Fundación BBVA

Documentos de Trabajo 5

2002

Michael Greenacre

The Use of Correspondence Analysis in the Exploration of Health Survey Data

CENTRE DE RECERCA EN ECONOMIA I SALUT-CRES
UNIVERSITAT POMPEU FABRA

Fundación BBVA

The Use of Correspondence Analysis in the Exploration of Health Survey Data

Michael Greenacre

UNIVERSIDAD POMPEU FABRA

Abstract

This working paper gives a comprehensive explanation of the multivariate technique called correspondence analysis, applied in the context of a large survey of a nation's state of health, in this case the Spanish National Health Survey. It is first shown how correspondence analysis can be used to interpret a simple cross-tabulation by visualizing the table in the form of a map of points representing the rows and columns of the table. Combinations of variables can also be interpreted by coding the data in the appropriate way. The technique can also be used to deduce optimal scale values for the levels of a categorical variable, thus giving quantitative meaning to the categories. Multiple correspondence analysis can analyze several categorical variables simultaneously, and is analogous to factor analysis of continuous variables. Other uses of correspondence analysis are illustrated using different variables of the same Spanish database: for example, exploring patterns of missing data and visualizing trends across surveys from consecutive years.

Key words

Correspondence analysis, health survey, principal component analysis, statistical graphics.

Resumen

Este documento desarrolla una amplia explicación de una técnica de análisis multivariante denominada análisis de correspondencias, aplicándola a datos de una encuesta nacional de salud, en este caso, a la Encuesta Nacional de Salud española (ENS). Primero se muestra cómo el análisis de correspondencias puede ser utilizado para interpretar una tabla de contingencia visualizándola en forma de un gráfico de puntos que representan las filas y columnas de la tabla. También pueden ser interpretadas diferentes combinaciones de las variables, codificando los datos de la manera apropiada. Además esta técnica se puede utilizar para obtener valores óptimos de escala para los niveles de una variable categórica, dándole, así, sentido cuantitativo a este tipo de variables. El análisis de correspondencias múltiple puede analizar varias variables categóricas simultáneamente, y es análogo al análisis de factores de las variables continuas. Otras aplicaciones del análisis de correspondencias se ilustran usando diferentes variables de la ENS; por ejemplo, para analizar pautas en los datos perdidos y visualizando tendencias entre encuestas de años consecutivos.

Palabras clave

Análisis de correspondencias, encuesta de salud, análisis de componentes principales, gráficos estadísticos.

La serie Documentos de Trabajo, así como información sobre otras publicaciones de la Fundación BBVA, pueden consultarse en: **http://www.fbbva.es**

DEPARTAMENTO EDITORIAL
DE LA FUNDACIÓN BBVA

DIRECTORA
Paz Pérez-Bilbao

COORDINADORA DE REDACCIÓN Y ESTILO
Mercedes Bravo

The Use of Correspondence Analysis in the Exploration of Health Survey Data

EDITA
© Fundación BBVA. Plaza de San Nicolás, 4. 48005 Bilbao

DISEÑO DE CUBIERTA
Roberto Turégano

DEPÓSITO LEGAL: M-46.658-2002
IMPRIME: Sociedad Anónima de Fotocomposición

La serie Documentos de Trabajo de la Fundación BBVA está elaborada con papel 100% reciclado, fabricado a partir de fibras celulósicas recuperadas (papel usado) y no de celulosa virgen, cumpliendo los estándares medioambientales exigidos por la actual legislación.

El proceso de producción de este papel se ha realizado conforme a las regulaciones y leyes medioambientales europeas y ha merecido los distintivos Nordic Swan y Ángel Azul.

C O N T E N T S

1. Introduction

THE National Health Survey ("Encuesta Nacional de Salud") conducted every three years in Spain is an example of a large complex social survey designed to provide a snapshot of the nation's state of health at a particular moment in time. In the 1997 survey of adults, which will be the subject of this working paper, there are 46 basic questions, some of which consist of possible multiple responses, pushing up the total number of questions effectively to 83. Added to this there are several questions which are conditional on the responses to the basic questions, giving an additional upper limit of 27 questions. Each of the 6,400 respondents interviewed thus provide between 83 and 110 items of information, so that the complete data file comprises approximately 640,000 numbers.

The usual way to summarize these data is to count frequencies of response and present these in the form of bar or line charts. The publication *Indicadores de Salud* (Regidor and Gutiérrez-Fisac 1999), for example, is a collection of tables where the 1997 data are compared to previous years, and only in a few cases are some graphical presentations given of the results as an aid to interpretation.

A second level of analysis is to explore relationships between different questions in the survey. There are various ways of doing this, some more complicated and more ambitious than others. One can, for example, postulate some functional relationship between two variables, say number of visits to the doctor and age. Since both variables are on simple numerical scale, the solution is fairly straightforward and, after inspecting the scatterplot of these two variables, one can establish a regression model relating expected number of visits with age. But when it comes to relating health status, which is a multicategory variable having five possible responses, and the intake of medicines, where there are as many as 17 categories of medicine, the way to proceed is less obvious.

In this working paper we aim to show how correspondence analysis (CA) can be used to explore relationships between variables in a complex health survey, and suggest models for these relationships.

CA is a method aimed specifically at quantifying categorical data, that is, assigning numerical scale values to the response categories of discrete variables, with certain optimal properties. These scale values have been shown to have interesting geometric properties and provide what are called *maps* of the relationships between variables.

After introducing the method, we will give a simple illustration using a cross-tabulation computed from the 1997 health survey. Further applications will be given using more complex cross-tabulations.

We also show how CA can be used to develop scales which synthesize the responses to several questions which have a common theme. This is of great use to the modeller, who can replace several categorical variables by a single scale, which can then be used in subsequent analyses, such as regression analysis, which require interval-scaled data.

Several other issues are dealt with; for example, the exploration of patterns of missing data and how to explore trends between surveys from different years.

2. Correspondence analysis

ALTHOUGH the theory is fully explained in several texts (see Bibliography), we present a practical introduction in the context of the health survey data analyzed in this working paper, as well as a theoretical summary in an Appendix.

In its simplest form, correspondence analysis (CA) applies to a two-way cross-tabulation, such as the one in Table 1. This table summarizes the distribution of perceived health status categories in different age groups. The ultimate aim of CA is to produce a *map* of this table, where each row and each column is represented by a point. The way CA works is quite similar to principal components analysis (PCA), in that the total variance of the table is defined and then this total is decomposed optimally along so-called "principal axes". For mapping purposes, it is usually hoped that a large percentage of total variance is accounted for by the first two principal axes, thereby allowing the table to be visualized in two dimensions.

CA contains three basic concepts: that of a point in multidimensional space, a weight (or *mass)* assigned to each point and, finally, a distance function between the points, called the *chi-square distance*. Once these three concepts are defined, the method tries to reduce the dimensionality of the points by projecting them onto a subspace, usually a two-dimensional plane as mentioned above. This subspace optimally fits the points by weighted least-squares, where each point is weighted by its respective mass, and measurement of distance between points and the subspace is in terms of chi-square distance.

Let us look at each of these three concepts in turn. Since CA is defined equivalently for rows or columns, we shall explain it in terms of the rows of Table 1, with the understanding that the whole explanation applies similarly to the columns, if we simply transpose the matrix.

The rows divided by their row totals are vectors called *profiles*. Whereas these are proportions adding up to 1, in Table 2 we have

TABLE 1: Cross-tabulation of age groups by perceived health status

AGE GROUP	*Very Good*	*Good*	*Regular*	*Bad*	*Very Bad*	SUM
16-24	243	789	167	18	6	1,223
25-34	220	809	164	35	6	1,234
35-44	147	658	181	41	8	1,035
45-54	90	469	236	50	16	861
55-64	53	414	306	106	30	909
65-74	44	267	284	98	20	713
75+	20	136	157	66	17	396
SUM	**817**	**3,542**	**1,495**	**414**	**103**	**6,371**

TABLE 2: Row percentages calculated from Table 1

AGE GROUP	*Very Good*	*Good*	*Regular*	*Bad*	*Very Bad*	SUM
16-24	19.9	64.5	13.7	1.5	0.5	100
25-34	17.8	65.6	13.3	2.8	0.5	100
35-44	14.2	63.6	17.5	4.0	0.8	100
45-54	10.5	54.5	27.4	5.8	1.9	100
55-64	5.8	45.5	33.7	11.7	3.3	100
65-74	6.2	37.4	39.8	13.7	2.8	100
75+	5.1	34.3	39.6	16.7	4.3	100
AVERAGE	12.8	55.6	23.5	6.5	1.6	100

expressed them in the more familiar form of percentages. It is these profile vectors which are the multidimensional points in CA. So our map will attempt to show us these points representing the rows, or age groups in this case, where each age group is described by the vector of five coordinates, its distribution across the health status categories.

Each row profile point will then be weighted by the row mass, which is the frequency of the row category divided by the grand total. For example, since age group 16-24 has 1,223 respondents out of the total of 6,371, then this row point is weighted by the mass 1,223/6,371 = .192. The row masses add up to 1, and are nothing else but the row marginal proportions of the table.

Finally we measure distance between row points by the chi-square distance, which is just a slight variant of the usual physical distance between points in vector space. Usually physical distance between two vectors $x = [x_1\ x_2\ \ldots\ x_n]$ and $y = [y_1\ y_2\ \ldots\ y_n]$ is measured as:

$$\text{physical distance} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2}$$

but the chi-square distance is a distance which weights each squared term as follows:

$$\text{weighted distance} = \sqrt{(x_1 - y_1)^2/v_1 + (x_2 - y_2)^2/v_2 + \ldots + (x_n - y_n)^2/v_n}$$

In PCA this type of distance function is already implied when each v_j is equal to the variance of the j-th variable (as a column) of a data matrix of numerical measurements. Specifically, the chi-square distance between *row* points weights each term inversely by the corresponding *column* marginal proportion c_j:

$$\text{chi-square distance} = \sqrt{(x_1 - y_1)^2/c_1 + (x_2 - y_2)^2/c_2 + \ldots + (x_n - y_n)^2/c_n}$$

where in our example (Table 1) c_1 = 817/6,371 = .128, c_2 = 3,542/6,371 = .556, and so on. The idea here is just like in PCA, in that this division compensates for the different variances in the columns of the profile matrix – we say it is "variance standardizing". Differences between the first column of Table 2 will tend to be smaller, since the percentages are smaller (they actually vary from 5.1 to 19.9, i.e. 14.8 percentage points), whereas differences in the second column will be greater because overall they are larger percentages (they vary from 34.3 to 65.6, i.e. 31.3 percentage points). Dividing by the column margin effectively equalizes out these inherent differences.

The total variance in correspondence analysis is measured by the so-called *inertia,* which is simply the usual Pearson chi-square statistic calculated on the cross-tabulation, divided by the total sample size n. It is this inertia which measures the degree of difference between the age groups that we are trying to represent optimally in the eventual map.

As we have said, the map – usually two-dimensional – is obtained by weighted least-squares (more specific details of the calculations involved are given in the Appendix). In practice, what happens is that the row profile points are projected onto the best-fitting plane. The coordinates of these points are called *principal coordinates,* because they are the coordinates with respect to the *principal axes* of the space. Each principal axis accounts for a certain amount of the total inertia, called the *principal inertia,* usually expressed as a percentage of the total.

In addition, we have points on the map representing the columns as well. There are two ways of representing the columns jointly with the rows. The easier of the two to understand, though not the one that is generally used, is the *asymmetric map* shown in Figure 1. In this map, the row profiles are depicted as described above, in principal coordinates, but the column points are depicted by projections of unit profile vectors onto the same space.

A unit profile vector is a vector of zeros and a single 1; for example, the unit profile vector [1 0 0 0 0] represents the column *Very Good* in the space of the row profiles. The practical problem with the asymmetric map is that the column points are spread out much more than the row points (see Figure 1 as an example). The more conventional joint map is the *symmetric map,* in which both row points and column points are represented in principal coordinates. As shown in the technical appendix, there is a simple scaling factor difference between principal and standard coordinates, which lends some theoretical credence to the symmetric map. One should remember, though, that the symmetric map really involves the projections of two sets of points in different spaces – row profiles in one space and column profiles in another. The interpretation of these maps is explained more fully below in the context of actual examples.

3. A simple illustration

As a first illustration of how CA operates, we look again at the table which cross-tabulates age with perceived health status (Table 1).

We needed to define age groups, which we chose as 16-24, 25-34, 35-44, etc., but this choice hardly affects our eventual results, as we shall point out later. The frequencies in Table 1 are not easy to interpret *per se,* because of the different marginal frequencies in the different age groups, so it is usual to calculate row percentages in order to compare the groups, as shown in Table 2.

The rows of Table 2 are the profiles of each age group across the health status categories. CA visualizes these profiles on a map which depicts the distance between each group and also shows how the health status categories should be scaled in order to visualize these distances optimally. There are two ways to report this map, the *asymmetric map* shown before in Figure 1 and the *symmetric map* (Figure 2).

The only difference is that in Figure 1 we see the age groups as a small bunch of points within the health status categories (asymmetric scaling), whereas in Figure 2 the two sets of points are mixed up with one another, since the spread of both sets of points is the same both horizontally and vertically. Notice from definitions (A.2) and (A.3) in the Appendix that the only difference between principal and standard coordinates is a scaling factor along each principal axis. Figure 2 is generally the map of choice, mainly because it simply looks better, but Figure 1 is perhaps easier to understand because the row and column points occupy the same profile space and are thus easier to interpret jointly. In Figures 1 and 2 we see the age groups lining up from right to left, with a slight arch formed by the middle age categories compared to the extremes. In Figure 1 the health categories lie in positions which can be considered fictitious age groups with responses in only one category; for example, the point "very good" (*muy bueno*) depicts a profile with a percentage of 100% in this category and 0% in the other health categories. Both from the small spread of age groups in the vertical direction and from the small percentage of inertia along this second principal axis (1.5%), we can deduce that any contrast of the health categories vertically is of minor

FIGURE 1: Asymmetric CA map of Table 1

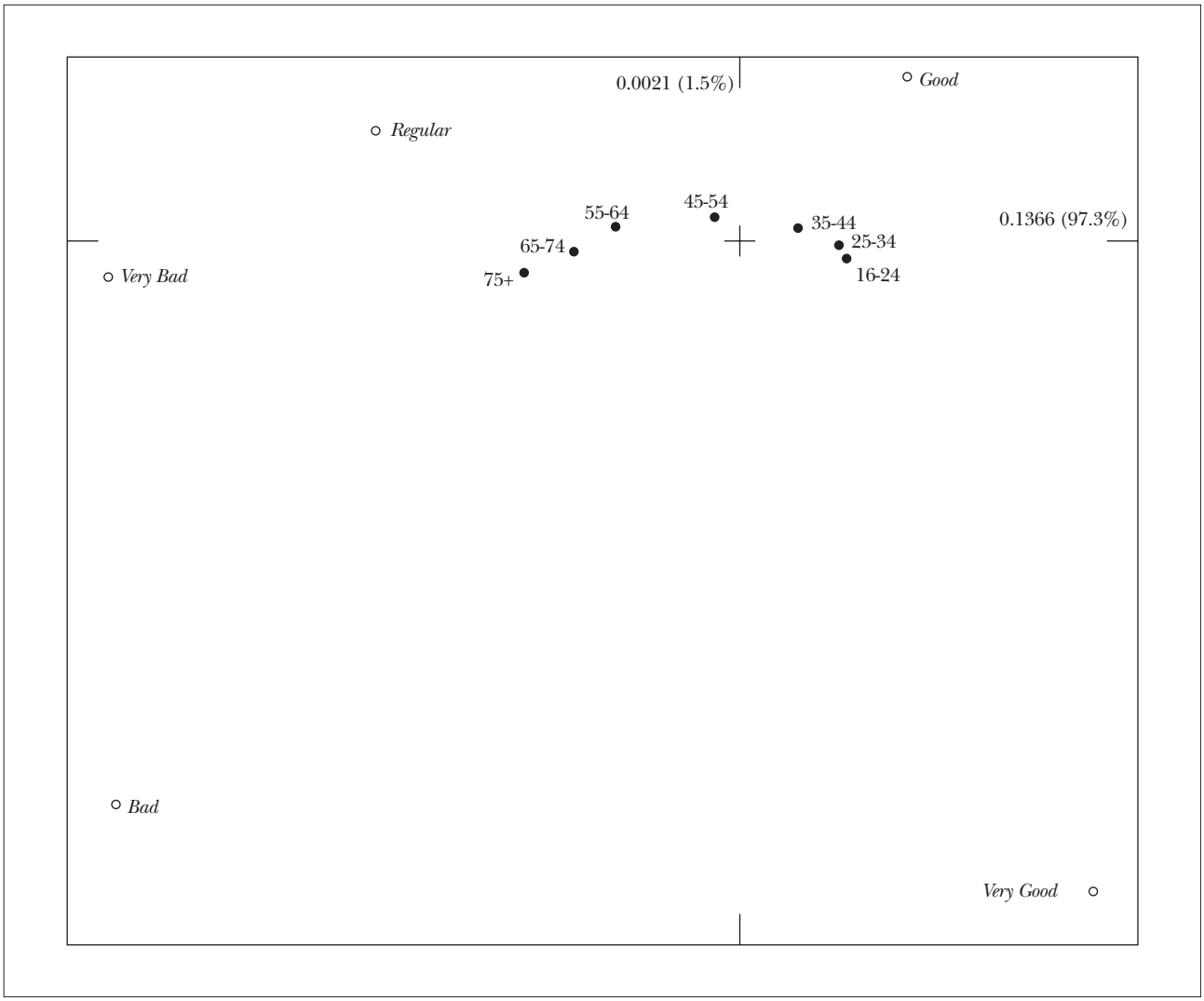
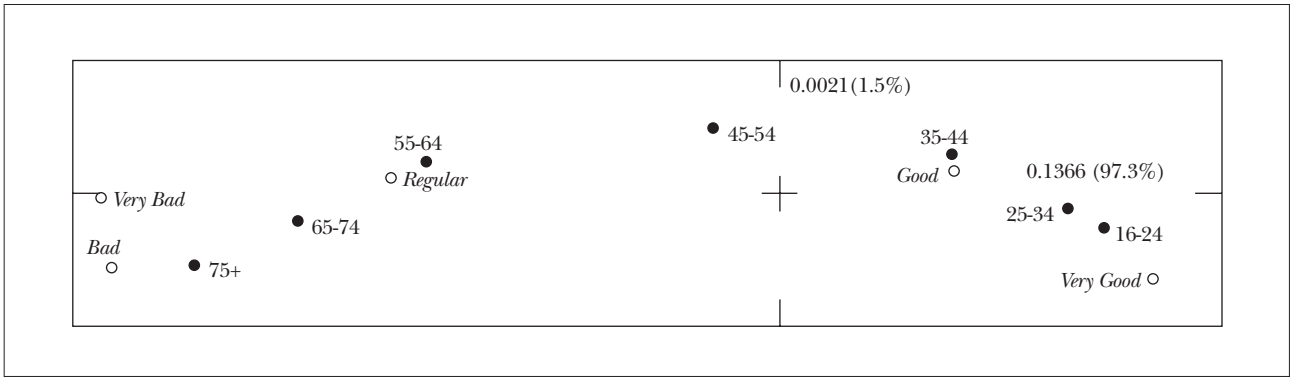


FIGURE 2: Symmetric CA map of Table 1



importance only. The essential information in the original table is captured by the horizontal spread of the points, and the percentage of inertia along this first axis actually puts a figure on the quality of the display in one dimension: 97.3%. Figure 2 tells the same story, showing the unimportance of the second dimension, with the health categories now scaled identically to the age groups along both axes.

What can we conclude from this graphical display? Looking at the right-to-left spread of the age groups, we see that there is only a small change from age group 1 to age group 2, then a larger step to age group 3, an even larger one to age group 4, then the biggest step of all to age group 5, and then smaller steps to group 6 and then group 7. The ordering of health status categories along this dimension agrees with the inherent order, from "very good" (*muy bueno*) to "very bad" (*muy malo*), and their actual relative positions give scale values which can be interpreted; for example, there is very little difference between "bad" (*malo*) and "very bad" (*muy malo*) but a huge difference between, say, "good" (*bueno*) and "regular", when it comes to distinguishing the responses between different age groups.

The health scale values (first principal coordinates) are centred and standardized but can be linearly transformed to any other more meaningful scale; for example, we could transform them to have endpoints equal to 0 and 100, with 0 representing "very bad" and 100 "very good":

Original scale:	–0.767	–0.755	–0.439	0.198	0.423
New scale:	0	1	27.6	81.1	100

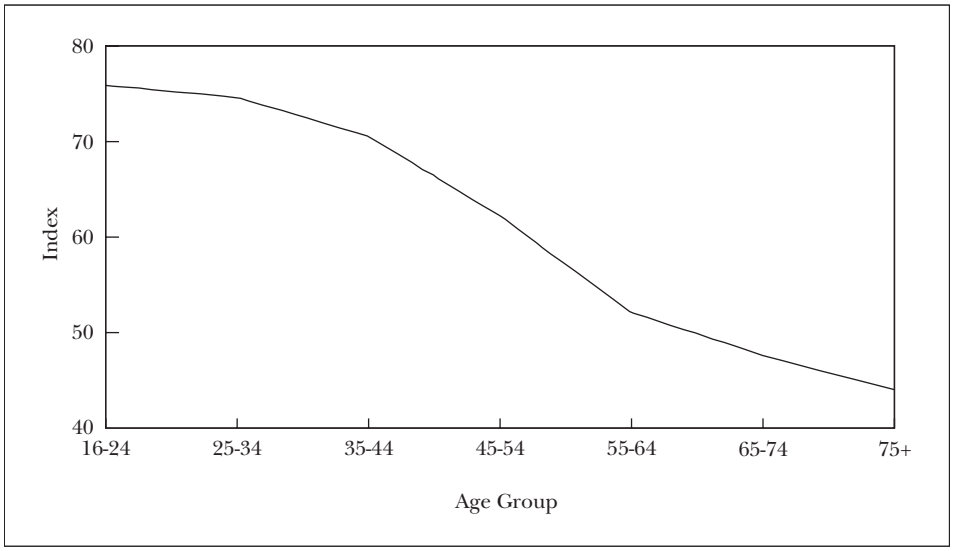
This is a quite different scale from what one would expect if the distances between the categories were equal, in which case the scale values would be 0, 25, 50, 75 and 100. The category "regular" is not in the middle of the scale, but very much towards the "bad" end of the scale, at least in the perceptions of respondents. Or, putting it another way, it is clearly a big step in a negative direction to admit one's health is "regular" as opposed to "good".

Using the above scale values one can establish average values for all those in the age groups:

16-24	75.97
25-34	74.69
35-44	70.63
45-54	62.25
55-64	52.17
65-74	47.67
75+	44.01

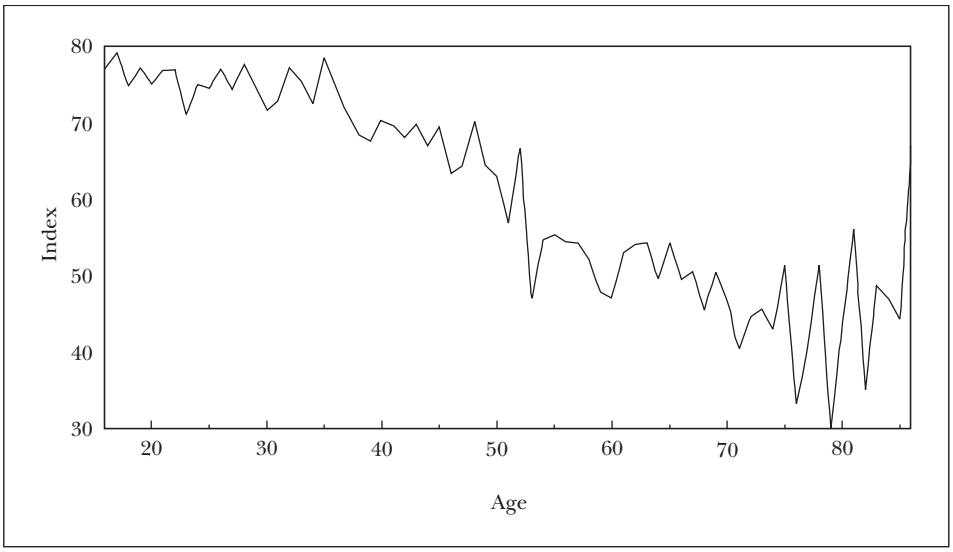
and then plot these against the midpoints of the age categories (Figure 3).

FIGURE 3: Plot of health status index (first dimension of CA) against age group



Since we have the exact ages of each respondent, we can get a more detailed plot by calculating and plotting the averages for each age (Figure 4).

FIGURE 4: Plot of health status index against age (up to 86)



There are some interesting patterns, such as the drop in perceived health in the years just preceeding 30, 40, 50 and 60, usually with a slight recovery in the years after.

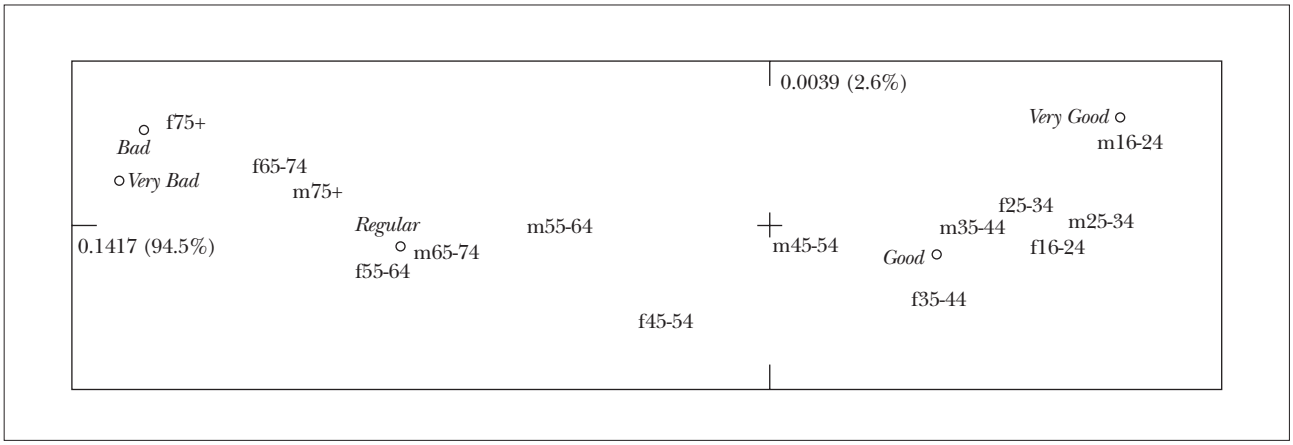
Because of the high sample size in this survey, we can explore the data at least one level further by splitting the age groups according to another variable. "Sex" is the most obvious one, and Table 3 shows the cross-tabulation of the seven age groups split between males and females, with the corresponding health categories.

TABLE 3: Age group and sex interactively cross-tabulated with health status

AGE GROUP	*Very Good*	*Good*	*Regular*	*Bad*	*Very Bad*	SUM
MALES						
16-24	145	402	84	5	3	639
25-34	112	414	74	13	2	615
35-44	80	331	82	24	4	521
45-54	54	231	102	22	6	415
55-64	30	219	119	53	12	433
65-74	18	125	110	35	4	292
75+	9	67	65	25	8	174
FEMALES						
16-24	98	387	83	13	3	584
25-34	108	395	90	22	4	619
35-44	67	327	99	17	4	514
45-54	36	238	134	28	10	446
55-64	23	195	187	53	18	476
65-74	26	142	174	63	16	421
75+	11	69	92	41	9	222
SUM	**817**	**3,542**	**1,495**	**414**	**103**	**6,371**

The symmetric map in Figure 5 shows that females consistently rate themselves as unhealthier than their male counterparts — the female points are always to the left of the male points of the corresponding age group, so that females of 65-74, for example, are rating their health worse than males of 75+.

FIGURE 5: Correspondence analysis of Table 3, symmetric map



4. Other applications to cross-tabulations

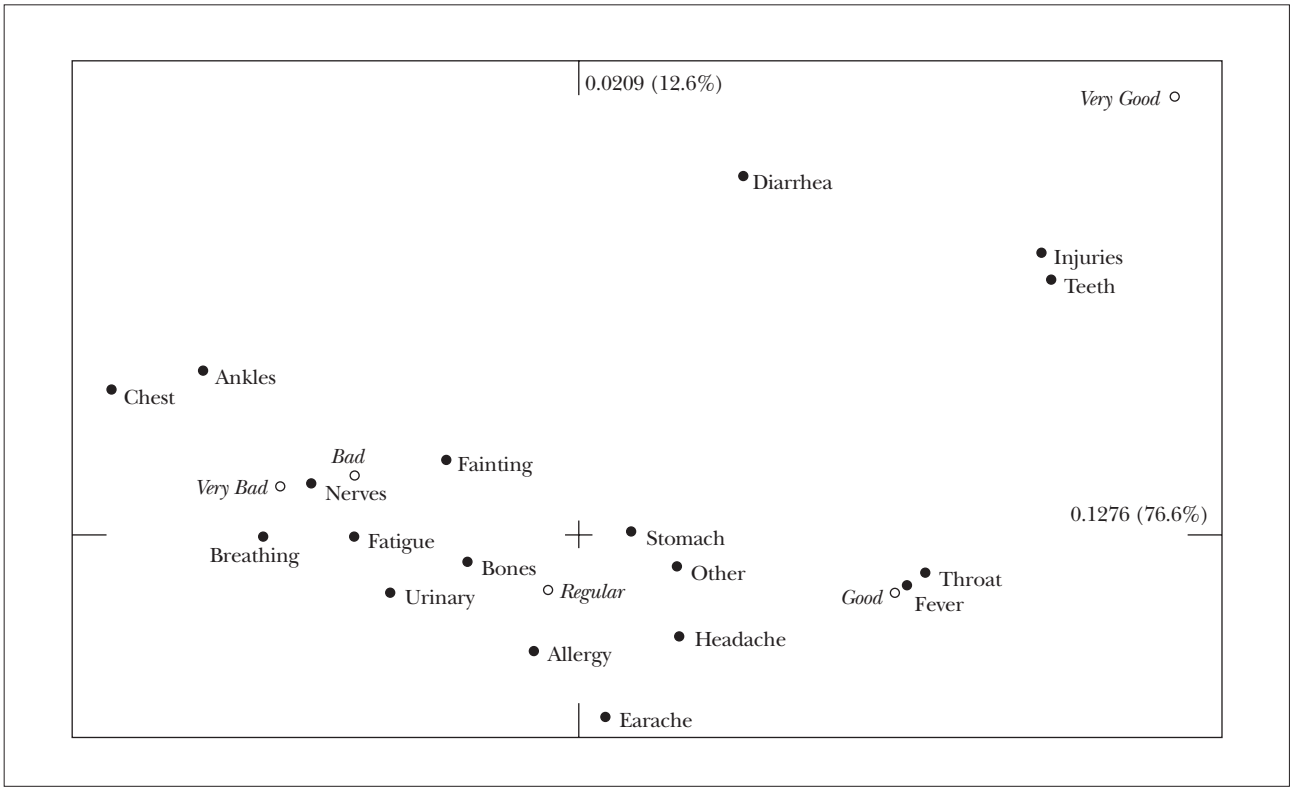
We present several other examples of how CA can be used to explore relationships between several variables. Question 5 of the survey asks respondents if they have had to reduce their normal leisure time activities because of some pain or other symptom. For those that answer "yes", there follows a list of 18 possible symptoms, 17 specific ones and an "other" category. In Table 4 we have tabulated the distributions of the five health status categories for each of the responses to these questions. Notice that the table is not a contingency table as in Tables 1 and 3, since multiple responses are possible to the "Which symptoms?" question.

TABLE 4: Ailments tabulated by perceived health

AILMENT	*Very Good*	*Good*	*Regular*	*Bad*	*Very Bad*	**SUM**
a. Bones, joints	5	64	132	104	30	335
b. Nerves, depression	0	13	24	39	9	85
c. Throat, cough	12	77	62	25	5	181
d. Headache	2	47	41	30	11	131
e. Cuts, injuries	8	21	13	8	2	52
f. Earache	0	4	7	4	0	15
g. Diarrhea	3	6	5	7	2	23
h. Allergies	0	5	8	6	1	20
i. Kidneys, urinary	0	6	12	7	7	32
j. Stomach	2	13	18	13	3	49
k. Fever	3	20	17	6	2	48
l. Teeth	2	5	4	2	0	13
m. Fainting	2	10	21	21	6	60
n. Chest	0	1	10	18	6	35
o. Ankles	1	1	13	15	7	37
p. Suffocation	0	5	27	22	10	64
q. Fatigue	1	9	35	26	10	81
r. Others	5	29	46	20	8	108
SUM	**46**	**336**	**495**	**373**	**119**	**1,369**

Figure 6 shows the symmetric map of this table. Again we find the five health status categories spread along the first principal axis with relative positions similar to those in the previous analyses. The symptoms are thus scaled from left to right in accordance with the associated health status: "chest problems", "ankles", "suffocation, respiratory problems" and psychiatric problems on the "bad" left side, and "teeth", "injuries", "throat" and "fever" on the "good" right side. The second axis is more important in this analysis, and is determined mostly by the status category "very good" and the three symptoms in the upper part of the map: "diarrhea", "injuries" and "teeth". This indicates a subgroup of people who do report problems, but who also tend to report higher than average "very good" health, tending to have one of these afflictions which is just a temporary problem. Notice the position of "diarrhea", which is associated with a mixed group of people: some who view their health at the "very good" end of the scale, and others at the opposite "very bad" end, but with fewer than expected people with "regular" health.

FIGURE 6: Correspondence analysis of Table 4, symmetric map



The next example concerns smoking, and here we have cross-tabulated question 19 about smoking habits with health status (Table 5).

TABLE 5: Smoking categories by perceived health

SMOKING CATEGORY	*Very Good*	*Good*	*Regular*	*Bad*	*Very Bad*	**SUM**
Smoke daily	288	1,309	398	102	18	2,115
Smoke, not daily	31	92	36	7	0	166
Used to smoke	107	519	234	74	20	954
Never smoked	391	1,622	831	228	64	3,136
SUM	**817**	**3,542**	**1,499**	**411**	**102**	**6,371**

The differences between smoking groups with respect to health status categories are not large – this fact can be deduced from the much smaller inertias along the principal axes in Figure 7. The small differences that exist, however, show that those who smoke have a slightly more positive view of their health. As an attempt to explain this finding, we investigated the relationship between smoking and age given in Table 6 and Figure 8. Clearly there is a strong tendency for younger people to smoke, so that the

FIGURE 7: Correspondence analysis of Table 5

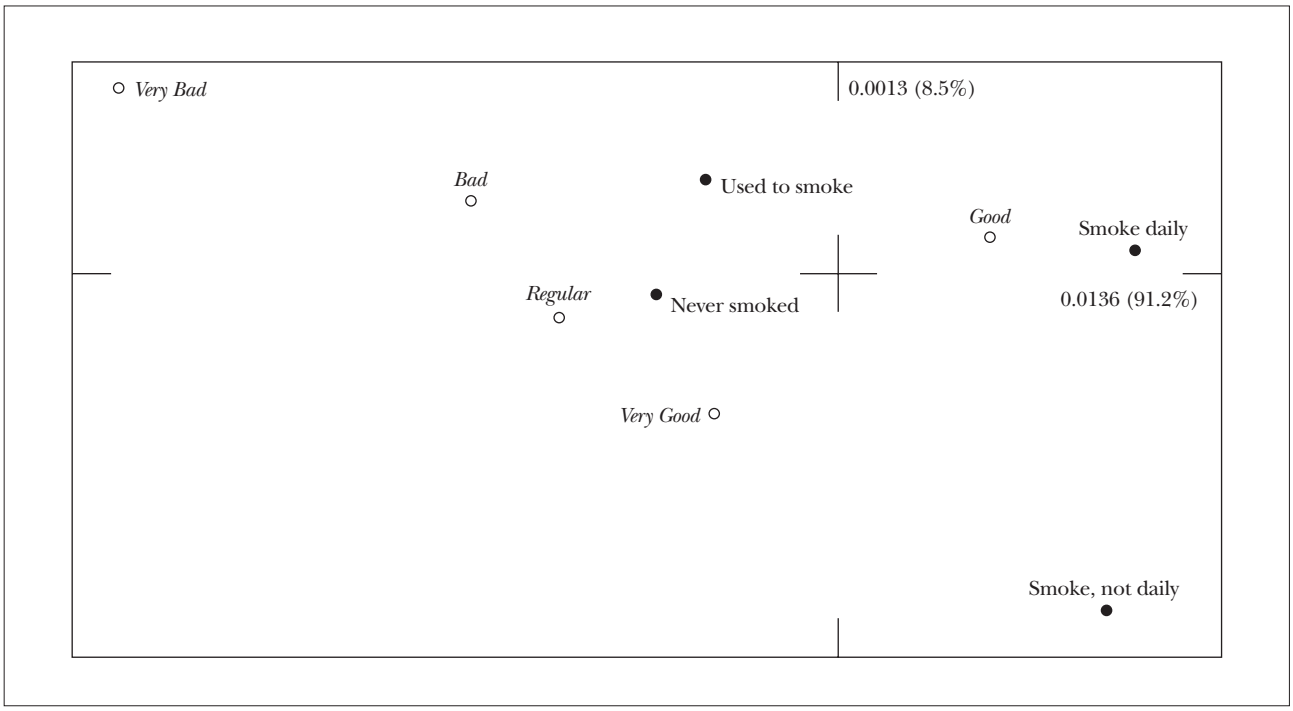
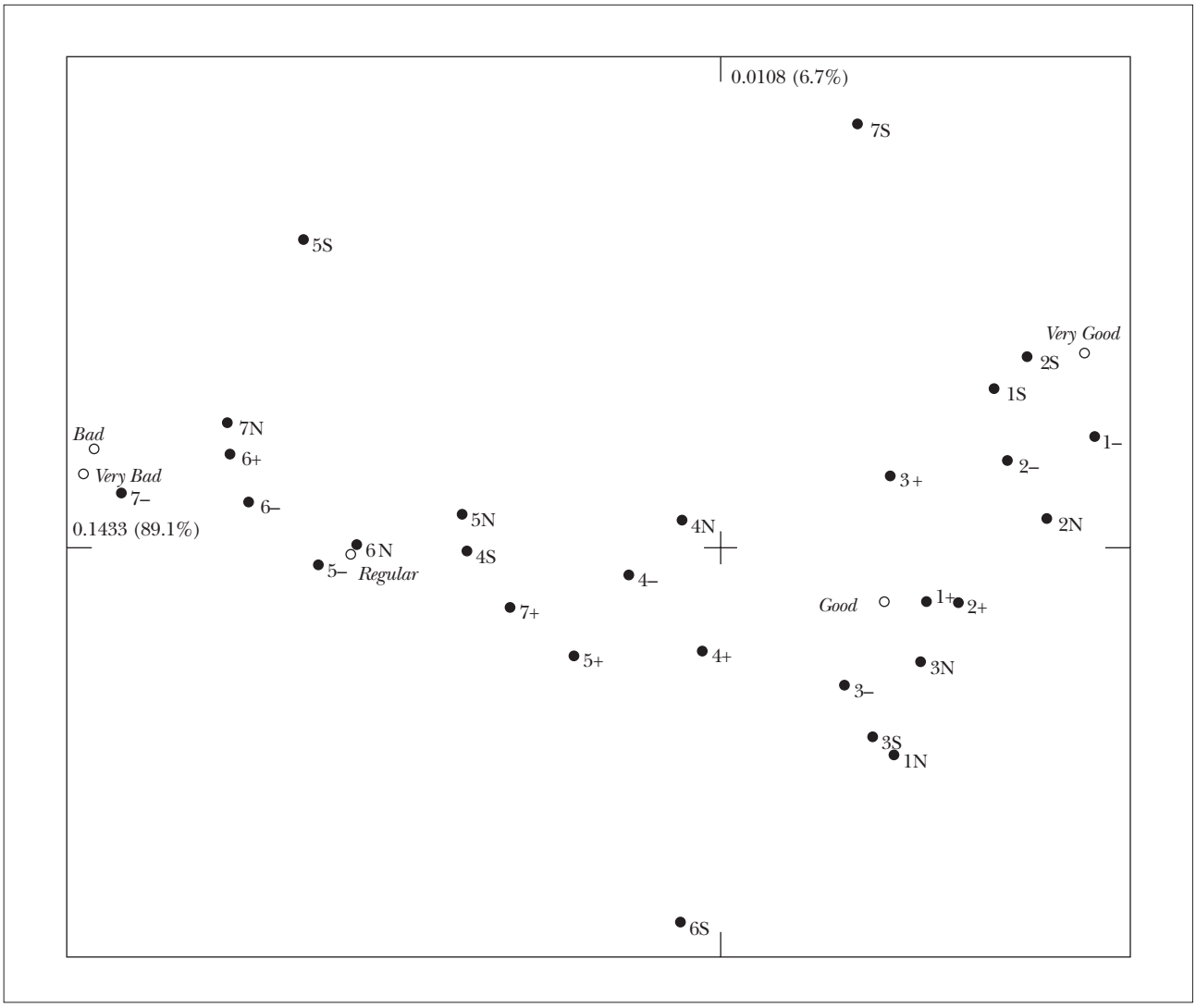


TABLE 6: Age groups and smoking habits interactively cross-tabulated with health status (age indicated by numbers 1 to 7, and smoking group indicated by + (smokes daily), S (smokes but not daily), N (doesn't smoke, but did), – (never smoked); e.g., 2– = age group 2, never smoked)

SMOKING CATEGORY	*Very Good*	*Good*	*Regular*	*Bad*	*Very Bad*	SUM
1+	63	282	71	10	4	430
1S	13	32	8	2	0	55
1N	5	44	10	2	0	61
1–	161	429	78	4	2	674
2+	95	431	90	20	2	638
2S	11	26	5	2	0	44
2N	30	101	18	3	0	152
2–	84	249	51	10	4	398
3+	88	285	80	24	4	481
3S	2	17	4	1	0	24
3N	21	118	27	4	2	172
3–	36	236	70	12	2	356
4+	26	165	76	13	3	283
4S	2	6	9	0	0	17
4N	19	77	36	10	3	145
4–	43	221	115	27	10	416
5+	10	100	46	17	3	176
5S	2	3	5	2	0	12
5N	15	81	50	20	7	173
5–	26	229	204	67	19	545
6+	4	28	24	16	1	73
6S	0	6	3	0	0	9
6N	10	61	58	16	4	149
6–	29	70	199	64	15	477
7+	2	14	9	2	1	28
7S	1	1	1	0	0	3
7N	7	36	34	19	4	100
7–	10	84	112	44	12	262
SUM	**815**	**3,532**	**1,493**	**411**	**102**	**6,353**

finding in Figure 6 can be attributed to the fact that the ex- and non-smoking groups have an older profile with worse perceived health. This leads us to consider the smoking categories within each age group (Table 6), giving a more detailed explanation of the relationship between perceived health and smoking (Figure 8). Here we can see that in the two youngest age groups, non-smokers do indeed feel better about their health than smokers. However, in the older age groups there is a tendency for those that smoke to feel in better health than those who do not.

FIGURE 8: Correspondence analysis of Table 6, symmetric map



5. Using correspondence analysis to develop scales

WE have already seen an example in Section 3 of what is called optimal scaling; the assignment of scale values to a categorical variable with optimal properties. We obtained values for the health status categories which lead to maximum separation, or discrimination, of the age groups. In general, we can use CA to obtain optimal scale values of a set of categorical variables which form a substantively homogeneous group.

For example, question 8a of the health survey asks respondents which of 17 different types of medicines they have taken during the previous two weeks (of the original 18 types, we excluded birth-control pills which only apply to women). More than half of the sample had not taken any medicines, so we excluded them from this analysis. This situation differs from the previous ones, because we are not looking at the relation between medicine consumption and another variable, such as age or smoking. Here we are trying to reduce the dimensionality of a set of variables in much the same way as factor analysis; that is, looking for common factors which capture the relationships between the variables by explaining a maximum amount of variability. The objective is identical to principal component analysis, apart from the fact that the variables are categorical in nature, so the missing link is the quantifications given to the categories.

Multiple correspondence analysis (MCA) – also known as homogeneity analysis (HA) – solves this problem by looking for the category scale values which lead to scores for each respondent which are maximally correlated with each respondent's scale values. To explain this, let us suppose that we make the *ad hoc* decision to assign the scale values 1 to each medicine taken and –0.5 to each medicine not taken, for each of the 16 medicines. Each respondent thus has a set of 16 scale values (which can be considered to form an $N \times 16$ matrix), and we can calculate his or her score by adding up the scale values, giving an

additional column of this notional matrix. Then we can calculate the correlation between the respondent scores and each of the 16 scales, and summarize in some way how well the scores reflect the 16 scales. In MCA this is done by calculating the average squared correlation between the score vector and the 16 scales. The objective of MCA is then to find out which scale values lead to a maximum value of this average squared correlation, so that in this sense the scores maximally explain each of the 16 scales. Once this "factor" has been identified, we proceed to find another set of scale values and associated scores, uncorrelated with the scores already identified, which again maximize the average squared correlation, and so on.

The basic numerical results of the MCA, given for the first three dimensions (i.e., factors) are given in Table 7.

In this table the squared correlations are called "discrimination measures" and the average squared correlation the "eigenvalue". Another way of thinking about the table is that the entries are coefficients of determination (R^2) giving the variance of each variable explained by each dimension (factor). Since the factors are uncorrelated, these R^2 can be added up row-wise to give explained variances by subsets of factors. The dimensions are ordered in

TABLE 7: **Eigenvalues and discrimination measures for each dimension of MCA**

	Dimension		
	1	**2**	**3**
Eigenvalue	.1031	.0815	.0745
Throat, cough	.183	.005	.395
Pain, fever	.127	.038	.537
Vitamins, minerals	.001	.000	.000
Laxatives	.025	.070	.010
Antibiotics	.044	.042	.025
Tranquillisers...	.144	.326	.006
Anti-allergy	.003	.010	.098
Diarrhea	.001	.069	.048
Rheumatism	.084	.002	.024
Heart	.277	.050	.003
Blood pressure	.311	.090	.002
Digestive remedies	.071	.080	.031
Antidepressants	.068	.421	.006
Slimming	.000	.000	.002
Lower cholesterol	.196	.014	.006
Diabetes	.115	.086	.000

TABLE 7 (continued)**: Eigenvalues and discrimination measures for each dimension of MCA**

Medicine	Response		Optimal Scale		
			dim 1	dim 2	dim 3
Throat, cough	*yes*	*1*	–.74	–.13	1.09
	no	*2*	.25	.04	–.36
Pain, fever	*yes*	*1*	–.50	.27	–1.03
	no	*2*	.25	–.14	.52
Vitamins, minerals	*yes*	*1*	–.11	–.06	–.04
	no	*2*	.01	.00	.00
Laxatives	*yes*	*1*	1.01	1.69	.64
	no	*2*	–.02	–.04	–.02
Antibiotics	*yes*	*1*	–.72	.70	.54
	no	*2*	.06	–.06	–.05
Tranquillisers...	*yes*	*1*	.95	1.43	–.19
	no	*2*	–.15	–.23	.03
Anti-allergy	*yes*	*1*	–.28	.49	1.51
	no	*2*	.01	–.02	–.07
Diarrhea	*yes*	*1*	.27	2.74	2.28
	no	*2*	–.00	–.03	–.02
Rheumatism	*yes*	*1*	1.03	.17	–.55
	no	*2*	–.08	–.01	.04
Heart	*yes*	*1*	1.65	–.70	.18
	no	*2*	–.17	.07	–.02
Blood pressure	*yes*	*1*	1.10	–.59	–.08
	no	*2*	–.28	.15	.02
Digestive remedies	*yes*	*1*	.84	.89	.56
	no	*2*	–.08	–.09	–.06
Antidepressants	*yes*	*1*	1.25	3.13	.39
	no	*2*	–.05	–.13	–.02
Slimming	*yes*	*1*	.14	–.48	1.06
	no	*2*	–.00	.00	.00
Control cholesterol	*yes*	*1*	1.86	–.50	.32
	no	*2*	–.11	.03	–.02
Diabetes	*yes*	*1*	1.31	–1.13	.00
	no	*2*	–.09	.08	.00

descending order of "eigenvalue", which is the average squared correlation, the quantity which is maximized by using the corresponding set of scale values given in the second part of the table. The optimum scale values are given, one for each "yes" and "no" response to the 16 types of medicine, for each dimension.

The first factor is a dimension which groups together the following medicines, in order of explained variance: medicines for blood pressure, for the heart, for lowering cholesterol and – to a lesser extent – for diabetes as well as tranquillisers and sleeping pills. It is interesting to note that medicines for minor ailments such as throat infection and flu, pains and fever, and antibiotics, have their signs of the scale values reversed. In other words, people who have been taking the former medicines for chronic health complaints are usually not taking these latter ones for less serious, transient, ailments.

The second factor groups mainly the following medicines: tranquillisers and sleeping pills and antidepressants, in other words the "psychiatric" dimension. Although not so well-explained by this factor, we also note high scale values for diarrhea and laxative medicines.

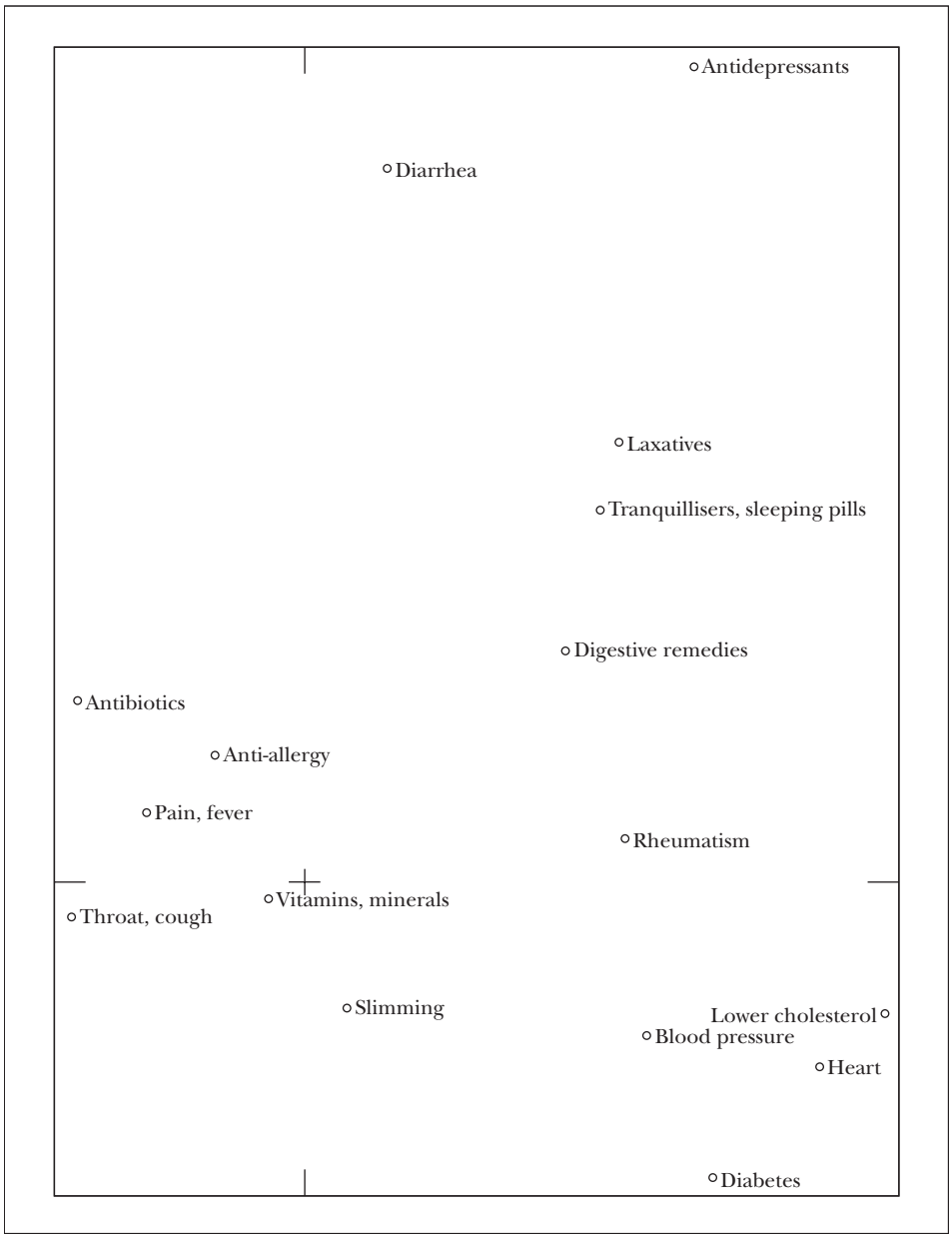
The first two dimensions can be plotted, as before, on a map (Figure 9). This gives an interesting view of the interrelationships between the medicines, with the grouping at bottom right of medicines for chronic diseases, at the top for psychiatric and digestive problems and on the left for more common ailments of a transient nature.

As a complementary analysis to the mapping procedure, we can perform a hierarchical cluster analysis of the 16 types of medicine. Figure 10 shows the cluster tree, based on complete linkage and using the Jaccard index to measure similarity between the medicines. We can see the same clusters as in Figure 9.

In the optimal scaling, we can continue to interpret the factors beyond the second. For example, the third factor separates out the medicines for flu, throat, pains and fever, by themselves. These are the respondents who have had a bacterial or viral infection in the previous two weeks, but are not taking any other medication.

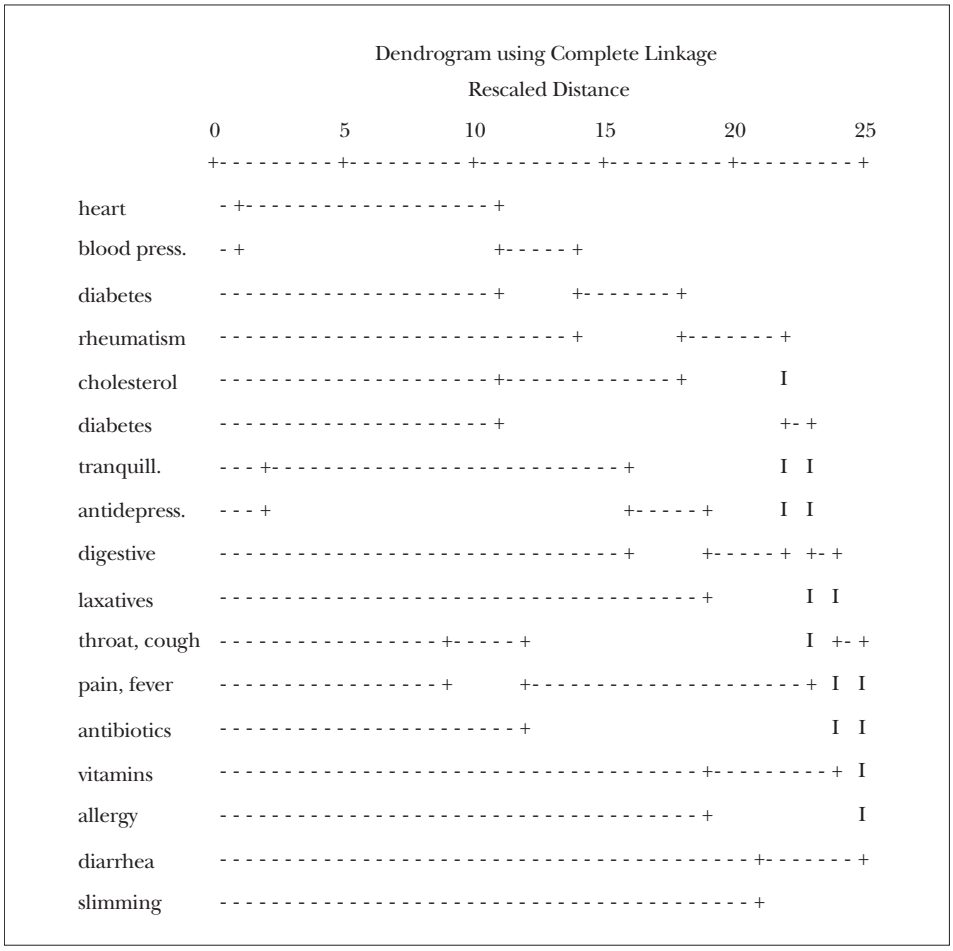
There is one final issue to resolve in this analysis, and it is a controversial one. If we retain two dimensions, or three, or whatever number, how much variance is being explained? This is a simple question in the case of PCA but in MCA it is plagued by difficulties, the main one being what we mean by total variance.

FIGURE 9: Multiple correspondence analysis, showing optimal scale values in two dimensions of "yes" responses to medicine types.



Originally, MCA was defined as the correspondence analysis of a matrix of dummy variables, called an *indicator matrix*. In our present example, where we have 16 variables and two categories of response per variable, the indicator matrix has N rows and 32 columns (where $N =$

FIGURE 10: Hierarchical clustering tree of medicine types



sample size), and contains only zeros and ones, with the ones indicating for each respondent his or her categories of response. The principal inertias of this matrix are exactly the eigenvalues of Table 7, so the way of evaluating variance explained would be to express each eigenvalue as a percentage of the total. The total inertia of an indicator matrix has been shown to be equal to a constant: $(J - Q)/Q$, where J = the total number of categories of response, and Q = the number of questions, i.e. in this example $(32 - 16)/16 = 1$. So the eigenvalues in this particular case are also the proportions of inertia, with the first three dimensions thus accounting for 10.13, 8.15 and 7.45% of the inertia. These percentages are very small, and give a pessimistic view of the value of the analysis.

A second way of defining MCA is to perform a CA not on the indicator matrix, usually denoted by Z, but on the so-called *Burt matrix,* B = Z^TZ. This is the super-matrix of all two-way cross-tabulations of the set of variables. It is well known that this CA leads to the same standard coordinates as before, but with principal inertias equal to the squares of those for the indicator matrix, so that the percentages of inertia are calculated on the squared eigenvalues. We calculated the sum of squared eigenvalues (there are 16 in total) to be 0.06609, so that the proportions of inertia explained by the first three dimensions are now computed as: $0.1031^2/0.06609$, $0.0815^2/.06609$, $0.0745^2/0.06609$, giving percentages of 16.1, 10.1 and 8.4% respectively. These look more optimistic than before, but they are actually still too low. This is explained by Greenacre (1989), who pointed out that neither of these ways of calculating the percentages of inertia have the simple two-variable situation described in Section 3 as a special case.

A more realistic alternative, which agrees with the two-variable case of simple CA, is to ascertain how well a solution is able to reconstruct the two-way association pattern of the variables. Greenacre (1993) explains how a simple calculation using the eigenvalues can give us this alternative measure. First, we adjust the total inertia of B as follows, to obtain the average inertia of all the two-way tables between all pairs of the 16 variables:

$$\text{average inertia} = \frac{Q}{Q-1} \times \left(\text{inertia of B} - \frac{J-Q}{Q^2} \right)$$

that is:

$$\frac{16}{15}\left(0.06609 - \frac{16}{16^2} \right) = 0.003830$$

(The difference between the previous total of 0.06609 and this one of 0.003830 is that part of the Burt matrix which we actually do not need to explain at all, and which creates the problem in the percentage calculation.) Second, we adjust the eigenvalues themselves, as follows:

$$\left(\frac{Q}{Q-1} \right)^2 \times \left(\text{eigenvalue} - \frac{1}{Q} \right)^2$$

for example:

$$\left(\frac{16}{15}\right)^2 \times (0.1031 - 0.0625)^2 = 0.001875$$

and express this as a percentage of the average inertia 0.003830, giving 49.0%. In the case of the second and third inertias, we similarly obtain percentages of 10.7 and 5.0% respectively. These percentages of inertia are more realistic reflections of the variance explained, and have more justification than the usual approaches.

We can thus conclude that the two-dimensional map of Figure 9 explains at least 59.7% of the total inertia in the 16 variables.

6. Exploring missing data

CA is frequently used to explore patterns of missing data in a survey, and to answer questions such as: is there a specific group of respondents tending to not answer questions? Or is non-response "correlated" between variables, i.e. can we say that certain groups of variables tend to have non-responses simultaneously? A way to answer these questions would be to set up a data matrix of binary information, where for each respondent we simply code whether the respondent has replied or not, using a one for a missing response and a zero for an actual response, whatever that may be. We would code the data this way because we are interested more in the occurrence of a non-response than a response, but if we wished to treat these two possibilities equally we would use the coding in MCA and introduce two columns for each variable; a dummy variable for non-response and a dummy variable for response. Thus, for N respondents and Q questions under investigation, the matrix would either be of order $N \times Q$ or $N \times 2Q$. The CA of these matrices will give an idea of which questions have non-responses by the same people, and also which respondents are associated with which non-responses.

In this particular survey, the level of non-response is very low, so such questions can not be investigated: but there is one variable – "Income" – which raises an interesting issue. There are 1382 non-responses to the question on income (denoted by I?), and we can use CA to investigate the relationship between this question, including categories of response and non-response, and other biographical variables which are answered by almost all the respondents. Income was thus cross-tabulated with the following variables: sex, marital status, level of schooling, work situation, breadwinner or not, and work situation of head of family. Although these are separate cross-tabulations, the fact that they have one question in common allows us to stack the tables one on top of each other (Table 8). CA of this set of tables will show as best as possible the relationship of each question with income, and we will be

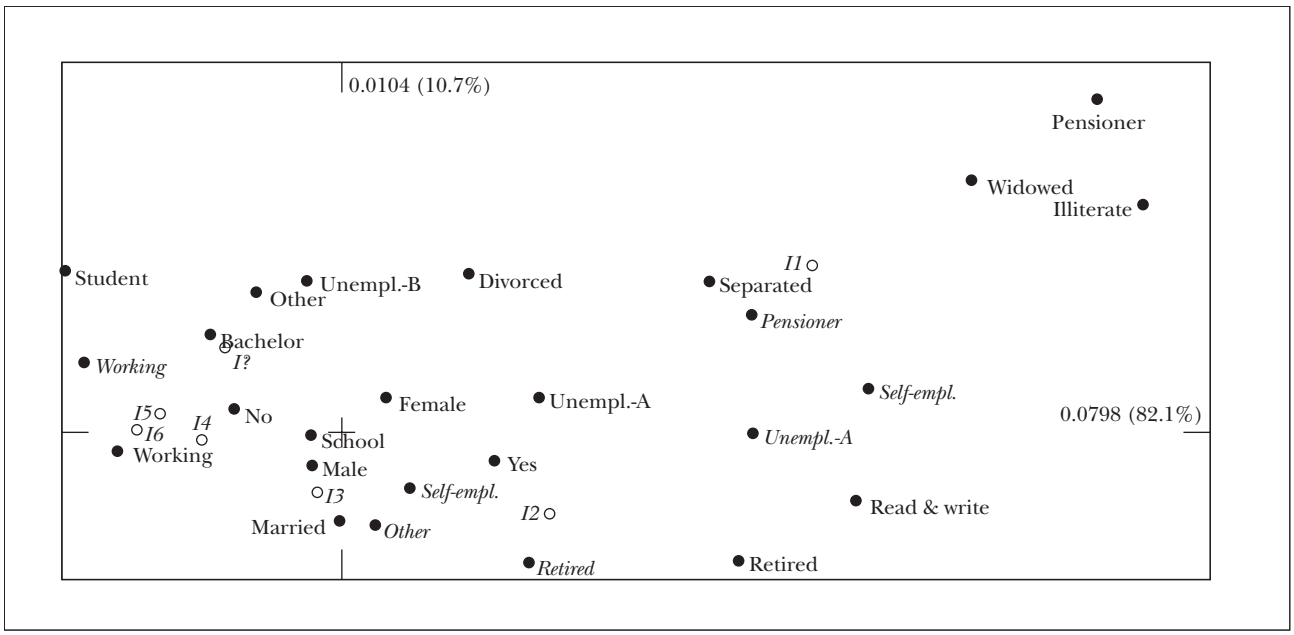
TABLE 8: Income categories and "non-response" cross-tabulated with biographical variables

	Income groups and missing category (I?)							
	I1	**I2**	**I3**	**I4**	**I5**	**I6**	**I?**	**sum**
Male	177	644	711	454	262	185	651	3,084
Female	300	684	728	433	250	156	731	3,282
Bachelor	104	294	410	305	204	123	602	2,042
Married	207	844	941	543	292	208	680	3,715
Separated	21	24	23	7	4	2	17	98
Divorced	8	5	10	9	3	3	5	43
Widowed	137	161	55	22	9	5	73	462
Illiterate	59	66	26	2	1	0	16	170
Read & write	35	86	37	7	5	0	26	196
School	383	1174	1376	878	506	341	1336	5,994
Working	49	304	559	466	298	219	521	2,416
Retired	143	394	228	60	27	22	120	994
Pensioner	92	95	24	8	4	1	38	262
Unemployed-A	84	140	153	71	35	11	108	602
Unemployed-B	9	19	23	8	9	8	37	113
Student	9	64	129	116	71	42	268	699
Self-employed	86	303	318	150	65	34	272	1,228
Other	3	8	5	8	3	4	13	44
Head of household: yes	323	711	658	373	184	125	434	2,808
Head of household: no	150	609	774	514	324	216	930	3,517
Working	43	233	552	424	270	190	722	2,434
Retired	63	287	179	72	48	23	151	823
Pensioner	17	24	17	8	1	0	15	82
Unemployed-A	24	58	24	9	4	2	27	148
Self-employed	1	2	1	0	0	0	1	5
Other	0	3	0	1	1	0	2	7

especially interested in the position of the income non-response category (I?).

Figure 11 shows the resulting map. The income categories, labelled I1 to I6 in the map, lie in their expected order, with the lowest income on the right and the highest income on the left (notice that we can change the sign of all the coordinates on the first axis so that higher income is on the right; this makes no difference to the CA results). It is interesting to see how the other categories are scaled from right to left in terms of their income profiles, from "illiterate", "pensioner" and "widowed" on the right to "head of household working", "working" and "student" on the left. The income non-response point (I? in the map) lies well to the higher income

FIGURE 11: Correspondence analysis of Table 8; the job status in italics refers to that of the head of household (last part of Table 8)



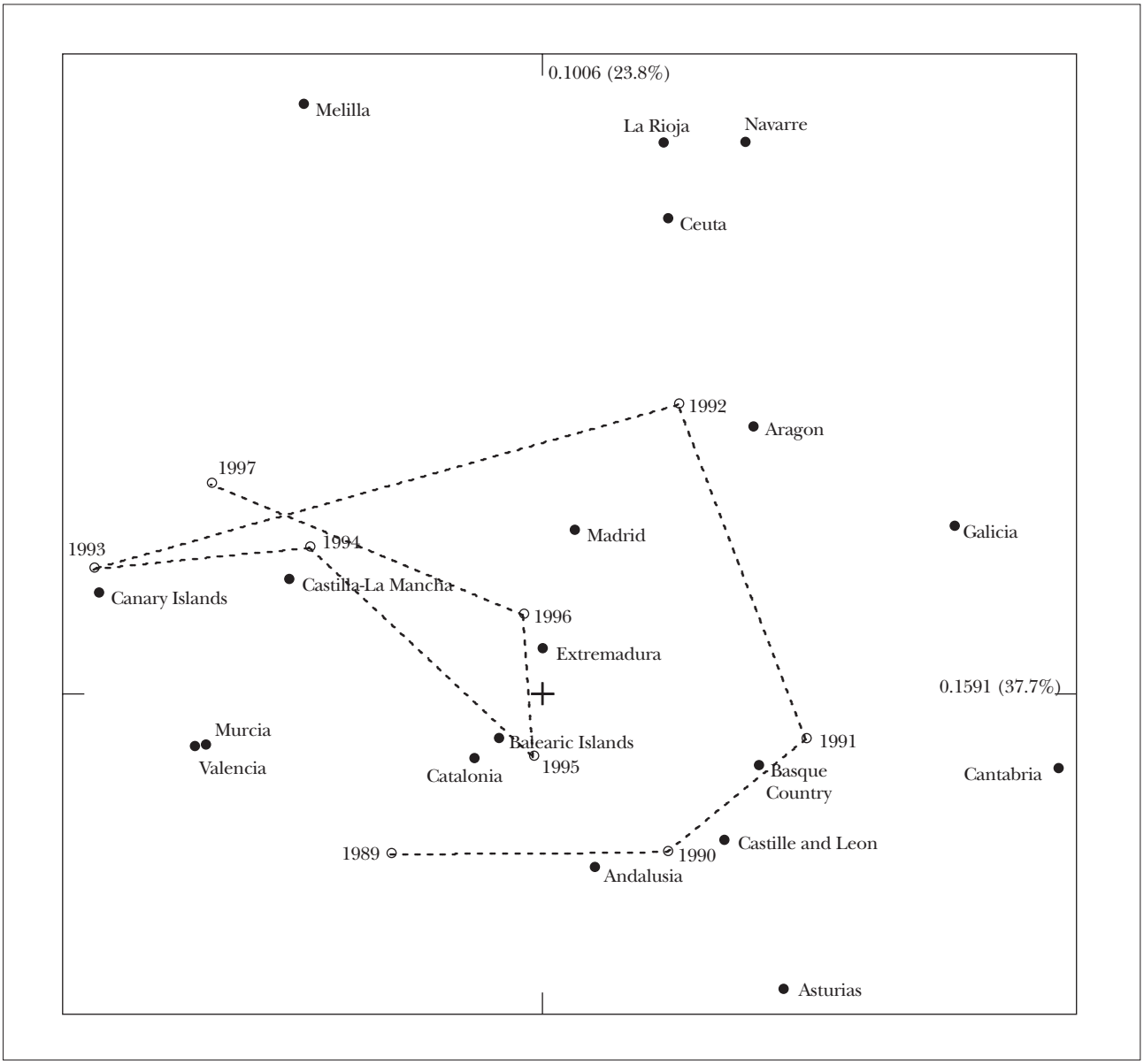
side, just below response 4 (150,000-200,000 pts./month) with respect to the first axis. This is an informal estimate of the position of this group with respect to the other income groups. But it should be remembered that this is an average position of the non-respondents, not a specific income group; and there is likely to be a high spread of incomes within the group. A more formal way of estimating the income in this group of non-respondents would be to set up a model at the individual respondent level of income group related to biographical variables, then estimate the most likely group for each non-respondent.

7. Visualizing trends

THE usual way to display trends is in the form of a line plot with the horizontal axis depicting the time line and the vertical axis depicting the variable which is being observed over time. Thus in Figure 5.1.1 of Regidor and Gutiérrez-Fisac (1999), the number of cases of measles reported in Spain is plotted over the years 1989 to 1997. But in the table on which this figure is based, Table 5.1.2 on page 95 of this publication, the reported cases for each autonomous region in Spain are given for each year, 19 regions in all. To visualize and compare these trends would be difficult since we would have to make 19 different line plots and then try to compare them amongst one other and with the overall trend given in Figure 5.1.1. CA can be used to interpret the different trends in the autonomous regions. The symmetric map of Table 5.1.2 of Regidor and Gutiérrez-Fisac (1999) is given in Figure 12. Notice that in this figure the centre of the display corresponds to the trend of the whole country, or average row profile. Thus a complete trend line is reduced to a point, and the points representing the autonomous regions will show how each region deviates from this overall pattern, with the year points facilitating the interpretation of these deviations.

First, notice the trajectory traced out by the nine consecutive years. A circle is traced out from years 1989 to 1993, then the years move towards the centre of the map (1994 to 1996) and then 1997 returns to a position near 1993 and 1994. The most outlying autonomous regions are those that show the greatest deviation from the average: Asturias in the initial years has more than average incidence, Cantabria in 1991, to Galicia, Aragon and then the group formed by Ceuta, La Rioja, Navarre and Melilla in 1992, and the Canary Islands in 1993. Regions near the centre such as the Balearic Islands and Extremadura do not differ as much from the average trend.

FIGURE 12: Correspondence analysis of measles trend data



8. Conclusions

IN this working paper we have tried to give a comprehensive overview of how correspondence analysis can assist in deciphering the complex information contained in a national health survey. From a simple cross-tabulation to a multiway table and a set of intercorrelated categorical variables, correspondence analysis provides a medium for exposing patterns in the data and suggesting hypotheses. It also facilitates the quantification of categorical data, which can assist with the model-building process. Optimal scales can be defined which capture a maximum percentage of variation and condense the data at the same time, and these scales can be used in other analyses which require interval scales. The method also allows investigation of missing data, which is a categorical item of information, and provides an alternative method for plotting trend data as a movement between points in multidimensional space.

Appendix: Correspondence analysis theory

1. Let N be the $I \times J$ table with grand total n and let P = $(1/n)$N be the *correspondence matrix,* with grand total equal to 1.
2. Let r and c be the vectors of row and column sums of P respectively and D_r and D_c the diagonal matrices with r and c on the diagonal.
3. Compute the singular value decomposition (SVD) of the centred and standardized matrix with general element $(p_{ij} - r_i c_j)/\sqrt{r_i c_j}$:

$$D_r^{-1/2}(P - rc^T)D_c^{-1/2} = UD_\alpha V^T \qquad (A.1)$$

   where the singular values are in descending order: $\alpha_1 \geq \alpha_2 \geq \ldots$
4. Compute the *standard coordinates* X and Y:

$$X = D_r^{-1/2}U \qquad Y = D_c^{-1/2}V \qquad (A.2)$$

   and *principal coordinates* F and G:

$$F = XD_\alpha \qquad G = YD_\alpha \qquad (A.3)$$

   Notice the following:

- The results of CA are in the form of a map of points representing the rows and columns with respect to a selected pair of principal axes, corresponding to pairs of columns of the coordinate matrices – usually the first two columns for the first two principal axes. The choice between principal and standard coordinates is described below.
- The total variance, called *inertia,* is equal to the sum of squares of the matrix decomposed in (A.1):

$$\sum_i \sum_j (p_{ij} - r_i c_j)^2 / (r_i c_j) \qquad (A.4)$$

  which is equal to the Pearson chi-squared statistic calculated on the original table divided by n.

- The squared singular values α_1^2, α_2^2, ..., called the *principal inertias,* decompose the inertia into parts attributable to the respective principal axes, just as in PCA the total variance is decomposed along principal axes.
- The most popular type of map, called the *symmetric map,* uses the first two columns of F for the row coordinates and the first two columns of G for the column coordinates, that is both in principal coordinates as given by (A.3).
- An alternative scaling, which has a more coherent geometric interpretation, but less aesthetic appearance, is the *asymmetric map;* for example, rows in principal coordinates F and columns in standard coordinates Y in (A.2) (or vice versa). The choice between a row-principal or column-principal asymmetric map is governed by whether the original table is considered as a set of rows or a set of columns, respectively, when expressed in percentage form.
- The positions of the rows and the columns in a map are projections of points, called *profiles,* from their true positions in high-dimensional space onto a best-fitting lower-dimensional space. A row or column profile is the corresponding row or column of the table divided by its respective total – in the case of a contingency table the profile is a conditional frequency distribution. Each profile is weighted by a *mass* equal to the value of the corresponding row or column margin, r_i or c_j. The space of the profiles is structured by a weighted Euclidean distance function called the *chi-square distance* and the optimal map is obtained by fitting a lower-dimensional space which fits the profiles by weighted least-squares.
- Equivalent forms of (A.4) which show the use of profile, mass and chi-square distance are:

$$\sum_i r_i \sum_j \left(\frac{p_{ij}}{r_i} - c_j \right)^2 / c_j = \sum_j c_j \sum_i \left(\frac{p_{ij}}{c_j} - r_i \right)^2 / r_i \qquad \text{(A.5)}$$

  Thus the inertia is a weighted average squared distance between the profile vectors (e.g., $\frac{p_{ij}}{r_i}$, $j = 1, ..., J$, for a row profile, weighted by the mass r_i) and their respective average (e.g., c_j, $j = 1, ..., J$, the average row profile), where the distance is of a weighted Euclidean form (e.g., with inverse weighting of the j-th term by c_j).

- An equivalent definition of CA is as a pair of classical scaling problems, one for the rows and one for the columns. For example, a square symmetric matrix of chi-square distances can be calculated between the row profiles, with each point weighted by its respective row mass. Applying classical scaling (also known as principal coordinate analysis) to this distance matrix, and taking the row masses into account, leads to the row principal coordinates in CA.
- We can write the SVD in (A.1) in terms of the standard coordinates in the following equivalent form, for the (i, j)-th element:

$$p_{ij} - r_i c_j = r_i c_j \left(1 + \sum_k \alpha_k x_{ik} y_{ik}\right) \qquad \text{(A.6)}$$

which shows that CA can be considered as a bilinear model (see chapter 6 by van der Heijden, Mooijaart and Takane in Greenacre and Blasius, 1994). For any particular solution, for example in two dimensions where the first two terms of this decomposition are retained, the residual elements have been minimized by weighted least squares.

Bibliography *

BENZÉCRI, J.-P. (1973): *Analyse des Données;* tome I: *Analyse des Correspondances,* tome II: *La Classification,* Paris, Dunod.

BLASIUS, J. and M. J. GREENACRE (1998): *Visualization of Categorical Data,* San Diego, Academic Press.

GIFI, A. (1990): *Nonlinear Multivariate Analysis,* Chichester, Wiley.

GREENACRE, M. J. (1984): *Theory and Applications of Correspondence Analysis,* London, Academic Press.

— (1989): "The Carroll-Green-Schaffer scaling in correspondence analysis: a theoretical and empirical appraisal", *Journal of Marketing Research,* 26, pp. 258-365.

— (1993): *Correspondence Analysis in Practice,* London, Academic Press.

— and J. BLASIUS (1994): *Correspondence Analysis in the Social Sciences,* London, Academic Press.

LEBART, L., A. MORINEAU and K. WARWICK (1984): *Multivariate Descriptive Statistical Analysis,* Chichester, Wiley.

NISHISATO, S. (1980): *Analysis of Categorical Data: Dual Scaling and its Applications,* Toronto, University of Toronto Press.

REGIDOR, E. and J. L. GUTIÉRREZ-FISAC (1999): *Indicadores de Salud. Cuarta Evaluación en España del Programa Regional Europeo Salud para Todos,* Madrid, Ministerio de Sanidad y Consumo.

VAN DER HEIJDEN, P., A. MOOIJAART and Y. TAKANE (1994): "Correspondence analysis and contingency table models", in M. Greenacre and J. Blasius (eds.): *Correspondence Analysis in the Social Sciences,* chapter 6, pp. 79-111, London, Academic Press.

* The most up-to-date and comprehensive references on correspondence analysis are the edited volumes by Greenacre and Blasius (1994) and Blasius and Greenacre (1998), containing both theoretical material as well as sociological applications.

ABOUT THE AUTHOR

MICHAEL GREENACRE, formerly Professor of Statistics at the University of South Africa, is presently a permanent foreign professor in the Department of Economics and Business at Pompeu Fabra University in Barcelona. His doctoral studies were at the University of Paris where he studied under Jean-Paul Benzécri, the originator of correspondence analysis. Greenacre has written two books on correspondence analysis and edited two other books on the topic, as well as giving short courses to audiences of marketing researchers, economists and environmental scientists in several countries, notably the USA, UK, Germany, Finland, Spain, Italy, Switzerland, South Africa and Norway.

Fundación **BBVA**

DOCUMENTOS DE TRABAJO

NÚMEROS PUBLICADOS

DT 01/02 *Trampa del desempleo y educación: un análisis de las relaciones entre los efectos desincentivadores de las prestaciones en el Estado del Bienestar y la educación*
Jorge Calero Martínez y Mónica Madrigal Bajo

DT 02/02 *Un instrumento de contratación externa: los vales o cheques. Análisis teórico y evidencias empíricas*
Ivan Planas Miret

DT 03/02 *Financiación capitativa, articulación entre niveles asistenciales y descentralización de las organizaciones sanitarias*
Vicente Ortún-Rubio y Guillem López-Casasnovas

DT 04/02 *La reforma del IRPF y los determinantes de la oferta laboral en la familia española*
Santiago Álvarez García y Juan Prieto Rodríguez

Documentos de Trabajo 5

Fundación BBVA

Sede Social
Plaza de San Nicolás, 4
48005 Bilbao

Sede en Bilbao
Gran Vía, 12
48001 Bilbao
Tel.: 94 487 52 52
Fax.: 94 424 46 21

Sede en Madrid
Paseo de Recoletos, 10
28001 Madrid
Tel.: 91 374 54 00
Fax.: 91 374 85 22

informacion@fbbva.es
www.fbbva.es