked: first, about women without children; second, about women with a pre-school child; third, about women with a child still at school; and fourth, about women whose children have left home. In all four questions respondents could choose between the answers "should work full-time", "should work part-time" or "should stay at home". Given these various responses by over 2000 people representative of the Spanish population, what defines the "attitude" of Spaniards towards the issue of women working, and how do Spaniards compare with other nations?

Actually, the situation here is not so different from the stock market example mentioned before: both situations are concerned with what statisticians call *multivariate* data. On the one hand, the stock market involves hundreds of shares, of which 35 have been selected and combined by economists in a substantively meaningful way into a single index that reflects the overall value of the market. On the other hand, many of the central concepts in social science, such as attitude, values and culture, can only be grasped through a number of questions pointing to different aspects of the underlying or latent concept of interest. The crucial question is how to combine all the responses to produce a unique or synthetic indicator, just like the IBEX 35 combines different variables into an index. A further question is whether one index will be sufficient to synthesize the data, in other words is it possible that there is more than one *dimension* underlying the responses about women working?

There are ways to build a unique indicator or variable (such as factor analysis, reliability analysis), but often it is more fruitful to engage in a thorough exploration of the data and particularly the relationships

Correspondence analysis map of attitudes to women working in 1994 and 2002, in different countries; the change in Spanish attitudes is indicated by the dashed arrow.



1 – women before having children
2 – women with baby
3 – women with child(ren) at school
4 – women whose children have left home

T – should work full-time
t – should work part-time
H – should stay at home

AUS – Australia
BG – Bulgaria
DE – (former) East Germany
DW – (former) West Germany
E – Spain
GB – Great Britain
H – Hungary
IL – Israel
IRL – Ireland
J – Japan
N – Norway
PL – Poland
RUS – Russia
SLO – Slovenia
USA – USA

(**CAPITAL** letters are 2002 data; *small* letters are 1994 data)

between the question responses expressing an attitude or complex view on an issue and their connections with other social attributes such as gender, age, education, country, etc. Data exploration, without too much subjectivity, is a critical step in the understanding of a question, problem or issue. Prime among the repertoire of exploratory tools is *correspondence analysis*.

Correspondence analysis looks for structure in a set of data in the form of a few dimensions into which a maximum amount of the information content is concentrated. These dimensions can be used to draw maps of points representing variables or their categories (for example, different response options). Instead of inspecting many tables either in sequence or at once, it is more productive to detect patterns or associations through the visual interpretation of the correspondence analysis maps, supported by some accompanying synthetic measures of the quality of the map. The dialogue between theoretical or conceptual interpretation and empirical data or evidence is facilitated by the proximities and distances between the points in the maps.

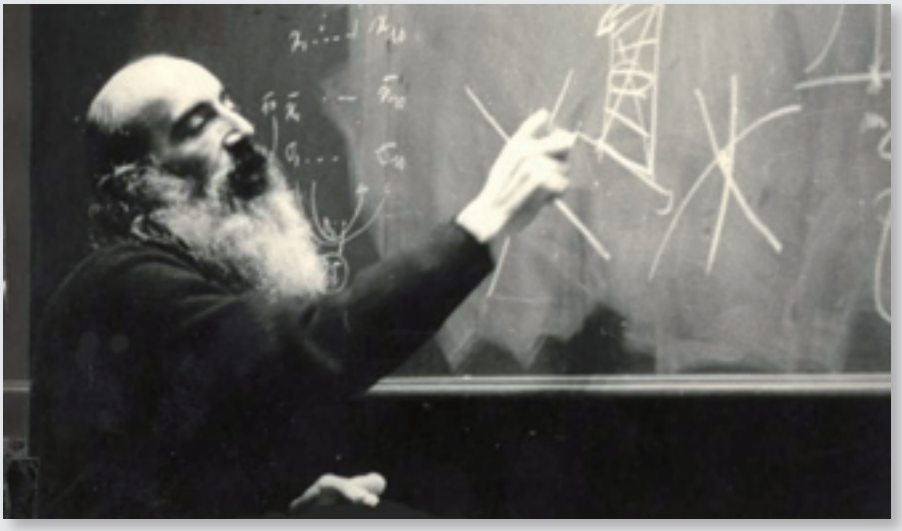
The data for correspondence analysis are usually counts, such as the counts of words in different texts (in linguistics) or the counts of different plant species at different locations (in botany). In our sociological example, the data are the counts of the different responses to the questions about women working, not only from the Spanish sample but also from respondents in many other countries that participated in the ISSP project, totaling tens of thousands of respondents. In addition, we have data from two waves of the survey, in 1994 and 2002, so we can see how attitudes in Spain and in the other countries have changed over this eight-year period.

Using the two principal dimensions of the data leads to the map shown in the accompanying figure. The proximity of two response categories is related to their association: for example, at upper left "women should work full-time when they have a pre-school child" (abbreviated as *2T*) is close to "women should work full-time when they have a child at school" (*3T*), because these responses are often given together, reflecting a highly liberal attitude. Moreover, these responses hardly ever coincide with, and are thus far away from, those on the right hand side: "women should stay at home before they have children" (*1H*) and "women should stay at home when the children have left home" (*4H*), reflecting a highly traditional attitude. In fact, one can see that the data naturally generate a spectrum of attitudes from liberal in the upper left corner to traditional in the upper right corner, between which the response categories follow a curved pattern with intermediate attitudes in the lower part of the map. The countries will now lie at positions according to their responses to the questions. This particular map is centered at the average position of countries in 1994 (the country points displayed by small letters). The countries in 2002 (points in large letters) have all moved to the left, that is towards the liberal side of the attitude map, especially countries like Russia, Poland and Slovenia, and Spain's movement is indicated by an arrow. The position of Spain in 1994 is of particular interest, since it lies within the curve of response categories – this is interpreted as a combination of liberal and traditional extremes, a type of *polarization* of attitudes. In fact, looking at Spain's data in 1994 it appears that 13% of Spaniards said that women should stay at home before having children (*1H*), compared to 5% across all countries, while another 13% said that women with a pre-school child should work full-time (*2T*), compared to 8% across all countries. This position has changed in 2002, the first percentage dropping to 6% and the second increasing to 17%, which accounts for Spain's movement to the left.

The ability of the two-dimensional map to summarize the data set is given by the two figures on the dimensions of the map: 49.0% and 21.6% respectively, totaling slightly over 70%. The positions of the countries on the horizontal axis would be the best single index, or indicator, of attitude, accounting for almost half of the information content of the data. This indicator would place Spain and (former) West Germany at similar positions both in 1994 and in 2002, on such a liberal-to-traditional scale, thus concealing the fact that Spain is more polarized than West Germany, which lies to the bottom of the map in the region of the intermediate compromise response categories.

Data maps like these have found many applications: in marketing research, showing the positions of brands with respect to certain attributes and other brands; in ecology, showing the relationship between species occurrences and environmental conditions; in archaeology, inferring a time gradient according to artifacts found in graves; and in linguistics, positioning different authors with respect to stylistic variables. We have illustrated the use of correspondence analysis to obtain a map of social attitudes, in which the countries are situated as points, where the dimensions of the map are similar in concept to the indices used in the stock market. By contrast with the financial indices, however, there is very little subjective element in correspondence analysis in the determination of the map's dimensions – the method allows the data to "speak for themselves".

IMAGE COURTESY OF M. GREENACRE



Originating in psychology, and popularized by the French mathematician and linguist Jean-Paul Benzécri (in the photo), correspondence analysis is used today in fields as diverse as marketing, sociology, ecology, genetics, music and archaeology. The data maps produced by this technique often appear in the popular press in France, for example in *Le Monde* and *L'Express*.