3rd International Conference on Industrial Engineering and Industrial Management
XIII Congreso de Ingeniería de Organización
Barcelona-Terrassa, September 2nd-4th 2009

Analysis of missing qualitative data values with Optimal Scaling features

Cecilio Mar Molinero[1], Fabiola Portillo[2]

[1] Kent Business School. University of Kent. Canterbury, Kent, CT2 7NZ, Reino Unido.
C.Mar-Molinero@kent.ac.uk
[2] Dp--to. de Economía y Empresa. Facultad de CC. Empresariales. Universidad de la Rioja. C/ Cigüeña, 60, 26004. Logroño. fabiola.portillo@unirioja.es

Keywords: optimal scaling, qualitative missing data, data-mining, internal labour markets

1. Introduction

The main objective of this paper is the study of managerial innovation diffusion in a dependent economy, in particular the introduction of internal labour markets in a major Spanish railway company, MZA, in the late 19th Century, and to suggest the way in which such an innovation might have arisen. MZA was established in Madrid in 1856 with French and Spanish capital. Until well into the 20th Century, MZA was either the largest, or the second largest, firm in Spain and amongst the top ten in Europe. Baker et al. (1994: 882) observed that "descriptions of firm policies" based on internal labour markets "have been subjected to little careful quantitative study"; and this appears to still be the case. This study attempts to provide new insights into the introduction of advanced managerial methods in Europe using the quantitative approach.

The data used were obtained from the Spanish Railway Foundation (FFE) which holds the personal files of the workers employed in the Madrid Atocha workshop of MZA. This data set contains personal information on 889 employees who were hired between 1882 and 1889. Some of the information contained in the files ─such as initial salary─ is quantitative, and some information ─such as reasons for leaving the firm─ is qualitative. To be able to properly analyze the multivariate information contained in the data set, it is necessary to jointly deal with qualitative and quantitative data and, therefore, with variables measured in several scale types, something that determines the statistical methods to be used. We quantify qualitative data through the application of Optimal Scaling (OS) methods. This permits the application of the standard techniques of multivariate analysis. In this case, we used Categorical Principal Component (CPC) and Property Fitting Analysis because they allow the graphical representation of the results and, by so doing, highlight the main features of the data set. Besides, there was a problem of missing data that had to be addressed: out of the 889 files, only 537 contained complete information. Missing data were estimated using an OS based method appropriate for the imputation of categorical data, MISTRESS, suggested by Van Buuren and Van Rijckevorsel (1992).

The predictions that the internal labour markets theory makes were supported by the data: it was found that most workers who left the firm in an abnormal way ─resignation, redundancy, or disciplinary dismissal─ did so not long after they had been hired, and that their reason for leaving was unrelated to both first salary and seniority at the time of joining. It was also found that workers who left the company through retirement, illness, or death had long employment careers. These results are consistent with the existence of some kind of internal labour markets in MZA in the 1880s. The hypothesis that management innovations found

their way to MZA through the French connection was also consistent with the data: French workers tended to be employed in more senior jobs, to have relatively short periods of employment in MZA, and were paid higher salaries than Spanish workers. This is consistent with their being employed as specialists or consultants, as would be expected to be the case of a company where the main source of expertise was French.

The paper is structured as follows. A brief discussion of internal labour markets forms the body of the next section, which also contains the hypotheses to be tested on the data. This is followed by a description of the data and methods used and its characteristics. The next section, results, is divided into various parts. The first analysis subsection describes the way in which categorical missing data estimation took place; the second and the third analysis subsections deal with the main body of the research using the techniques of CPC and Property Fitting, respectively. The paper ends with a concluding section.

2. Theory and Hypotheses

Much recent work on internal labour markets has focused on their evolution and actual implications (Osterman 1994, 2006; Elvira and Graham 2002). The presence and development of internal labour markets implies that firms internalize labour processes that are normally external. Such internalization can be analyzed from a variety of theoretical perspectives (Miyazaki 1977; Elbaum 1983; Osterman 1987), and is often associated with large modern organizations that operate in well defined and relatively stable markets, and are able to hire workers with long term contracts (Fitzgerald 1988). This situation applies particularly to large railway companies (Howlett 2000, 2004).

Basically, internal labour markets consist of a set of explicit or implicit rules and procedures governing labour relations, particularly in respect to the hiring of staff, internal promotion opportunities, and wage policies (Lazear and Oyer 2003). In an internal labour market, new workers join the firm at the so-called "ports of entry", which tend to be associated with low qualification levels, and are paid the market rate for the job (Doeringer and Piore 1971). Access to higher level jobs would be limited to employees who have been in the firm for some time, in order to use internal promotion as a reward, something that the transaction costs model would predict (Wachter and Wright 1990; Siebert and Addison 1991). Pay policies are consistent with the efficiency wage model in that internal workers receive relatively high rewards with respect to external labour markets (Idson and Feaster 1990); such rewards may take the form of welfare programs.

Internal labour markets bring benefits both to the workers and to the owners of the firms. These benefits can be analyzed from the perspective of the risk aversion model, from the perspective of the transaction costs model, and from the perspective of the specific human capital model. Risk adverse workers see that internal labour markets reduce the uncertainties associated with the economic cycle, and with possible periods of inability to work (Bailey 1974). Companies also derive benefits from internal labour markets, as these generate a stable and permanent labour force, therefore reducing the transaction and search costs associated with the provision of an appropriately skilled labour force, whilst, at the same time, facilitating firm-specific investments in human capital via on-the-job training, since the training is more likely to result in the long-term benefit of the firm (Siebert and Addison 1991; Mackinnon 2004).

Several studies, both theoretical and empirical, confirm that internal labour markets bring important efficiency gains to large firms (Williamson 1985). This is explained by the fact that large firms rely more on specific investment in human capital and have more difficulties in controlling workers. Internal labour markets would solve both problems by providing in-firm

training and by establishing systems for collective regulation and grievance procedures that are less costly than the alternatives (Siebert and Addison 1991). This requires the introduction of bureaucratic procedures that facilitate workers' control (Mayer and Nickerson 2005). Firms could also benefit from more positive attitudes on the part of workers who get enhanced job security and better promotion prospects (Elbaum 1983).

Summarizing, the theoretical models predicts:

- Hypothesis 1. Entry into the firm will take place at low levels of qualification, new workers taking up relatively junior jobs with relatively low initial wages.
- Hypothesis 2. Exit from the firm will be more common amongst workers who joined recently, since those workers who build up some experience would loose their non-wage benefits if they were to leave.
- Hypothesis 3. Starting wage, being determined by external labour markets is related to the skills and experience that the worker brings into the firm but it is not related to length of stay or reasons for leaving.
- Hypothesis 4. Reasons for leaving such as resignation, dismissal, or redundancy will be associated with short employment periods.
- Hypothesis 5. Workers who survive the initial qualifying period will leave the firm for reasons such as illness, retirement, or death.

3. Data and Methods

The data was collected from the personal files of workers who joined the Spanish railway company MZA during the period 1882 to 1889. This information was originally hand written; it was coded into an ACCESS file, and errors were made in those processes. When an obvious error was identified, the worker was excluded from the data set. For instance, there was a worker whose date of birth was recorded as 1884 but who was said to have started working in the firm in 1882, two years before birth. The clean data set contained 889 workers.

The original information was recoded into eight variables: enrolment age, experience in the firm, initial wage, first job performed, leaving reason, marital status, place of birth, and section in which they carried out their activity. The first three variables are quantitative, while the last five are qualitative. Leaving reason was often left blank, and in the end only 537 files contained complete information. The gender of the worker could be inferred from the name, available in the data, but there were only two females in the sample, and gender was not included in the analysis.

The data set contains a mixture of variables measured in nominal, ordinal, and ratio scales. OS is appropriate in this situation since it quantifies qualitative data and, by doing so, makes it possible to apply the standard methods of multivariate analysis. The OS analysis was performed twice. In the first instance, only the 537 employees for whom complete information was available were included (Portillo et al. 2006). The quantification of qualitative variables generated, as a sub product, the matrix of correlations between variables. This was important information for the imputation of missing observations. A great deal of effort went into the study of missing values, so that these could be estimated without affecting the statistical properties of the data set. Having imputed values to missing variables, the OS analysis was performed a second time, and the results were interpreted using CPC analysis and Property Fitting techniques.

The existence of missing data is a common problem when working with data bases. Which method is appropriate to deal with missing values depends on the missingness pattern and mechanism that generates the absences (Little and Rubin 2002). The missing data pattern describes which values are missing in the data matrix, and the mechanism describes the relationship between absent data and variable values. In this particular case, the only variable with missing values is leaving reason, hence the missing data pattern is univariate. The absence of leaving reason is significantly related to the values of other variables —experience in the firm, first job, place of birth, and section–, therefore the mechanism is not Missing Completely at Random (MCAR); for example, leaving reason is missing more frequently amongst low experienced workers employed in low level jobs. Common sense suggests that MCAR does not apply in this case. One can only imagine that, if a worker does not turn up to work on a particular day, there will be an inclination to wait to see if the absence is temporary or permanent, and that permanent absences may just go unrecorded. On the other hand, death during service is a traumatic event and will probably be immediately recorded. Thus, the absence of the value of a variable may depend on the value of another variable, against the assumption of the MCAR process. When the data is MCAR, ignoring observations with incomplete data results in correct inferences, although there is some information loss. When the presence or absence of an observation in a particular variable depends on the value of another variable, one should take into account, during model estimation, the process by which observations are missing (Rubin 1976).

If all the variables were measured on a ratio scale, we would be taking advantage of correlations between variables in order to estimate the most likely values for the missing data. In this particular data set, because some of the variables, including leaving reason, are qualitative, imputation of missing values is more complex (Von Hippel 2004). A several step procedure was followed for the estimation of missing values: we first applied the OS model to the data that excluded employees with missing observations; second, correlations between variables were studied; third, the relationships between variables were explored using Path Analysis (Retherford and Choe 1993); and, finally, a homogeneity analysis based procedure due to Van Buuren and Van Rijckevorsel (1992) was followed in order to obtain the imputations that were used in the final OS analysis of the data.

4. Results

4.1. Categorical Missing Data

The application of OS to the data set that contains no missing values –537 observations– generated a set of correlations between variables. Having established that the missing values do not follow a MCAR process, a study of the correlation structure of the data was performed using Path Analysis. The AMOS routine of SPSS was run using as endogenous variables experience in the firm, initial salary, and leaving reason. The remaining variables –enrolment age, first job, marital status, place of birth, and section of work– were treated as exogenous. The only two relevant path coefficients associated with leaving reason that took values significantly different from zero were the ones that linked this variable with enrolment age, and experience in the firm, thus confirming the insights gained from the previous steps.

The final step in the estimation of missing values was to use the imputation procedure proposed by Van Buuren and Van Rijckevorsel (1992). This procedure maximizes consistency in the data set as measured by Guttman's η^2–statistic (Guttman 1941). The MISTRESS technique is implemented in the SAS environment, and combines correspondence analysis methods with the k-means clustering algorithm. The variables found to be associated

with leaving reason in the previous analyses –enrolment age, and experience in the firm– were used to obtain estimates of missing values, whose distribution is presented in Table 1.

Variable		Count			Percentage			χ^2-statistic[a] (p–value)
		Observed	MCAR Expected	MISTRESS Imputed	Observed	MCAR Expected	MISTRESS Imputed	
Leaving reason	Missing	352	0	0	39.6	0.0	0.0	68.720 (<0.01)
	Resignation	228	378	472	25.6	42.5	53.1	
	Dismissal	80	132	98	9.0	14.9	11.0	
	Redundancy	27	45	59	3.0	5.0	6.6	
	Transfer	38	63	76	4.3	7.1	8.5	
	Illness	38	63	39	4.3	7.1	4.4	
	Death	86	142	95	9.7	16.0	10.7	
	Retirement	40	66	50	4.5	7.4	5.6	

Note. [a] χ^2 test for the difference between the expected distribution under MCAR mechanism and MISTRESS estimates.

Table 1. Missing data imputation for leaving reason

The missing values was 39.6% with respect to the total in leaving reason, and 4.9% with respect to the database. These percentages are within the acceptability limit of unknown entries required to avoid consistency bias, obtained by Van Buuren and Van Rijckevorsel (1992) using the bootstrap. The η^2-statistic took the value 0.68, which is higher than the 0.50 limit required to maximize consistency, also obtained by these two authors.

The last column in Table 1 reports the results of a χ^2-test to explore the differences between the expected values under the MCAR mechanism and the imputed values generated by MISTRESS. These differences were found to be significant at the 1% level, confirming that the MCAR was not appropriate, and that MZA employees whose leaving reason was unknown were likely to have resigned.

4.2. Optimal Scaling and Categorical Principal Component

After imputing missing values, the data for the eight variables on 889 workers were analyzed using CPC. The first two categorical principal component account for about 50% of the variation in the data, and the first five components account for almost 90%.

The scores assigned to the categories of each non-numerical variable are given in Table 2. We will now discuss score values, as these are important for interpretation purposes. Scores in first job increase as seniority increases: the lowest score in first job is associated with apprentices and the highest score is attached to supervisors. Moving on to the scores of the various categories of leaving reason, it is to be noticed that abnormal reasons –resignation, dismissal, redundancy– show negative scores, while categories that are a normal end to an employment career –illness, death, and retirement– show positive scores. The ordering of the scores for the different categories of place of birth is to be noted. The highest score is attached to the category "France". This makes sense in the context of the MZA firm, since it is suspected that French workers are specialists who joined the firm for a short period of time, only to return to their original job when they had transmitted the knowledge that required their presence in Madrid. If we turn now our attention to the section that workers joined, we see that sections that involved heavy engineering work –forge, lathe, boiler making, and so on– are associated with negative weights; sections such as upholstery, paintwork, vigilance, and so on, are associated with positive weights.

Variable		Frequency $n = 889$	Optimal Scaling	
			Scores	F-statistic[a] (p–value)
First job	Apprentice	24	–2.049	26.358 (<0.01)
	Unskilled	160	–1.125	
	Work assistant	180	–1.125	
	Skilled	520	0.815	
	Supervisor	5	1.546	
Leaving reason	Resignation	472	–0.601	25.698 (<0.01)
	Disciplinary dismissal	98	–0.438	
	Redundancy	59	–0.224	
	Transfer	76	–0.158	
	Illness	39	2.021	
	Death	95	1.599	
	Retirement	50	2.421	
Place of birth	Madrid	283	0.692	12.335 (<0.01)
	Rest of Spain	559	0.015	
	France	30	4.257	
	Other	17	3.511	

Note. [a] Fisher´s (1938) test for the global significance of OS parameters.

Table 2. Optimal Scaling scores for qualitative variables

Table 2 also shows the result of Fisher's (1938) test for the significance of estimated scores for qualitative variables. It can be observed that the F-statistics for all variables are significant at the 1% level, indicating that all variables included in the model are relevant. We conclude from the result of these tests that there are highly significant differences between the workers hired by MZA, and that these differences are captured by their first job, leaving reason, marital status, place of birth, and section in which they were employed. Fisher (1938) highlights the importance of this result, since it is only under this condition that the scores estimated by the OS model are meaningful.

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. The variables that load highest in the first component are enrolment age, marital status, first job, and initial salary. Taking into account the quantifications reproduced in Table 3 and the loadings, we can deduce that mature workers who joined as supervisors with a high initial salary, score high in the first principal component; while young workers who joined as apprentices, with low salaries, achieve low scores in this component. The first principal component can, therefore, be interpreted in terms of experience before the worker joined the firm. The variables that load high in the second component are experience in the firm, and leaving reason. At the positive extreme of the second component we find those workers who had been in the firm for a long time, and retired from the job; while, at the negative extreme, we find workers who were sacked from the firm after a short employment period. The second component is, therefore, associated with permanence in the firm.

4.3. Graphical Representation: Property-Fitting Analysis

Property Fitting is a regression based technique that can be located within the context of Biplots (Gower and Hand, 1996). Details of the particular form of the algorithm implemented here can be found in Kruskal and Wish (1978). Schiffman et al. (1981) give a simple introduction to the method. Mar Molinero and Mingers (2007) provide the mathematical justification. Property Fitting draws normalized vectors —of unit length— through the space

of the principal components in the direction in which a property of the data grows. In this case, the values of the various variables have been taken as properties. The vectors are shown in Figure 1. Take, for example, the variable "enage" —enrolment age—, the associated vector points towards the right hand side of the Figure, indicating that the older a worker was when joining, the more to the right of the configuration he will be plotted.

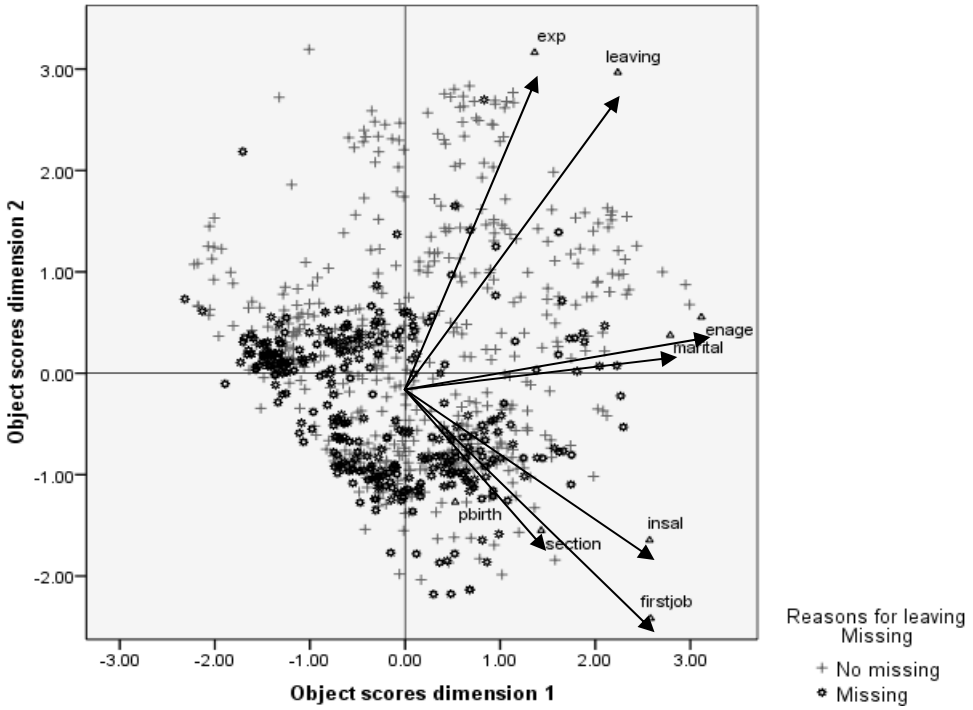

Figure 1. Workers plotted in the first two categorical principal components with indication of missing data, and projection of Property Fitting vectors.

The position of the vectors is calculated using regression analysis. The coordinates of vector in each two-dimensional representation are the normalized βs, which can be interpreted as directional cosines. The analysis also generates a measure of quality of fit, R^2. Vectors whose R^2 is low —less than 0.5— are not normally represented. In this case, the results were excellent; the lowest adjusted R^2 value was 0.83. The longer the projection of the vector, the more relevant is the two dimensional plot to the interpretation of the results. The angle between any two vectors depends on the correlation between the variables involved. Acute angles indicate positive correlation, the smaller the angle the higher the correlation. Orthogonal vectors are associated with a lack of correlation.

A further feature of the data highlighted in Figure 1 is whether the observation contains imputed values or not. Most workers for whom the leaving cause was estimated are plotted towards the left and towards the bottom of the Figure. Using the information contained in the oriented vectors we deduce that these are workers who did not stay in the firm for long; who were rather young when they joined the firm; and who were single. We also see that enrolment age, initial salary, place of birth, and section are not related to the absence in the worker's file of the leaving reason. This confirms the validity of the missing data procedure used in the analysis.

In conclusion, Figure 1 jointly represents the workers and the variables measured on them. There is also a measure of the relevance of the variable to the two-dimensional representation —through the length of the vector—, and an indication of the association between the

variables —through the angle between the vectors. We will now proceed to interpret the information in Figure 1 in the light of the hypotheses put forward in the text.

5. Conclusions

In this paper we have addressed methodological and business questions. The methodological questions relate to missing values and to the correct procedures to deal with a mixture of qualitative and quantitative variables. We will address these in turn. The treatment of missing data is not straightforward. A model based only on observations for which complete information is available may not be representative of the population unless the data is Missing Completely at Random. In this study we have established that the data was not Missing Completely at Random, and that the missing data contained valuable information about the process under study. We followed a several step process for missing value imputation that culminates on the maximization of internal consistency. This made it possible to operate with the full data set without dropping any observations.

A further difficulty was the coexistence of qualitative and quantitative variables. OS procedures are appropriate in these cases. Optimal Scaling was introduced by Fisher (1938), but it has seldom been applied to the analysis of business problems. In this paper we have used OS in order to transform qualitative information into quantitative. After this, we have applied standard tools of multivariate analysis. We have also used hypotheses tests originally developed by Fisher (1941) in order to establish the relevance of the variables in the model.

The analysis confirmed that most workers were hired for low level jobs. It is also clear, that many workers stayed in the firm for a very short time period. We have also established that workers for whom the leaving cause was unknown were similar to workers with low level of qualification who left after a short employment period. We also observed the presence of a relatively important contingent of French workers in the company; these were older, better paid, did not stay long in MZA; and occupied better jobs than the average Spanish worker. All this is consistent with the structure of capital in MZA and suggests that French workers were hired as advisors or consultants.

The application of Optimal Scaling techniques has resulted in the estimation of highly significant parameter scores whose values conform to prior expectations. Following this, all variables —five qualitative variables and three quantitative variables— were entered in the Categorical Principal Components algorithm. It was found that the first two categorical components account for about 50% of the variability in the data. The first component was interpreted as “prior experience”, and the second was interpreted as “permanence in the firm”.

In summary, having applied the methods of Optimal Scaling to study the possible existence of some kind of internal labour market in the Spanish railway company MZA at the beginning of the Second Industrial Revolution, we observed two characteristics predicted by the internal labour market theory: the existence of “ports of entry” for low levels of qualification, and long term labour relations. We conclude that the company MZA was already operating under some form of internal labour market in the 19th Century. This is consistent with the presence of internal labour markets in the UK and Australia during the same period.

References

Bailey, M.N. 1974. Wages and unemployment under uncertain demand. Review of Economic Studies 41(1) 37-50.

Baker, G., M. Gibbs, B. Holmstrom. 1994. The internal economics of the firm: Evidence from personnel data. The Quarterly Journal of Economics. CIX(November) 881-919.

Chandler, A.D. 1977. The visible hand. The managerial revolution in American business. Belknap-Harvard, Cambridge, MA.

Doeringer, P.B., M.J. Piore. 1971. Internal labour markets and manpower analysis. Lexington Books, Heath, Lexington, MA.

Elbaum, B. 1983.The internalization of labour markets: Causes and consequences. American Economic Review 73(2) 260-265.

Elvira, M.M., M.E. Graham. 2002. Not just a formality: Pay system formalization and sex-related earnings effect. Organization Science 13(6) 601-617.

Fisher, R.A. 1938, 1941. Statistical methods for research workers (7th and 8th editions). Oliver and Boyd, Edinburgh.

Fitzgerald, R. 1988. British Labour Management and Industrial Welfare. Croom H, London.

Gower, J.C., D.J. Hand. 1996. Biplots. Chapman & Hall, London.

Guttman, L. 1941. The quantification of a class of attributes: A theory and method of scale construction. P.H. Horst eds. The prediction of personal adjustment. Social Sciences Research Council, New York, 319-348.

Howlett, P. 2000. Evidence of the existence of an internal labour market in the Great Eastern Railway Company, 1875-1905. Business History 42(1) 21-40.

Howlett, P. 2004. The internal labour dynamics of the Great Eastern Railway Company, 1870-1913. Economic History Review LVII(2) 396-422.

Idson, T., D. Feaster. 1990. A selectivity model of employer-size wage differentials. Journal of Labour Economics 8(1) 99-122.

Kruskal, J.B., M. Wish. 1978. Multidimensional Scaling. Sage, London.

Lazear, E.P., P. Oyer. 2003. Internal and external labour markets: a personnel economics approach. Working Paper 10192, NBER, Cambridge, MA.

Little, R.J.A., D.B. Rubin. 2002. Statistical analysis with missing data. Wiley & Sons Inc, Hoboken, New Jersey.

Mackinnon, M. 2004. Trade Unions and employment stability at the Canadian Pacific Railway, 1903-1929. D. Mitch, J. Brown, M.H.D. van Leeuwen, eds. Origins of the modern careers. Ashgate, Aldershot, 126-144.

Mar Molinero, C., J. Mingers. 2007. Mapping MBA Programmes: an alternative analysis. Journal of the Operational Research Society 58 874-886.

Mayer, K.J., J.A. Nickerson. 2005. Antecedents and performance implications o contracting for knowledge workers: Evidence from information technology services. Organization Science 16(3) 225-242.

Miyazaki, H. 1977. The rat race and internal labour markets. The Bell Journal of Economics 8(2) 394-418.

Osterman, P. 1987. Choice of employment systems in internal labour markets. Industrial Relations 26(1) 46-67.

Osterman, P. 1994. How common is workplace transformation and who adopts it?. Industrial and Labour Relations Review 47(2) 173-178.

Portillo, F., C. Mar Molinero, T. Martínez Vara. 2006. Interpreting a data base of railway workers using optimal scaling techniques. Working Paper 127, Kent Business School, Canterbury.

Retherford, R.D., M.K. Choe. 1993. Statistical Models for Causal Analysis, chapter 4. John Wiley & Sons, New York.

Rubin, D.B. 1976. Inference and missing data. Biometrika 63 581-592.

Schiffman, S.S., M.L. Reynolds, F.W. Young. 1981. Introduction to Multidimensional Scaling: Theory, methods and applications. Academic Press, London.

Siebert, W.S., J.T. Addison. 1991. Internal labour markets: Causes and consequences. Oxford Review of Economic Policy 7(1) 76-92.

Von Hippel, P.T. 2004. Biases in SPSS 12.0 missing value analysis. The American Statistician 58 160-164.

Van Buuren, S., J.L.A. Van Rijckevorsel. 1992. Imputation of missing categorical data by maximizing internal consistency. Psychometrika 57 567-580.

Wachter, M.L., R.D. Wright. 1990. The economics of internal labour markets. Industrial Relations 29(2) 240-262.

Williamson, O.E. 1985. The economics institutions of capitalism. Collier Macmillan, New York.