Multiple Correspondence Analysis of Subsets of Response Categories

Michael Greenacre[1] *and Rafael Pardo*[2]

[1]Departament d'Economia i Empresa
Universitat Pompeu Fabra
Ramon Trias Fargas, 25-27
08005 Barcelona. Spain

E-mail: michael@upf.es

[2]Fundación BBVA
Paseo de Recoletos, 10
28001 Madrid. Spain

E-mail: rpardoa@fbbva.es

Acknowledgement The support of the BBVA Foundation (Fundación BBVA) in Madrid in sponsoring this research is gratefully acknowledged.

Multiple Correspondence Analysis of Subsets of Response Categories

Michael Greenacre and Rafael Pardo

Abstract: In the analysis of multivariate categorical data, typically the analysis of questionnaire data, it is often advantageous, for substantive and technical reasons, to analyse a subset of response categories. In multiple correspondence analysis, where each category is coded as a column of an indicator matrix or row and column of Burt matrix, it is not correct to simply analyse the corresponding submatrix of data, since the whole geometric structure is different for the submatrix . A simple modification of the correspondence analysis algorithm allows the overall geometric structure of the complete data set to be retained while calculating the solution for the selected subset of points. This strategy is useful for analysing patterns of response amongst any subset of categories and relating these patterns to demographic factors, especially for studying patterns of particular responses such as missing and neutral responses. The methodology is illustrated using data from the International Social Survey Program on Family and Changing Gender Roles in 1994.

Keywords: Correspondence analysis, exploratory data analysis, missing data, multivariate categorical data, principal components analysis, questionnaire data.

1 Introduction

In the social sciences the principal application of multiple correspondence analysis (MCA) is to visualize the interrelationships between response categories of a set of questions in a questionnaire survey, for example a set of statements to which the respondents answer on the following scale: “strongly agree”, “somewhat agree”, “neither agree nor disagree”, “somewhat disagree”, “strongly disagree”. Once the relationships between the questions, or items, are visualized in a spatial map and interpreted, the method additionally allows the display of explanatory demographic variables such as age, education and gender in order to enrich the interpretation. There are invariably many non-responses as well, and this absence of a response is a potential category that also needs to be considered.

It may be interesting from a substantive point of view to focus on a particular subset of response categories, for example the categories of agreement only. Or we might want to focus on the categories of agreement and disagreement alone, excluding both the non-responses and the fence-sitting “neither agree nor disagree” responses. A further analysis of interest would be just of the non-responses by themselves, in order to understand how these are correlated between items as well as how item non-response is correlated with demographic variables. The response categories “neither agree or disagree” provide another interesting subset which could be analysed alone, to see if there are specific questions to which respondents are giving this unsure response and how the pattern of these responses is related to the demographic characteristics.

MCA is generally defined in two practically equivalent ways: either as (i) the simple correspondence analysis (CA) of the individual response data in the format of an indicator matrix, where all response categories form the columns of the indicator matrix, or (ii) the CA of all cross-tabulations concatenated in the so-called Burt matrix, a symmetric matrix which has the response categories as both rows and columns. The maps obtained by MCA are frequently overpopulated with points, making their printing difficult and interpretation complicated. There are strategies for rectifying this situation, such as not plotting points that contribute weakly to the principal axes of the map, but this would be undesirable when we are truly interested in each category point across all the questions. Furthermore, it is commonly found that the principal dimensions of MCA tell an obvious and unsurprising story about the data at hand, while the more interesting patterns are hidden in lower dimensions. Exploring further dimensions is not a simple task, since all the category points appear on and contribute to every dimension, to a greater or lesser extent. The basic problem is that the MCA map is trying to show many different types of relationships simultaneously and these relationships are not isolated to particular dimensions. While the technique does the best it can to visualize all the response categories, the maps may not be easily conducive to visualizing those relationships of particular interest to the researcher.

The methodology we expose here allows subsets of categories to be analyzed and visualized, thus focussing the map on relationships within a chosen subset, or between a subset and another subset. Thus, this approach would allow, for example, a direct analysis and interpretation of the non-responses, how they interrelate, how they relate to other response categories and to demographic variables.

We shall illustrate the methodology on a number of questions from the survey on Family and Changing Gender Roles II in the International Social Survey Programme (ISSP, 1994). We shall use the German data from this study as an example, for both (former) West and East Germany, involving a total sample of 3,291 respondents (a few respondents had to be omitted owing to missing data for the demographic variables of interest). We consider eleven questions (Table 1) related to the issue of single or dual earners in the family, mostly concerning the question of women working or not, which we shall call the substantive variables. To simplify our presentation we have combined the two response categories of agreement, "strongly agree" and "agree somewhat", into one, and similarly have combined the two corresponding categories of disagreement into one. In Table 1 we also list five demographic variables, referred to as the exogenous variables, which will be used to interpret the patterns of response found amongst the former eleven substantive variables. The raw response data of interest are thus of the form given in Table 2(a), showing the first four substantive and first two exogenous variables as examples, while Table 2(b) shows the same data coded as zero-one dummy variables in the columns of an indicator matrix. The Burt matrix corresponding to these data would be equal to the transpose of the indicator matrix multiplied by itself – part of the Burt matrix is shown in Table 2(c).

Insert Tables 1 and 2 about here

There are two possible analytical strategies in this situation: firstly, using MCA, that is CA of the indicator matrix of dummy variables corresponding to the substantive variables (or the corresponding cross-tabulations in the Burt matrix), with the categories of the exogenous categories displayed as supplementary points; or secondly, CA of the cross-tabulations of the variables with the exogenous variables, that is an analysis of several concatenated tables (see, for example, Greenacre, 1994). Here we treat the former case of MCA, which is more concerned with the interrelationships between the substantive variables, with the exogenous variables visualized *a posteriori.*

In order to motivate our approach, first consider the usual MCA map of these data in Figure 1, showing all the categories of the substantive variables and of the exogenous variables, the latter displayed as supplementary points. This result is typical of analyses of survey data such as these where non-response categories have been introduced into the analysis: the non-response categories are highly associated across questions and have separated out from the actual response categories. The latter response categories form a diagonal strip of points, with the "polar" categories of agreement (1) and disagreement (3) generally at the extremes and the unsure categories (?) in the middle, while the supplementary points (indicated by a diamond symbol, without labels) form another band of points just to the right of the response categories. In many similar examples, the non-response categories are aligned with the first principal axis and we can thus eliminate most of their effect by mapping the points with respect to the plane of the second and third dimensions. In this particular example, however, there is an added complication that the non-responses separate out diagonally in the plane, which would necessitate a rotation to "reify" the solution, an operation which is perfectly feasible in correspondence analysis but not regularly done nor incorporated into CA software.

In order to investigate the spread of points in the cloud of points in the upper left-hand side of Figure 1, we have several possible courses of action. One possible strategy would be to remove all cases which have non-responses, called "listwise deletion", but this would entail a reduction in sample size from 3291 to 2479, that is a loss of 812 cases, or 25% of the sample. A second strategy would be to try to isolate the effect of the non-responses on the first principal axis by rotating the solution appropriately, as mentioned above, and then consider dimensions from two onwards. But the non-response points will still contribute, albeit much less, to these subsequent dimensions, thereby complicating the interpretation. A third way to improve the visualization of the data would be to omit the non-response categories from the indicator matrix, and then apply CA. But then the totals of the rows of the indicator matrix are no longer equal and the profiles have different nonzero values (and different masses) depending on the number of non-missing responses.

Insert Figure 1 about here

A better solution, as we intend to show, will be provided by applying a variant of CA, called subset correspondence analysis, to the indicator matrix, or to the Burt matrix. Greenacre and Pardo (2004) showed how subset correspondence analysis can improve the interpretability of CA maps by focussing on subsets of points. Our proposal is thus to apply this subset methodology to the present case of MCA. This approach may be applied to *any* subset of the response categories, but usually a subset will consist of the same response categories across the questions: for example, one might simply want to analyse all the categories excluding non-responses, or the "agreement" categories alone or even just the non-response categories by themselves. The main idea is to analyse the subset of the original profile matrix, in this case the row profile matrix, and not re-express the profiles relative to their new totals within the subset. Furthermore, the row and column masses used in any subset are the same as those masses in the original data matrix of which a subset is being analyzed. A further benefit of this approach is that, if we partition the categories completely into mutually exclusive subsets, then we obtain a decomposition of the total inertia of the indicator matrix into parts accounted for by each subset. This decomposition of total inertia into parts is even more interesting when we think of MCA as an analysis of the Burt matrix rather than that of the indicator matrix.

2 Correspondence analysis of a subset of an indicator matrix

CA, and thus MCA too, is a particular case of weighted principal components analysis (see, for example, Greenacre, 1984, chapter 3). In this general scheme, a set of multidimensional points exists in a high-dimensional space in which distance is measured by a weighted Euclidean metric and the points themselves have differential weights, these latter weights being called masses to distinguish them from the dimension weights. A two-dimensional solution, (in general low-dimensional), is obtained by determining the closest plane to the points in terms of weighted least-squared distance, and then projecting the points onto the plane for visualization and interpretation. The original dimensions of the points can also be represented in the plane by projecting unit vectors onto the plane – these are usually depicted as arrows rather than points, since they may be considered as directions in the biplot style of joint interpretation of row and column points (Gower & Hand, 1996; Greenacre, 1993, 2004). In the context of MCA, however, when the rows represent many, often thousands, of respondents, we are generally interested in the column points only, and groups of respondents, for example age groups or social class groups represented as supplementary column points or, equivalently, by the centroids of the respondent points that fall into these groups.

The most general problem and solution is as follows. Suppose that we have a data matrix **Y** ($n\times m$), usually centred with respect to rows or columns or both. We assume that the rows represent respondents and that the columns represent variables, which in our context are categories of response. Let $\mathbf{D}_r$ ($n\times n$) and $\mathbf{D}_w$ ($m\times m$) be diagonal matrices of row masses and column weights respectively, where the masses give differentiated importance to the rows and the column weights serve to normalize the contributions of the variables in the weighted Euclidean distance function between rows. With no loss of generality the row masses are presumed to have a sum of 1. The rows of **Y** are thus presumed to be points with varying masses, given by the diagonal of $\mathbf{D}_r$, in an m-dimensional Euclidean space, structured by the inner product and metric defined by the weight matrix $\mathbf{D}_w$. The solution, a low-dimensional subspace which fits the points as closely as possible using weighted least-squares, minimizes the following function:

$$\mathrm{In}(\mathbf{Y}-\hat{\mathbf{Y}}) = \sum_{i=1}^{n} r_i\,(\mathbf{y}_i-\hat{\mathbf{y}}_i)^{\mathsf{T}}\mathbf{D}_w\,(\mathbf{y}_i-\hat{\mathbf{y}}_i) \qquad (1)$$

where $\hat{\mathbf{y}}_i$, the i-th row of $\hat{\mathbf{Y}}$, is the closest low-dimensional approximation of $\mathbf{y}_i$ (equivalently, $\hat{\mathbf{Y}}$ is the best optimal low-rank matrix approximation of **Y**). The function In(·) stands for the *inertia*, in this case the inertia of the difference between the original and approximated matrices. The *total inertia*, a measure of dispersion of the points in the full m-dimensional space, is equal to In(**Y**).

The solution can be obtained compactly and neatly using the generalized singular value decomposition (GSVD) of the matrix **Y** (see, for example, Greenacre, 1984, Appendix A). Computationally, using the ordinary SVD algorithm commonly available in software packages such as R (Venables & Smith, 2003), the steps in finding the solution are to first transform the matrix **Y** by pre- and post-multiplying by the square roots of the weighting matrices, then calculate the SVD and then post-process the solution using the inverse transformation to obtain principal and standard coordinates. The steps are summarized as follows:

1. $$\mathbf{S} = \mathbf{D}_r^{1/2}\mathbf{Y}\mathbf{D}_w^{1/2} \qquad (2)$$

2. $\mathbf{S} = \mathbf{U?V}^{\mathsf{T}}$ (3)

3. Principal coordinates of rows: $\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U?}$ (4)

4. Principal coordinates of columns: $\mathbf{G} = \mathbf{D}_w^{1/2}\mathbf{V?}$ (5)

Step 2 above is the SVD, with the (positive) singular values in descending order in the diagonal matrix D, and left and right singular vectors in the matrices **U** and **V** respectively. A two-dimensional solution, say, would use the first two columns of **F** and **G**, where the principal coordinates in **F** and **G** are the projections of the rows (respondents) and columns (variables) onto principal axes of the solution space. An alternative scaling for the columns is to plot standard coordinates, that is formula (5) without the post-multiplication by D – the points are then projections onto the principal axes of the unit vectors representing the column variables and usually depicted by vectors from the origin of the map to the points. The total inertia is the sum of squares of the singular values $d_1^2+d_2^2+\ldots$, the inertia accounted for in a two-dimensional solution is the sum of the first two terms $d_1^2+d_2^2$, while the inertia not accounted for (minimized in formula (1)) is the remainder of the sum: $d_3^2+d_4^2+\ldots$.

Regular MCA is the above procedure applied to an indicator matrix **Z** of the form illustrated by the left hand matrix in Table 1(b), that is of the Q=11 questions (variables). The matrix **Y** is this indicator matrix divided by Q (i.e., the row profile matrix), centred with respect to the averages of its columns. The averages of the columns are the column totals of **Z** divided by **Z**'s grand total nQ, where n is the number of rows (respondents), and hence are exactly the proportions of respondents giving the corresponding categories of response, divided by Q. Thus, if a particular category of response is given by 2675 of the total of 3,291 respondents (as in the case of A+, see Table 2(c)), and since there are 11 questions, then the corresponding column average is equal to the proportion of response 2675/3291 = 0.8128 divided by 11, that is 0.0739. In matrix notation:

$$\mathbf{Y} = \frac{1}{Q}\mathbf{Z} - \frac{1}{nQ}\mathbf{1}\mathbf{1}^{\mathsf{T}}\mathbf{Z} = (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathsf{T}})(\frac{1}{Q}\mathbf{Z})$$

The row masses in this case are all equal, being the row totals of $\mathbf{Z}$, all equal to Q, divided by its grand total nQ, hence the masses are all $1/n$. The column weights used (inversely) in the chi-square distance function are in the vector of column averages $(1/nQ)\mathbf{1}^\mathsf{T}\mathbf{Z}$.

We now wish to analyse and visualize a chosen subset of the indicator matrix. The subset version of simple CA of Greenacre and Pardo (2004), applied to the indicator matrix, implies that we maintain the same row and column weighting as in classical MCA described above, but the matrix to be analysed is the chosen subset of the profile matrix $\mathbf{Y}$, not of the original indicator matrix. That is, suppose that $\mathbf{H}$ is a selected subset of the columns of $\mathbf{Y}$, already centred, and that the corresponding subset of column weights (masses) is denoted by $\mathbf{h}$. Then subset MCA is defined as the principal components analysis of $\mathbf{H}$ with row masses $\mathbf{r}$ in $\mathbf{D}_r$ as before and metric defined by $\mathbf{D}_h^{-1}$ where $\mathbf{D}_h$ is the diagonal matrix of $\mathbf{h}$. Hence the subset MCA solution is obtained using steps (1)-(5) with $\mathbf{Y}$ equal to $\mathbf{H} - \mathbf{1}\mathbf{h}^\mathsf{T} = (\mathbf{I} - \mathbf{1}\mathbf{r}^\mathsf{T})\mathbf{H}$, $\mathbf{D}_r$ equal to the present $\mathbf{D}_r$ and $\mathbf{D}_w$ equal to $\mathbf{D}_h^{-1}$. The matrix (2) that is decomposed is thus:

$$\mathbf{S} = \mathbf{D}_r^{1/2}(\mathbf{I} - \mathbf{1}\mathbf{r}^\mathsf{T})\mathbf{H}\mathbf{D}_h^{-1/2} \tag{6}$$

and the row and column principal coordinates from (4) and (5) are thus:

$$\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}? \qquad\qquad \mathbf{G} = \mathbf{D}_h^{-1/2}\mathbf{V}? \tag{7}$$

All the usual numerical diagnostics (or contributions) of ordinary CA apply as before, since the total inertia, equal to the sum-of-squares of (6), can be broken down into parts corresponding to points and to principal axes, thanks to the SVD decomposition (see Greenacre, 2004).

3 Application to women's participation in labour force

Insert Figure 2 about here

Figure 2 shows the subset MCA map of the response categories agree (+), neither agree nor disagree (?) and disagree (-) for the questions A to K, thus omitting the non-responses (x). All the unsure (?) points line up on the vertical second axis of the solution, while most of the agreements and disagreements are spread horizontally along the first axis. Amongst the 11 statements, there are four that are worded in a reverse sense compared to the others: in statements A, F, G and H agreement represents a favourable attitude to women working, whereas in the other statements it is the opposite. Notice that in Figure 2 the disagreements to these four statements are on the side of the agreements to the others, which is what we would expect.

In order to interpret the agreement and disagreement categories without having to cope with the dimension of "neither…nor" responses, Figure 3 shows the subset MCA map of what we shall loosely call the "polar" responses, that is agreements (+) and disagreements (-), without the "non-polar" responses, that is without non-responses (x) and "neither…nor"s (?). The spectrum of agreements and disagreements is now easier to interpret, and it is clear that there is a closely correlated group items B, C, E and I , to a lesser extent J, along the horizontal axis which are being responded differently from F, G, H and to a lesser extent A, which are separating out vertically in the map(the solution could be slightly rotated to improve the interpretation along axes). We see now that the reversely worded statements are not always reversely answered – if they were then A+, F+, G+ and H+ should lie amongst the bunch of "-" categories of the other statements, similarly A-, F-, G- and H- should lie amongst the "+" categories of the others.

Insert Figure 3 and Table 3 about here

To investigate why A, F, G and H behave differently, we constructed a table where we counted how many times respondents had responded positively or negatively to a statement in this group and to a statement in the group formed by B, C, D, E and I which are all worded clearly in a sense unfavourable to women working (Table 3). If agreements to the first set of questions generally co-occurred with disagreements to the second set, then we would expect much larger percentages in the top right and bottom left of the table, compared to their corresponding margins. However, this table shows that whereas 38.2% of co-occurring responses are in the upper right corner of the table (agreements to A, F, G, H and disagreements to B, C, D, E, I), there are as many as 30.1% co-occurring agreements to statements from both sets. There is a similar effect in the last row: although the overall level of disagreement to A, F, G, H is lower, there are relatively many co-occurrences to disagreements to statements from both sets (4.0%) compared to the inverted responses where disagreement to statements in the first set coincides with agreement to statements in the second set (7.3%).

Insert Figure 4 about here

In spite of the separation of the A, F, G, H set from the others, for the abovementioned reason, we do see a lining up along the horizontal first axis in Figure 3 of attitudes favourable to women working on the left hand side, with unfavourable attitudes on the right. Figure 4 shows the categories of the exogenous demographic variables, which are represented as supplementary points with respect to the dimensions of Figure 3, but displayed separately. The spread is generally along the horizontal dimension, showing East Germans to be much more in favour of women working than West Germans: taking two examples from the original data, we calculated that 93.5% of East Germans think that both husband and wife should contribute to the household income (statement H), compared to 66.9% for West Germans; and 11.2% of East Germans think that the household is the wife's job (statement I), compared to 37.2% for West Germans). Also more in favour are the younger age groups and more educated groups, as well as single and divorced people, while older, less educated and the "widowed" group are less in favour. As expected, females are on average in favour of women working while men are less in favour. The interpretation using supplementary points could be enriched by indicating the positions of interactive subgroups of respondents, for example male East Germans, female East Germans, male West Germans, and male West Germans, or East Germans in different age groups and so on.

Finally, we performed a subset MCA of the non-response (x) categories alone, which we had originally omitted, and the resulting map is shown in Figure 5. All the category points are on the positive side of the first axis, so that the first axis is a dimension of overall non-responses and would be highly correlated with a count of non-responses for each respondent. However, we see an interesting dichotomy in the items, with non-responses to A, B and C clearly separated from the rest – these are exactly the items that include the word "mother" and relate to the working women's relationship with her children and family, for which there must exist a special pattern of interrelated non-responses in some respondents. The origin of the map in Figure 5 represents the average non-response point for all 11 questions. Demographic categories are also shown in Figure 5 and those to the left of centre will thus have less than average non-responses and categories to the right more than average. Hence, higher educational groups have fewer non-responses, as do East Germans compared to West Germans. Both the youngest and oldest age groups have higher than average non-responses, and so on. As in Figures 3 and 4, the contrast in attitudes reflected by the second dimension is not correlating strongly with the demographic categories.

Thanks to the fact that the margins in each subset analysis are always determined from the full data table, there is an interesting decomposition of total inertia across the subset analyses. A subset MCA of the complete indicator matrix, which is just an ordinary MCA, has a total inertia equal to $(J – Q)/Q$ = (44–11)/11 = 3, where Q = number of questions, J = total number of categories. In the analysis of the subset without the non-response categories (Figure 2), the total inertia is 2.046, while in the analysis of the non-response categories alone (Figure 5), the total inertia is 0.954. Hence in subset MCA the total inertia of all categories is decomposed into parts for each of the two mutually exclusive but exhaustive subsets. This breakdown is summarized in Table 4, including the percentages of inertia on the first two dimensions of each analysis reported previously. When the "neither…nor"s (?) are removed from the analysis of Figure 2, the total inertia for the "polar" responses (Figure 3) is equal to 1.163. From these results one can deduce that the inertia of the "neither…nor" categories is equal to 2.046–1.163 = 0.833. Thus the inertia from "non-polar responses" (? and X) is equal to 0.954+0.833 = 1.837, more than half of the inertia in the original MCA.

Insert Table 4 about here

The percentages of inertia indicated in each map but are all underestimates of the true variance explained, which is the same issue that affects the percentages of inertia in MCA. In Figure 1, where all categories are analyzed and which is thus a regular MCA, the principal inertias (eigenvalues) can be adjusted – Table 4 shows the adjusted percentages in the first row of the table. When analyzing the subset, we could still calculate the percentages of inertia explained conditional on the coordinates obtained in the subset analysis, but it is not clear whether adjusted estimates can be obtained easily from the subset MCA solution, as in MCA.

4 Subset MCA applied to the Burt matrix

The relationship between subset MCA of the indicator matrix and a subset analysis of the corresponding Burt matrix gives another perspective on the problem. As an example, suppose that we divide the question response categories into "polar responses" (PRs) and "non-polar responses" (NPRs) – as in our application, the categories "+"and "–" on the one hand, and the categories "?" and "X" on the other. Then the Burt matrix **B** can be subsetted into four parts:

	PR	NPR
PR	$\mathbf{B}_{11}$	$\mathbf{B}_{12}$
NPR	$\mathbf{B}_{21}$	$\mathbf{B}_{22}$

If λ_1, λ_2, … are the principal inertias of the indicator matrix $\mathbf{Z}$, with sum $(J–Q)/Q$ then λ_1^2, λ_2^2, … are the principal inertias of the Burt matrix **B**. Table 5 gives the total inertias and first two percentages of inertia for several subset MCAs of **B**, including the complete MCA in the first line. For example, in the complete MCA in Figure 1, the first principal inertia is 0.3143, which if squared is equal to 0.09878. This is the first principal inertia of the Burt matrix, and represents 26.0% of the total inertia 0.3797 of **B**, which checks with Table 5. This property carries over to the subset analyses as well. For example, in Table 3, the first principal inertia in the subset MCA of the PRs is 0.2718, which if squared is equal to 0.07388. Expressed as a percentage of the inertia 0.1366 of the submatrix $\mathbf{B}_{11}$ of the Burt matrix which analyses the PRs, a percentage of 54.1% is obtained, again agreeing with the percentage reported in Table 5. The connection between the principal inertias in the subset MCA of the indicator matrix and the subset MCA of the corresponding part of the Burt matrix, holds for exactly the same reason as in the complete MCA: the matrix $\mathbf{B}_{11}$ analysed in the latter case is exactly $\mathbf{Z}_1^{\mathsf{T}}\mathbf{Z}_1$, where $\mathbf{Z}_1$ is the submatrix of the indicator matrix analysed in the former case.

Insert Table 5 about here

From the above partitioning of **B** into four sub-matrices, the total inertia of **B** is equal to the sum of inertias in the subset analyses of $\mathbf{B}_{11}$, $\mathbf{B}_{22}$, $\mathbf{B}_{12}$ and $\mathbf{B}_{21}$ (notice that the last two are transposes of each other, so we could just count the inertia of one of them twice – the caption of Table 5 gives the calculation to verify this assertion). As we have just said in the previous paragraph, the subset analyses of $\mathbf{B}_{11}$ and $\mathbf{B}_{22}$ give results whose principal inertias are exactly the squares of those in the respective subset analyses of the corresponding indicator matrices of PR and NPR responses. But there is an "extra" subset analysis, namely that of $\mathbf{B}_{12}$, that is manifest in the Burt matrix but not in the indicator matrix. In our illustration, the submatrix $\mathbf{B}_{12}$ captures the associations between PRs and NPRs. In Table 5, which gives the corresponding decomposition of inertia for these subset analyses, we can see that the level of association between the PRs and NPRs is much less than within PRs and within NPRs. It should be remembered, however, that all these components of inertia are inflated by fragments of the "diagonal blocks" from the Burt matrix, as in the complete MCA.

In the case of $\mathbf{B}_{11}$ and $\mathbf{B}_{22}$, these are inflated by subsets of the diagonal matrices on the diagonal of $\mathbf{B}$. In the case of $\mathbf{B}_{12}$ or $\mathbf{B}_{21}$, the elements of the matrix corresponding to the same question account for the inflation, and consist of blocks of zeros since there is zero contingency between the PRs and NPRs of the same question. It is a question of continuing research if there exist simple ways for adjusting the eigenvalues and their percentages, as is possible in the MCA of the complete Burt matrix.

4. Discussion and conclusions

One of Benzécri's (1973) basic principles of *Analyse des Données* (Data Analysis) is that one should analyse all the available information, a principle which implies that every possible category of response, including missing responses, be analysed together. In the case of MCA this means analysing the so-called *"tableau disjonctif complet"* (complete disjunctive table, or indicator matrix) which has as many ones in each row as there are variables indicating the categories of response. When analysing several variables, however, it is almost always the case that the interpretation is hampered by the large number of category points in the map, all of which load to a greater or lesser extent on every dimension, so that interpretation and conclusions are limited to broad generalities. We have shown that there is great value in restricting the analysis to subsets of categories, which may be visualized separately and thus with better quality than they would have been the case in a complete MCA. The method allows exploration of several issues of prime importance to social scientists:

- analysing substantive responses only, ignoring non-responses,
- studying the pattern of non-responses by themselves and how they relate to demographic variables,
- focusing on the role played by neutral responses, how they are related to one another, how they are related to the non-responses and whether any patterns correlate with the demographic variables.

In the first visualization of the data (Figure 1), that is the complete MCA, the points representing the missing data were so strongly associated that they forced all the other categories into a group on the opposite side of the first axis. Even though the frequency of non-responses was fairly low, they dominate this map, leaving little remaining "space" to understand the relationships between the other responses categories. The subset analysis allowed this effect to be removed, showing the separation of the "non-missing" response categories more clearly in Figure 2. The neutral categories could also be removed (Figure 3) to further clarify the associations between the agreement and disagreement poles of questions that were worded in a favourable and unfavourable direction towards working women. The subset could also consist of just one response category across the questions, as illustrated by the map in Figure 5, which showed the missing data categories only. In all cases, supplementary points could be added to show the relationship between the analysed response categories and the demographic variables (Figures 4 and 5).

The subset variant of simple CA has been extended here to MCA and maintains the geometry of the masses and chi-square distances of the complete MCA, the only difference being that we do not re-express the elements of the subset with respect to their own totals, but maintain their profile values with respect to the totals of the complete data set. This approach ensures the attractive property that the total inertia is decomposed into parts for each of the subsets of categories. The same idea has already been used in an MCA context by Gifi (1990) to exclude non-responses, where the approach is called "missing data passive" (see also Michailidis and de Leeuw, 1998), and similarly by Le Roux and Rouanet (2004). These uses are limited, however, to the exclusion of missing data, whereas in our application we consider a much wider number of possible subsets, including the analysis of missing data alone when all the substantive responses are excluded.

Software notes

Programs for CA, MCA, subset CA and subset MCA are available as functions in the R language (www.r-project.org) as well as in XLSTAT (www.xlstat.com). The first author can be contacted for further information about this software.

References

Benzécri, J.-P. (1973). *Analyse des Données. Tôme 2: Analyse des Correspondances*. Paris: Dunod.

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

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

Greenacre, M.J. (1993). *Correspondence Analysis in Practice*. London: Academic Press.

Greenacre, M.J. (1994). Multiple and joint correspondence analysis. In *Correspondence Analysis in the Social Sciences* (eds. M.J.Greenacre & J. Blasius), pp.141 -161. London: Academic Press.

Greenacre, M.J. (2004). Weighted metric multidimensional scaling. Paper presented at *German Classification Society Meeting*, Dortmund, March 2004. Working Paper 777, Dept Economics and Business, Universitat Pompeu Fabra, Barcelona. . `http://www.econ.upf.es/eng/research/onepaper.php?id=777`

Greenacre, M.J. and Blasius, J. (1994). *Correspondence Analysis in the Social Sciences.* London: Academic Press.

Greenacre, M.J. and Pardo, R. (2004). Subset correspondence analysis: visualization of relationships among a selected set of categories from a questionnaire survey. Working Paper 791, Dept Economics and Business, Universitat Pompeu Fabra, Barcelona. . `http://www.econ.upf.es/eng/research/onepaper.php?id=791`

ISSP (1994). Family and Changing Gender Roles II. International Social Survey Program. University of Cologne: Zentralarchiv für Empirische Scozilaforschung.

Le Roux, B. and Rouanet, H. (2004). *Geometric Data Analysis: From Correspondence Analysis to Structured Data.* Dordrecht: Kluwer.

Michailidis, G. and de Leeuw, J. (1998). The Gifi system for descriptive multivariate analysis. *Statistical Science*, **13**, 307-336.

Venables, W.N. and Smith, D.M. (2003). *An Introduction to* R. `www.r-project.org`.

Table 1

List of variables used in this study, taken from the survey on Family and Changing Gender Roles II as part of the International Social Survey Program (ISSP, 1994).

A : A working mother can establish just as warm and secure a relationship with her children As a mother who does not work.

B : A pre-school child is likely to suffer if his or her mother works

C : All in all, family life suffers when the woman has a full-time job

D : A job is all right, but what most women really want is a home and children.

E : Being a housewife is just as fulfilling as working for pay.

F : Having a job is the best way for a woman to be an independent person.

G : Most women have to work these days to support their families.

H : Both the man and woman should contribute to the household income.

I : A man's job is to earn money; a woman's job is to look after the home and family

J : It is not good if the man stays at home and cares for and the woman goes out to work.

K : Family life often suffers because men concentrate too much on their work.

Response scales for each question: strongly agree, agree somewhat, neither agree nor disagree, disagree somewhat, strongly disagree; in our application we have merged the two categories of agreement into one, and the two categories of disagreement into one

Exogenous variables:

German region	2 regions: DW (West Germany), DE (East Germany)
Sex	2 categories: M, F
Age	6 groups: A1 (up to 25), A2 (26-35), A3 (36-45) A4 (46-55), A5 (56-65), A6 (66 and over)
Marital status	5 groups: MA (married), WI (widowed), DI (divorced), SE (separated), SI (single)
Education	7 groups: E0 (none), E1 (incomplete primary), E2 (primary), E3 (incomplete secondary), E4 (secondary), E5 (incomplete tertiary), E6 (tertiary)

Table 2

Raw data in two different but equivalent forms: (a) the original response pattern data for the first four questions and the first two exogenous variables region and sex; (b) the indicator (dummy variable) form of coding. Response categories for questions A – K are: 1. strongly agree or agree combined (+) , 2. neither agree nor disagree (?), 3. disagree or strongly disagree combined (-), 4. non-response (x); for region: 1. former West Germany (DW), 2. former East Germany (DE); for sex: 1. male (M), 2. female (F). Data is shown only for first six respondents (out of *n* = 3291). (c) A part of the Burt matrix, showing some cross-tabulations with question A only.

(a)

	A	B	C	D	...	Reg	Sex	...
	1	1	1	1	...	1	1	...
	3	1	1	3	...	1	1	...
	1	1	1	1	...	1	1	...
	1	3	3	3	...	1	2	...
3291	1	2	2	3	...	1	2	...
cases	1	1	3	2	...	1	1	...
	.	.	.	.	...	.	.	...
	.	.	.	.	...	.	.	...
	.	.	.	.	...	.	.	...
	.	.	.	.	...	.	.	...

(b)

A				B				C				D				Region		Sex		...
+	?	–	x	+	?	–	x	+	?	–	x	+	?	–	x	DW	DE	M	F	
1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	1	0	...
0	0	1	0	1	0	0	0	1	0	0	0	0	0	1	0	1	0	1	0	...
1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	1	0	...
1	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	1	0	0	1	...
1	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	1	0	0	1	...
1	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	1	0	1	0	...
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	...
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	...
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	...
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	...

(c)

		A				B				...	Region		...
		+	?	–	x	+	?	–	x		DW	DE	
	+	2675	0	0	0	1328	374	901	72	...	1685	990	...
	?	0	111	0	0	85	17	8	1	...	92	19	...
A	-	0	0	525	0	472	13	31	9	...	461	64	...
	x	0	0	0	110	61	3	4	42	...	86	24	...
	.	.	.	.	.	.	.	.	.	...	.	.	...
	.	.	.	.	.	.	.	.	.	...	.	.	...
	.	.	.	.	.	.	.	.	.	...	.	.	...
	.	.	.	.	.	.	.	.	.	...	.	.	...

Table 3

Percentages of co-occurrences of agreements (+), unsures (?) and disagreements (-) between the two sets of statements with opposite wording with respect to women working: A, F, G and H are favourable to women working, while B, C, D, E and I are unfavourable. Non-responses have been omitted (using pairwise deletion) when compiling these percentages.

		B, C, D, E, I		
		+	?	-
A, F, G, H	+	30.1%	10.6%	38.2%
	?	3.9%	1.7%	2.9%
	-	7.3%	1.2%	4.0%

Table 4

Total inertias of different subsets of categories, and the percentages of inertia along the first two dimensions of the analyses reported in Figures 1 to 4. For the first analysis of all the categories (MCA) the adjusted percentages are given in parentheses.

Subset analysed	***Total inertia***	***Percentages***	
		Axis 1	***Axis 2***
all response categories (Figure 1) +,?,–,x	3.000	10.4% *(50.5%)*	9.1% *(34.3%)*
without non-reponses (Figure 2) +,?,–	2.047	13.7%	8.0%
without NPR's (Figure 3) +,–	1.165	23.4%	10.7%
non-responses (Figure 5) x	0.953	30.1%	9.3%

Table 5

Total inertias of different subsets of categories in the analysis of the submatrices of the Burt matrix, and the percentages of inertia along the first two dimensions. It is readily checked that 0.3797=0.1366+0.2077+2(0.0177), as described in the text, corresponding to the inertias of the submatrices of the Burt matrix. For the first analysis of all the categories (MCA) the adjusted percentages are given in parentheses, as in Table 4.

Subset analysed	***Total inertia***	***Percentages***	
		Axis 1	***Axis 2***
all response categories +,?,–,x	0.3797	26.0% *(50.5%)*	19.5% *(34.3%)*
polar reponses (PRs) +,–	0.1366	54.1%	11.2%
non-polar responses (NPRs) ?,x	0.2077	41.9%	9.0%
PRs by NPRs (Figure 6)	0.0177	35.1%	15.7%

Figure 1

MCA map of Table 3, showing the four response categories for each of the 11 questions A to K (see Table 1). The unlabelled points with diamond symbols are the supplementary points for the exogenous variables.

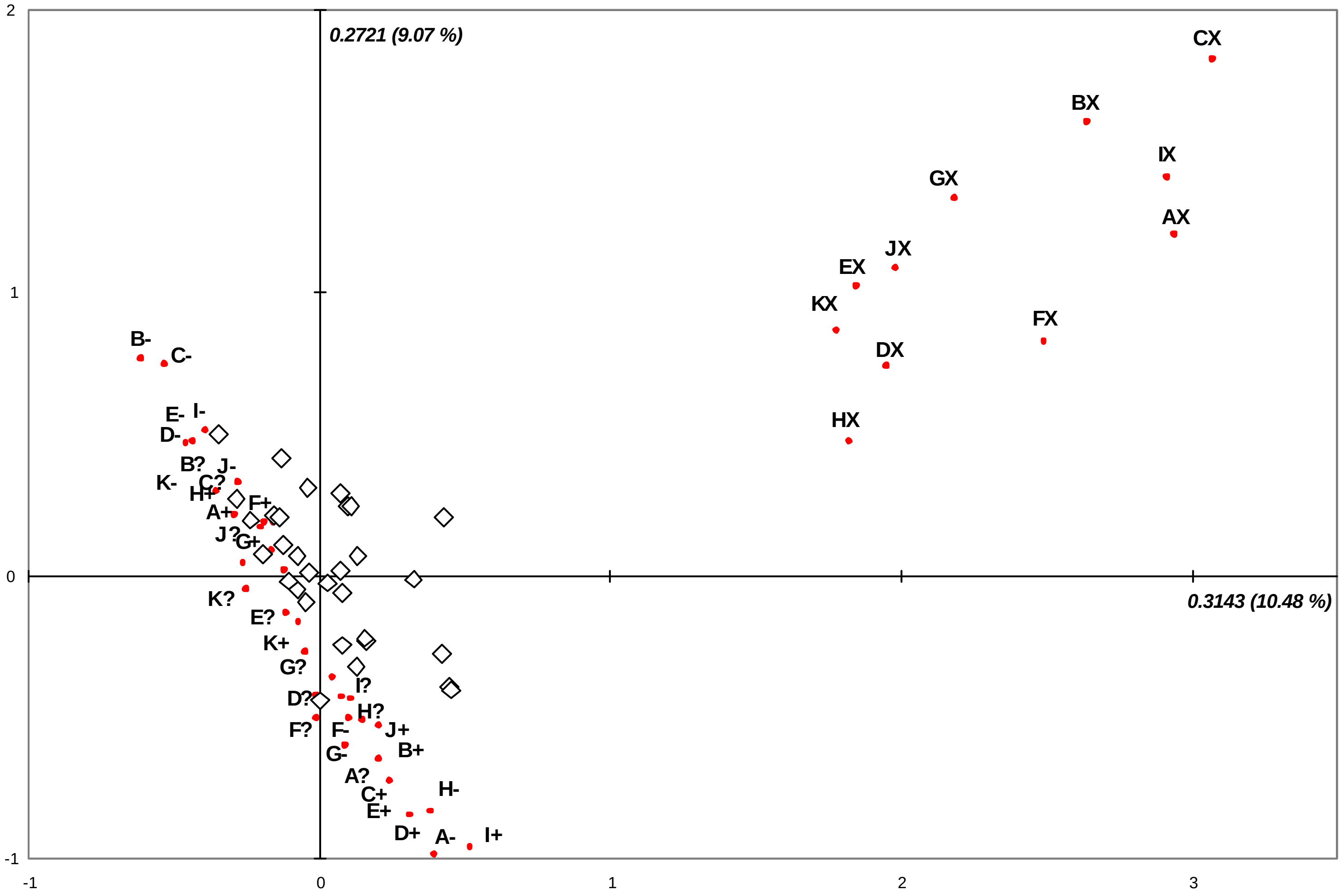


Figure 2

Subset MCA map of the response categories omitting the non-response categories.

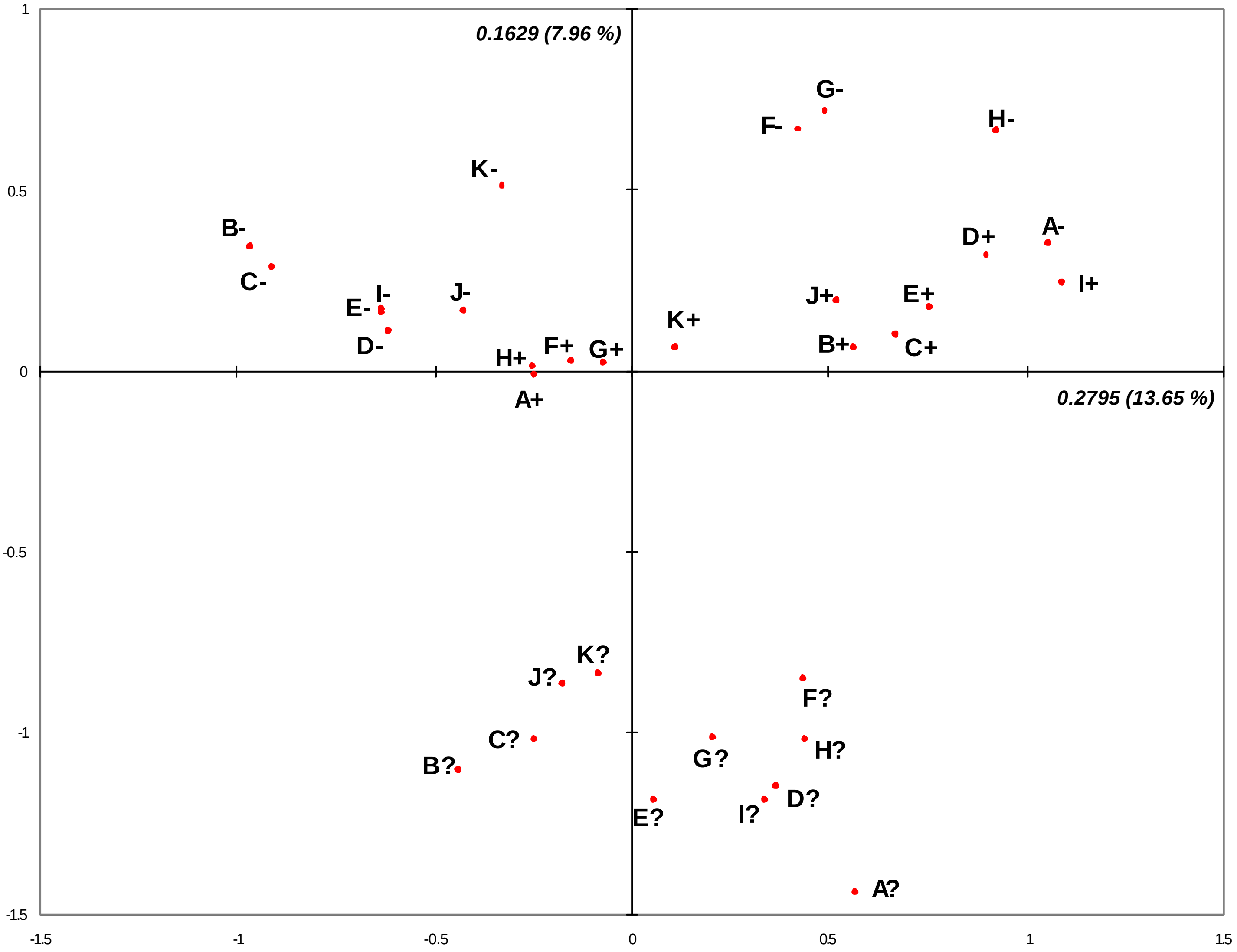


Figure 3

Subset MCA map of the agreement and disagreement categories only (A+, A- to K+, K-), without NSRs ("neither agree nor disagree" and non-responses).

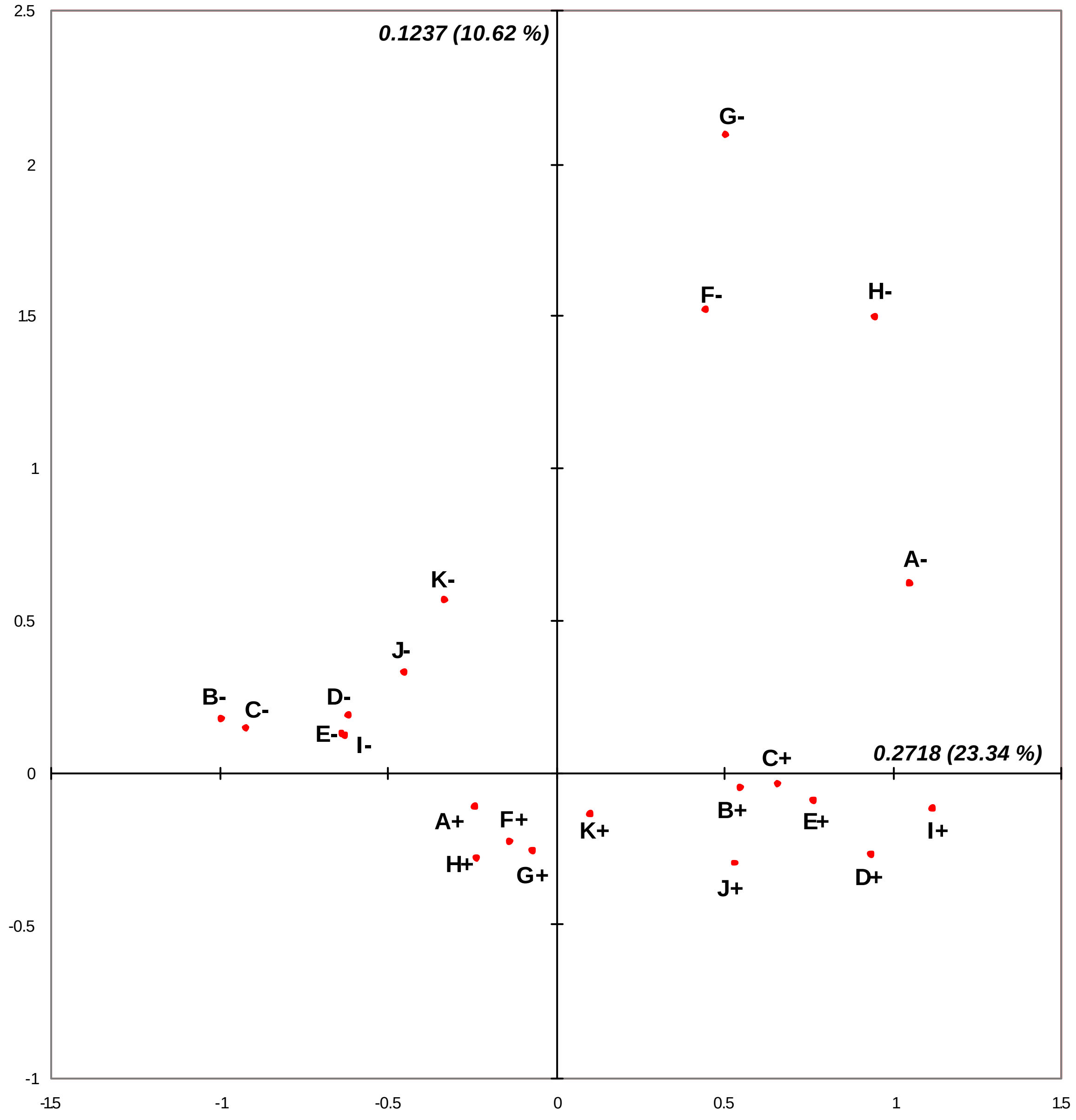


Figure 4

Positions of the supplementary points in the map of Figure 3.

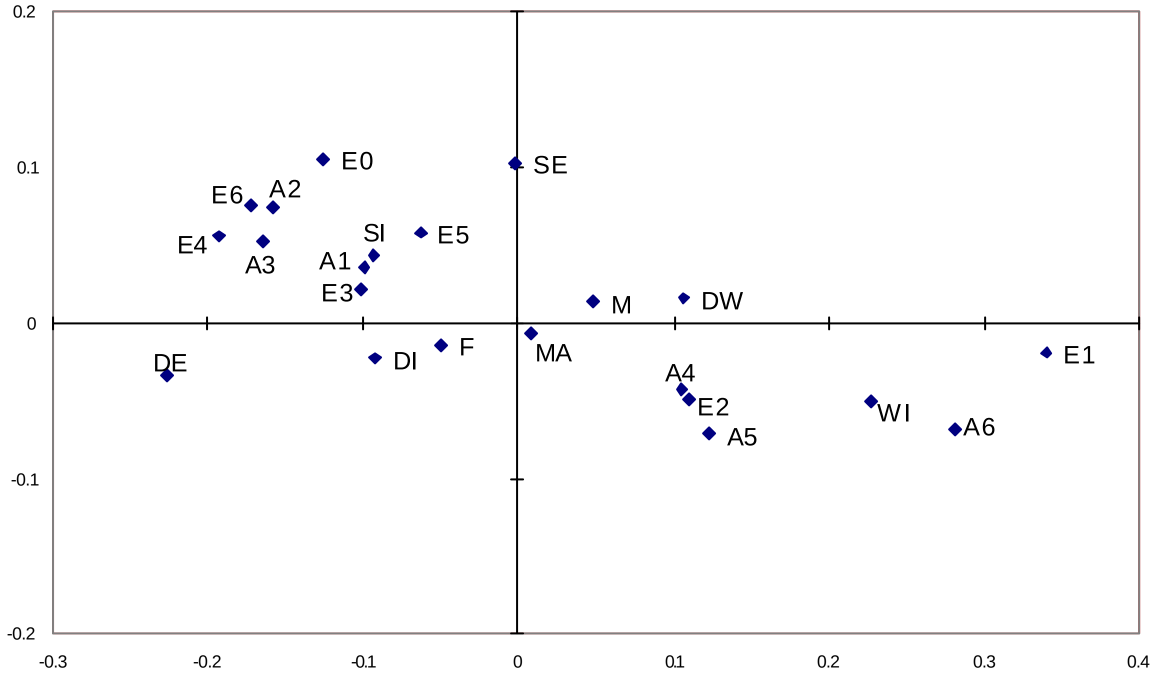
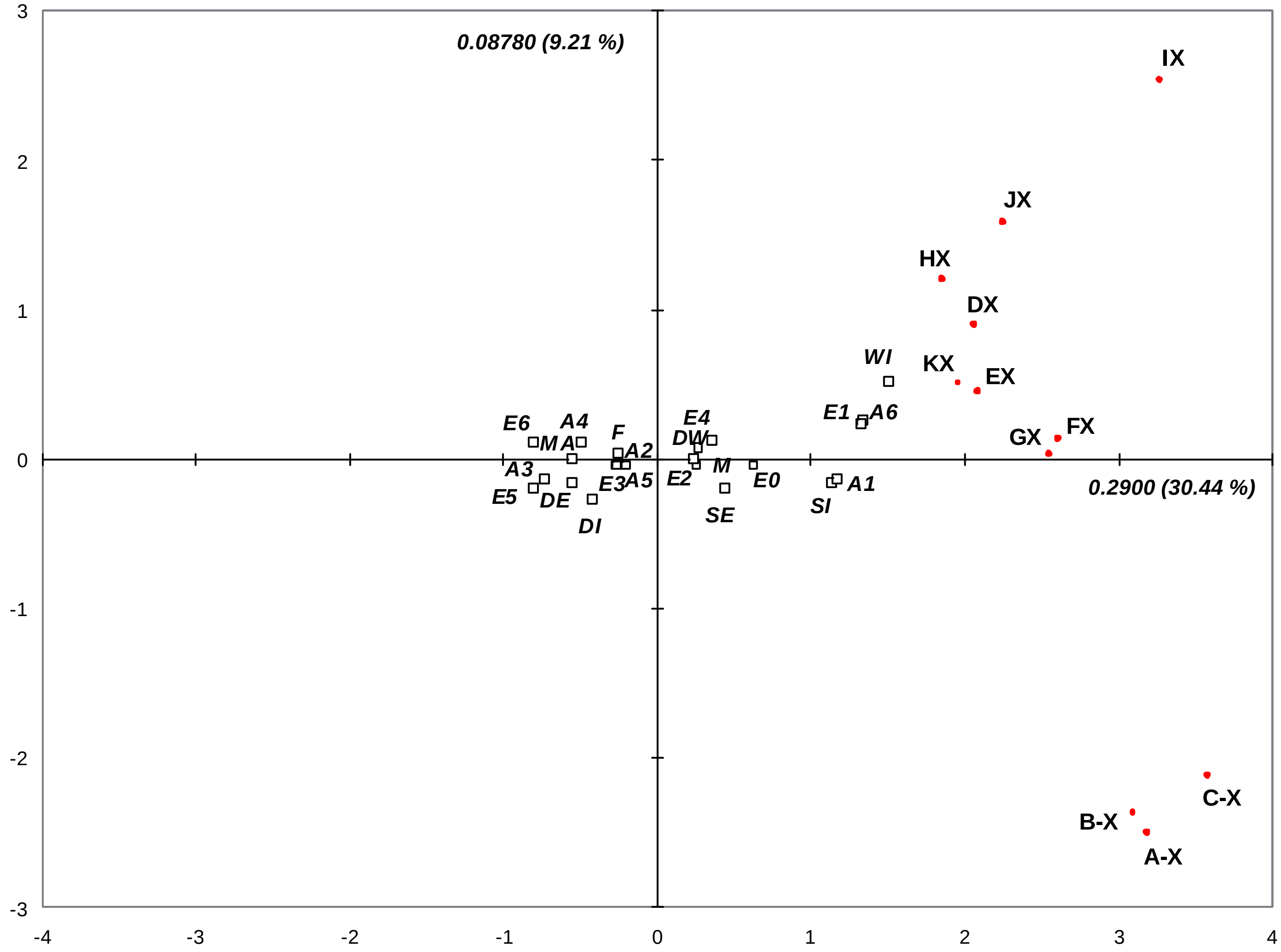


Figure 5

Subset MCA map of the non-response categories only (AX to KX), showing supplementary demographic categories


0.08780 (9.21 %)
0.2900 (30.44 %)
IX
JX
HX
DX
WI
KX
EX
E1
A6
FX
GX
E6
A4
F
E4
DW
MA
A2
A3
M
E3
A5
E2
E0
A1
E5
DE
SI
SE
DI
C-X
B-X
A-X