ISSN 1748-7595 (Online)

Kent Business School

Working Paper Series

Interpreting a Data Base of Railway Workers using Optimal Scaling Techniques

Fabiola Portillo
Universidad de la Rioja, Spain

Cecilio Mar Molinero
Kent Business School

Tomas Martinez Vara
Universidad Complutense de Madrid, Spain

KBS
UNIVERSITY OF KENT

Working Paper No.127
Jun 06

INTERPRETING A DATA BASE OF RAILWAY WORKERS USING OPTIMAL SCALING TECHNIQUES

By

Fabiola Portillo
Departamento de Economia y Empresa
Universidad de la Rioja, Spain

Cecilio Mar Molinero*
Kent Business School
University of Kent, UK
C.Mar-Molinero@kent.ac.uk

Tomas Martinez Vara
Escuela de Ciencias Empresariales
Universidad Complutense de Madrid, Spain

* corresponding autor.

This version June 2006.

ABSTRACT

Optimal scaling techniques, in particular Categorical Principal Components, are used in order to interpret the information contained in a database of railway workers. The data set consists of eight variables, three quantitative and five qualitative, measured on 527 workers who joined the Spanish railway company MZA during the period 1882 to 1885. The analysis revealed that workers whose place of birth was not Spain tended to be employed in more senior jobs and were paid higher salaries than workers whose place of birth was Spain. It also revealed that most workers who left the company in an abnormal way (redundancy, or disciplinary dismissal) did so not long after they had joined. It was also found that the reason for leaving was unrelated to both first salary and seniority at the time of joining.

Keywords: economic history, optimal scaling, categorical principal components, history of railways.

INTRODUCTION

Traditional statistical analysis techniques assume that all the data is quantitative, or in technical terms, that all measurements are taken in an interval or a ratio scale; Stevens (1951). This is seldom the case. It is quite common for data to be qualitative; i.e., measured on nominal or ordinal scales. Although the techniques that are applied in the statistical analysis of numerical data are well known and studied as part of the core programme of any Social Sciences faculty, the methods that can be used to the analysis of qualitative, or mixed qualitative/quantitative data are less well known.

In this paper we analyse the information contained in a database of railway workers who were employed by the MZA (Madrid Zaragoza Alicante) railway company, one of the two large railway companies in Spain during the second half of the 19th Century and the first half of the 20th. MZA was established in Madrid in 1856 and continued in existence until 1941 when it was nationalised and made part of the Spanish national railway network, RENFE. In 1889, the company owned 2041 Kms of railways.

For each of the employees of the company we have information on the age at which they entered employment, the number of years they worked for the company, and the initial salary. All three of them are quantitative. We also have qualitative information: marital status, type of job performed, reasons for leaving employment, and place of birth. All these are of a qualitative nature. We could ask questions about the relationship between pairs of variables; for example, how is the initial salary related to the age at which the employee joined the company. But analysing the relationships that exist between pairs of variables ignores the multivariate nature of

the data, and to fully exploit this multivariate nature we need to be able to work with both qualitative and quantitative variables.

The approach that has been followed in this paper is to transform qualitative values into quantitative using the technique of optimal scaling, which can be traced to RA Fisher (1938). Once the quantification of qualitative variables has been achieved, we can apply the standard methods of multivariate analysis. In this case, we have used principal component analysis because it allows the visualisation of the results and, by so doing, makes them accessible to the non-specialist. All the calculations have been performed with the computer package SPSS, version 14.

It will be shown that that workers whose place of birth was not Spain tended to be employed in more senior jobs and were paid higher salaries than workers whose place of birth was Spain. It was also found that most workers who left the company in an abnormal way (redundancy, or disciplinary dismissal) did so not long after they had joined, and that the reason for leaving was unrelated to both first salary and seniority at the time of joining.

.

The paper starts with a description of the data set. This is followed by an introduction to the technical apparatus used. The results of the analysis are shown next. An interpretation and discussion section follows. The paper ends with a concluding section.

THE DATA

The data was collected from the personal files of 533 workers who joined the Spanish railway company MZA during the period 1882 to 1885 and for whom there was no missing information- personal files existed for 940 individuals. This information was originally hand written, and was coded into an ACCESS file. Some errors were made in the process of transferring it to electronic form. When an obvious error was identified, the worker was excluded from the data set. To give just one example, there was a worker whose data of birth was recorded as 1884 but who was said to have started working in the company in 1882, two years before birth! After such obviously erroneous cases were removed, the total number of workers left was 527.

The gender of the worker could be inferred from the name, but there were only 2 females in the sample, and gender was not included in the analysis.

Besides name and surname, the database gave date and place of birth, job that the worker undertook when first employed, initial salary, marital status, year in which the worker left employment, and cause for leaving the company (which included death). This information was recoded into eight variables: age when the worker first joined the company, number of years that the worker stayed in the company, initial salary, place of birth, marital status, first job performed, section in which they performed their activity, and reasons for leaving the company. The first three variables are quantitative, while the last five are qualitative.

Place of birth was recorded in the form of four categories: born in Madrid, born elsewhere in Spain, born in France, and others. Madrid was separated from the rest of Spain because the largest group of workers was from Madrid. Separating Madrid from the rest of Spain made it possible to check if workers from Madrid were engaged in different activities than workers from the rest of Spain, perhaps because they were employed in workshops or offices located in Madrid, the place where the headquarters of the company were located. MZA was set up with French capital, and 17 workers in the sample had been born in France, the largest non-Spanish contingent. It was felt that French employees might be skilled workers brought to Spain for their specialised knowledge, an issue to be explored in the analysis. "Born elsewhere" encompassed workers with a variety of birthplaces, from Cuba to Great Britain.

Marital status contained only three categories: single, married, widower. There was no missing data.

The name of first job performed was carefully recorded in the original personal files, and the listing of such jobs is certainly entertaining; sometimes even funny. These jobs were recoded into three categories: apprentices or similar, skilled and established working positions, and supervisors. It is clearly not the same thing to be a skilled carpenter or a skilled mechanic, but it was felt that this difference would be captured by the section in which the work was performed. This variable is clearly of an ordinal nature, since it is better to be a supervisor than an established ordinary worker, and better to be an established ordinary worker than an apprentice.

The data named twelve sections: boiler making, forge, foundry, fitting, assembly, lathe, offices, plumbing and electric installations, paintwork, upholstery, vigilance, and carriages.

The reasons for leaving the company produced a catalogue of social relations in the company, from "being drunk and disorderly" to "claiming to be ill while working in another firm". These were classified into: death, retirement, illness, resignation, transfer, redundancy, and disciplinary dismissal. The data set did not explain why a worker who had been "transferred" had ceased to work for the company. Perhaps such workers had started work for a subsidiary of MZA. Table 1 gives summary statistics for each variable.

Variable	**Category**	**Frequency (n = 527)**
Reason for leaving	Resignation	145
	Disciplinary dismissal	174
	Redundancy	10
	Retirement	38
	Illness	39
	Death	84
	Transfer	37
First job	Apprentice	195
	Skilled	329
	Supervisor	3
Section	Boiler making	63
	Forge	31
	Foundry	30
	Fitting	36
	Assembly	54
	Lathe	36
	Offices	3
	Installations	8
	Paintwork	105
	Upholstery	21
	Vigilance	3
	Carriages	137

Marital status	Single	270
	Married	246
	Widower	11
Place of birth	Madrid	191
	Rest of Spain	309
	France	16
	Other countries	11

Variable	**Mean**	**Median**	**Minimum**	**Maximum**
Age when joining (years)	27,91	26	13	53
Experience in the firm (years)	9,30	3	0	53
Initial salary (Pesetas)	2.747	2.500	500	10.000

Table 1.- Descriptive statistics for the variables in the data set

A BRIEF DESCRIPTION OF OPTIMAL SCALING

The idea underlining the assignment of quantitative values to qualitative concepts can be traced to an example in the 7th edition of "Statistical methods for research workers" by RA Fisher (1938). Its subsequent development and the debates that ensued are discussed by Welsh and Robinson (2004). De Leeuw (1976) published an algorithmic approach for "optimal scaling" based on what has become know as alternating least squares that was implemented in the computer package SPSS. The methodology was generalised so that any form of quantitative analysis with qualitative data can now be contemplated; for a full account see Young (1981). For an application of optimal scaling in the area of management analysis see Serrano et al. (2004). We will now give a summary description of the principles involved with the help of an example.

Consider the following cross-classification- Table 2- where the rows indicate the height (two categories) and the columns indicate the weight (two categories). Each cell gives values for the weight that has been observed in particular individuals.

	Thin	Fat
Short	50, 55	70
Tall	75	90, 100

Table 2.- Weight of five individuals classified by means of two criteria

We see in Table 2 that there are two thin and short individuals, one weighs 50, and the other weighs 55. There is a short and fat individual who weighs 70, and so on.

We would like to estimate an additive model in such a way that we start by attaching a weight to being short and thin. The weight of a short and fat individual would be obtained by adding a given amount on account of the extra weight induced by fatness. In the same way, the extra weight induced by tallness would be obtained by adding another amount to the weight of a short and thin person. Finally, to estimate the weight of a fat and tall person we would start with the weight of a short and thin, add the increase produced by fatness and the increase produced by tallness. In this model we do not consider interactions.

Try, for example, 52 as the estimated weight of a short and thin person. Of course, this does not coincide with the observed and there are errors: one individual is 2 under and the other individual is 3 over the estimate. If we estimate 20 as the extra weight of fatness, a fat and short individual would weigh 52+20=72. The only short and fat individual weighs 70, an error of 2. If, proceeding in the same way, we estimate that

tallness adds an extra weight of 25, the weight of a thin and tall person should be 52+25=77, and not 75 as observed, an error of 2. Finally, the weight of a fat and tall person would be 52+20+25=97. Since we have two tall and fat individuals, we are making two errors, one of 3 (for the individual who weighs 100), and one of 7 in the case of the individual who weighs 90.

In summary, we have attached a value to being tall; we have attached a value to being fat, and we have a set of errors associated with our estimates. We could now change the estimated values in order to minimise some statistic of the errors, until the "best" set of errors is obtained. If the optimisation criterion is to minimise the sum of squared errors, this method is just analysis of variance.

Imagine now that, instead of the weights of the individuals, we are given the indication that some of them are male, some are female, and some are children. This would produce the following table- Table 3:

	Thin	Fat
Short	C, F	M
Tall	C	F, M

Table 3- Weight of five individuals classified by means of two criteria. C stands for "child", M stands for "male", and F stands for "female".

Fisher's idea was to proceed in two steps. First, guess the weight of a child, the weight of a male, and the weight of a female. Second, find the impact of fatness and the impact of tallness using least squares. Having done this, we go back to the estimated weight of a child, the estimated weight of a male, and the estimated weight of a female, change them and start again. This procedure is repeated until optimal scores are found for the C, F, and M categories, for the impact of tallness, and for the

impact of fatness. Because the process alternates between an estimate and least squares, it is called "alternating least squares". There are ambiguities that are resolved through normalisation.

In the above example, we have assumed that weight is a consequence of tallness, fatness, and person category. In the data we have there is no cause and effect. However, the principles involved are very much the same, although the procedure in the second step (minimising the sum of squares) is based on Kruskal's (1964) ordinal multidimensional scaling algorithm (MDS). For an introduction to MDS see Kruskal and Wish (1978).

OPTIMAL SCALING ANALYSIS OF RAILWAY WORKERS DATA

The data on the values of the eight variables for the 527 employees of MZA was entered in the computer package SPSS.

The analysis was performed using the Categorical Principal Components routine (CATPCA). This routine proceeds as follows. First, all the data- including the continuous variables- is categorised. Next, the alternating least squares algorithm is used to quantify each of the categories. After this step all the data becomes quantitative and the standard Principal Components (PCA) algorithm is applied.

The quantification of each of the categories is given in Table 4.

Variable	Categories (quantification)
Place of birth	Madrid (-0.42), Spain (-0.02), France (0.87), Other (6.61)
Marital status	Single (-0.93), Married (0.89), Widower (2.97)
First job	Apprentice (-1.30), Skilled (0.75), Supervisor (2.18)
Section	Boiler making (-0.76), Forge (-1.30), Foundry (-0.66), Fitting (0.01), Assembly (-1.42), Lathe (-0.42), offices (-0.24), plumbing and electric installations (-0.88), paintwork (1.61), upholstery (0.99), vigilance (1.82), carriages (0.08)
Reasons for leaving	Death (1.03), Retirement (2.52), Illness (1.18), Resignation (-0.65), Transfer (-0.51), Redundancy (-0.56), Disciplinary dismissal (-0.63)

Table 4 .- Quantification of qualitative categories using optimal scaling.

CATPCA was run with eight dimensions, the maximum possible when eight variables are present. The importance of a dimension is measured by the value of the associated eigenvalue. Under the Joliffe (1972) criterion, a dimension is important if the eigenvalue is greater than 0.8, although the usual approach is to consider a dimension as important if the eigenvalue is greater than 1. The results are shown in Table 5 where it can be seen that five eigenvalues meet the Joliffe condition and account for over 90% of the variance in the data.

Dimension	Variance Accounted For	
	Eigenvalue	% of total variance
1	**2,103**	26.29
2	**1,883**	49.82
3	**1,394**	67.25
4	**1,042**	80.28
5	**,812**	90.43

Table 5.- Eigenvalues and total variance explained

The results in the above table suggest that four or, perhaps, five independent characteristics can describe an employee of MZA. We will later consider the interpretation of such characteristics. For the moment it is just sufficient to observe that the first two characteristics account for just under 50% of the variation in the data. We should expect that most of the relevant information contained in the data can be found by looking at the first two characteristics (dimensions).

To interpret the meaning of the dimensions, it is usual to look at the component loadings. If a variable "loads high" on one of the components, it is relevant to the interpretation of that component. Component loadings are shown in Table 6. High loadings have been highlighted in Table 6.

	Dimension					
	1	**2**	**3**	4	5	6
Marital status	**,626**	-,010	**,647**	-,148	-,122	-,371
Place of birth	,104	-,191	-,099	**,858**	-,452	-,045
Enrolment age	**,724**	,109	**,523**	-,020	-,175	,369
First job	**,616**	**-,515**	-,457	-,052	,091	-,173
Section	,334	-,432	-,**484**	-,445	-,457	,106
Initial salary	**,653**	-,255	-,141	,289	**,583**	,104
Reason for leaving firm	,412	**,795**	-,315	,017	-,067	-,014
Experience in firm	,299	**,828**	-,361	,007	-,020	-,070

Table 6. - Component loadings. "High loadings" have been highlighted.

The variables that load high in the first principal component are: initial salary, enrolment age, marital status, and first job. A relatively older married person, who joined the company quite late in life to work as a supervisor would have a high value under this component. A low value would be associated with a single young person joining as an apprentice. We can label this component as "seniority" at the start.

The second component is associated with experience in firm, first job and reason for leaving the firm. This component takes high values for who join as apprentices and stay until retirement age. We can label it as "experience in the firm".

The third component appears to be related to marital status, enrolment age, and section. It takes high values for married workers who join the firm quite late in life and work in paintwork, upholstery, or vigilance. It is difficult to label this component, but we could describe it as "maturity at the time of joining the firm".

The fourth component is just "place of birth". The fact that this variable forms a component by itself indicates that it is not correlated to the remaining variables, nor to the other components. One may conclude that place of birth is not a relevant piece of information, but it is possible for a particular group of workers to display a distinctive feature related to birth. This is something we will explore later with respect to non-Spanish workers.

GRAPHICAL INTERPRETATION OF THE RESULTS

CATPCA, in the same way as PCA, calculates the values of the principal components for each individual. In this way, each individual can be represented in the space of the components. Figure 1 shows the representation of the data set in the first two principal components.

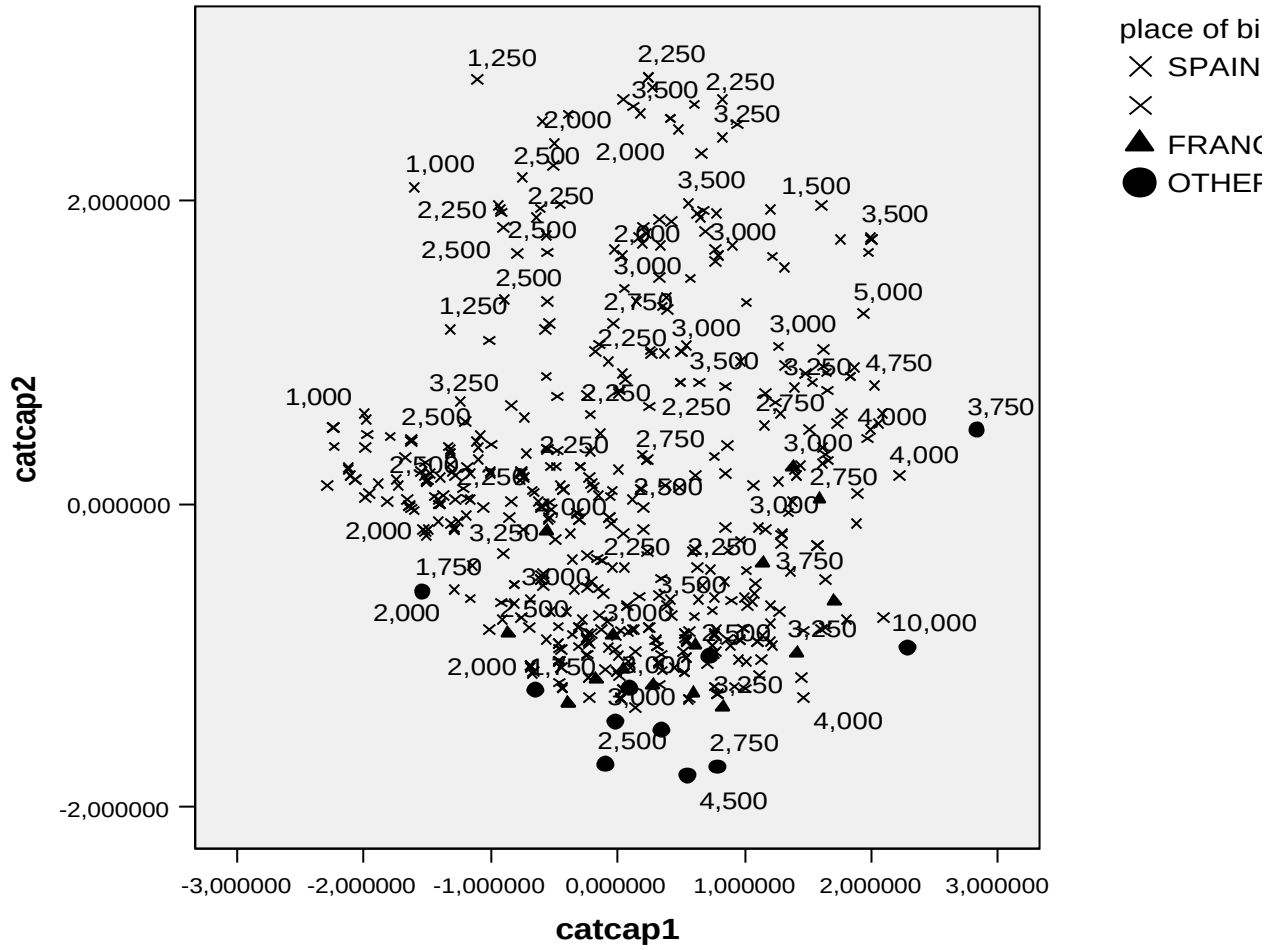

Figure 1.- Representation of the data in the first two principal components with an indication of first salary and place of birth.

In Figure 1 we have added to the representation an indication of salary, and of place of birth, making a distinction between Spanish born workers and workers born outside Spain. We can see that initial salary increases from left to right of the picture, indicating that seniority when joining the company is reflected in higher salaries, as one would expect. We also observe that almost all the workers who were born outside Spain are to be found on the South East corner of the representation, indicating that they share common characteristics. We deduce that workers born outside Spain are employed with salaries that are higher than average, although there are Spanish workers who are employed with salaries that are just as high.

In Figure 2, the same data is represented, but this time next to the point that is associated with each individual we have written the age at time of joining, and we have also indicated, by means of a symbol, the reason for leaving employment.

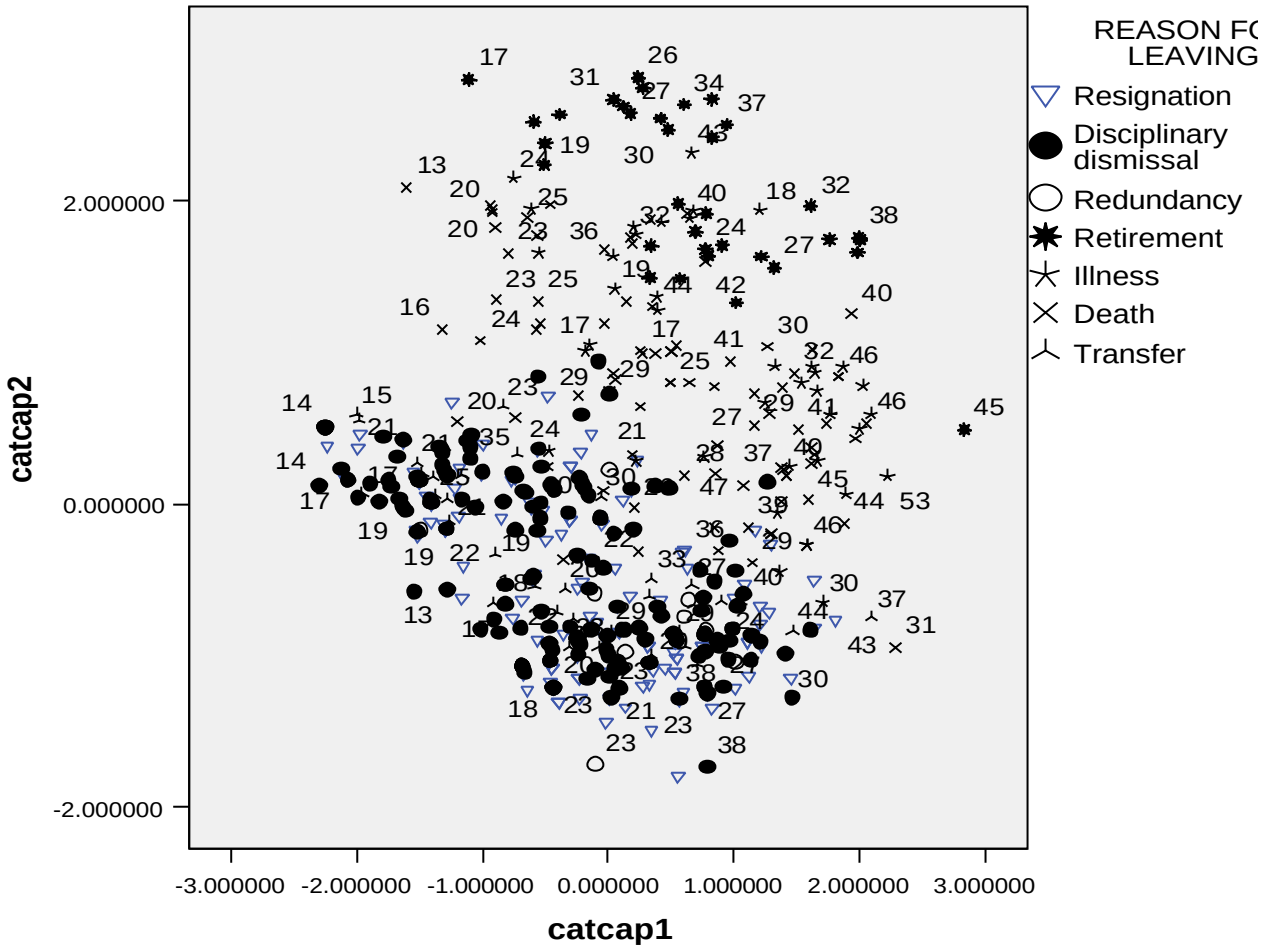

Figure 2.- Representation of the data in the first two principal components with an indication of age at time of joining and reason for leaving the firm.

We notice that age increases from left to right in the figure. The way in which the reason for leaving plots is particularly interesting. Towards the top of Figure 2 we find what would be described as "normal" exit: retirement, death, and illness. At the bottom of the figure we find redundancy, disciplinary dismissal, resignation, and transfer. This may indicate that social distress increases from top to bottom of the

figure. With what variables would this social distress be associated? We will explore this by adding information on the figure in the same way in which North-South, and East-West directions are superimposed on geographical maps. For each characteristic of the data we will draw a line in such a way that it points in the direction in which the characteristic increases. This technique is known as Property-Fitting, (Pro-Fit). For a description of how Pro-Fit works see Schiffman et al (1981). Pro-Fit, as originally developed, works with quantitative variables only, but it can be shown to work with dichotomical variables; see Mar Molinero and Mingers (2006).

In order to apply Pro-Fit, for each variable, we run a regression in which the observations are the individuals. The dependent variable is the characteristic of the data that interests us; for example, age at the time of joining the company. The explanatory variables are the values of the categorical principal components for each individual. When the dependent variable is a dichotomy- for example, "normal" versus "abnormal" exit-, the regression takes the form of binary logistic. Being regression based, the usual statistics are available to measure the quality of fit. Pro-Fit produces a series of oriented lines through the representation, but we prefer to represent these lines by means of a normalised, oriented vector. Normalisation has the advantage that the length of the vector indicates its relevance in the interpretation of the figure. The following variables have been used as properties for the purpose of this analysis: age at time of joining, starting salary, experience in the firm, reason for leaving (normal versus abnormal), marital status (single/not single), place of birth (born in Spain/not born in Spain), and first employment (apprentice/other). The last four variables were entered as dichotomies. The technical details are given in Table 7, and the projection of the normalised vectors can be seen in Figure 3. We can see

that in every case, the coefficient of determination took very high values, indicating that Pro-Fit is an appropriate technique to interpret the figures.

VARIABLE	β_1	β_2	β_3	β_4	β_5	β_6	R^2
Joining age	0.74	0.10	0.53	-0.01	-0.17	0.37	0.93
Initial salary	0.68	-0.26	-0.11	0.40	0.55	0.06	0.82
Experience in firm	0.27	0.87	-0.40	0.01	-0.02	-0.11	0.95
Reason for leaving: normal/abnormal	0.42	0.84	-0.32	0.02	-0.09	-0.03	0.96 (Nagelkerke)
Marital status: single/not single	0.50	0.00	0.77	-0.12	-0.15	-0.35	1.00 (Nagelkerke)
Place of birth: Spain/not Spain	-0.07	0.19	0.11	-0.85	0.47	0.04	1.00 (Nagelkerke)
First job: Apprentice/other	0.48	-0.68	-0.51	-0.08	0.05	-0.22	0.99 (Nagelkerke)

Table 7.- Results of Property Fitting analysis: directional cosines and coefficients of determination. In the case of bivariate logit, the R^2 is Nagelkerke's.

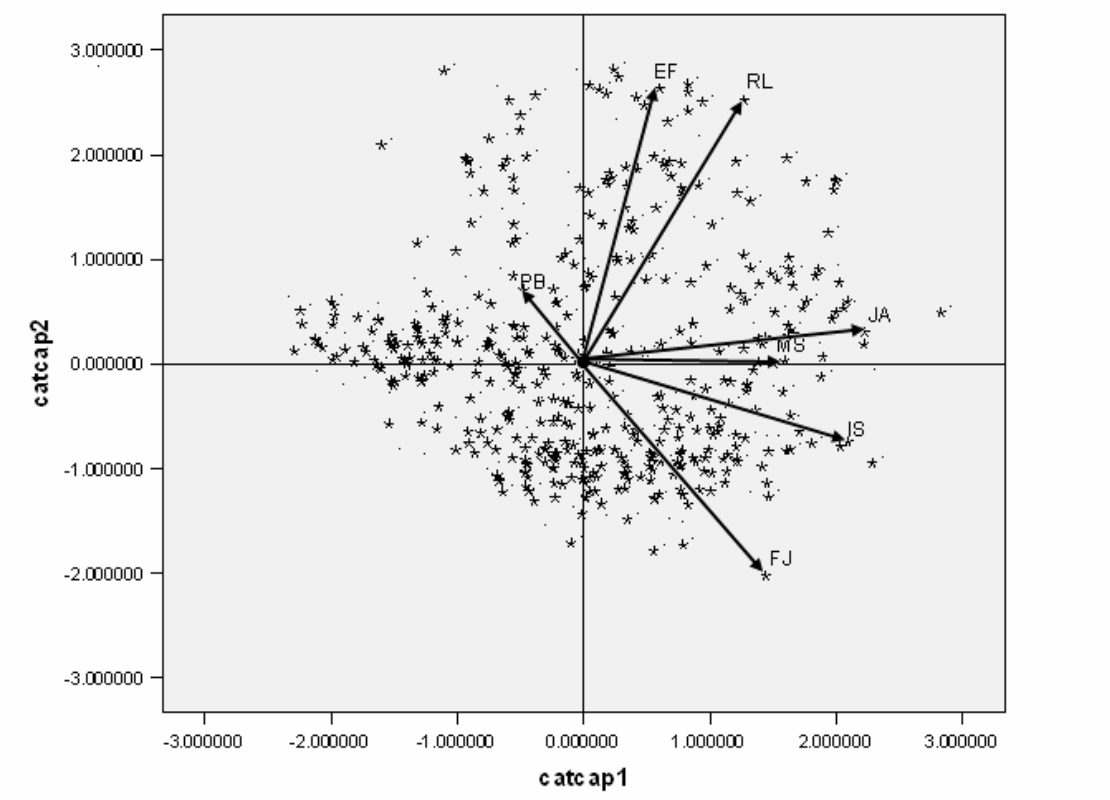

Figure 3.- Results of Property Fitting analysis. Place of Birth (PB), Experience in Firm (EF), Reason for Leaving (RL), Joining Age (JA), Marital Status (MS), Initial Salary (IS), First Job (FJ).

In Figure 3, vectors that appear at acute angles are highly associated in a direct way. We see that high associations exist between EF (experience in firm) and RL (reason for leaving), indicating that "normal" exit is associated with spending many years in the firm; in other words, workers who have been in the firm for a long time tend not to be sacked. The converse is also true; workers who are dismissed in an abnormal way tend to be those who have not been in the firm for very long. JA (joining age) and MS (marital status) are also positively associated, indicating that married workers tend to be older than unmarried ones. JA and MS are both positively associated with IS (initial salary), this can be interpreted to mean that older workers tend to be paid more when they join. Initial salary appears not to be related to experience in the firm (the two vectors are at right angles) or with reason for leaving. Seniority at first job is related with initial salary, very much as one would expect. Seniority at first job is negatively related with being born in Spain, as can be ascertained from the fact that FJ and PB are almost in the same line but in the opposite direction.

It would be possible to further explore the data set by bringing into the picture the third categorical principal component, but an examination of the size of the directional cosines in Table 6 indicates that most of the story behind the data is the first and the second components.

CONCLUSION

The analysis of most data sets is complicated by the fact that they contain a mixture of qualitative and quantitative information. In this paper we have shown how optimal scaling can be used to attach quantitative scores to qualitative categories, so that the standard multivariate analysis tools can be applied.

The data set contained information about 527 railway workers in one of the two main Spanish railway companies of the 19th Century. This company has been extensively studied from the point of view of its management, but nobody appears to have studied a company from the point of view of the workers who were employed in it.

The data has revealed several features of work in the company. Some of the workers in the data set had not been born in Spain. These tended to obtain higher salaries and more senior positions than the average worker who had been born in Spain, although some Spaniards received, when joining, salaries just as high, and were employed in jobs that were just as senior.

Of particular interest during this period is social distress. This is revealed by the way in which workers left the company. The analysis clearly shows that most workers who leave the company in an abnormal way (redundancy, disciplinary dismissal), do so not far after they have joined. The reason for leaving is unrelated to both first salary and seniority at the time of joining.

REFERENCES

De Leeuw, J. (1976) Additive structure in qualitative data: an alternating least squares method with optimal scaling features. Psychometrika, 41, 471-503.

Fisher, R.A (1938) Statistical methods for research workers, 7th edition. Oliver and Boyd, Edinburgh, UK.

Joliffe. I.T. (1972): Discarding variables in Principal Components Analysis. Applied Statistics, 21, 160-173.

Kruskal, J.B. (1964). Multidimensional Scaling by optimizing goodness of fit to a non-metric hypothesis. Psychometrika, 29, 1-27.

Kruskal, J.B. and Wish, M. (1978). Multidimensional Scaling. Sage. London. UK.

Mar Molinero, C.; and Mingers, J. (2006). Mapping MBA Programmes: an alternative analysis. Forthcoming in the Journal of the Operational Research Society.

Schiffman, S.S., Reynolds, M.L. and Young, F.W. (1981): Introduction to Multidimensional Scaling: Theory, Methods and Applications. Academic Press, London

Serrano, C.; Mar Molinero, C.; and Chaparro, F. (2004). Spanish savings banks: a view on intangibles. Knowledge Management Research & Practice, 2, 103-117.

Stevens, S.S. (1951) Mathematics, measurement, and psychophysics. In S.S. Stevens (Ed.), Handbook of Experimental Psychology. New York: Wiley.

Welsh, A.H.; and Robinson, J. (2005) Fisher and inference for scores. International Statistical Review, 73, 131-150.

Young, F.W. (1981) Quantitative analysis of qualitative data. Psychometrika, 46, 357-388.

UNIVERSITY OF KENT

http://www.kent.ac.uk/kbs/research-information/index.htm