Subset Correspondence Analysis: Visualizing Relationships Among a Selected Set of Response Categories from a Questionnaire Survey

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

(corresponding author)

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

E-mail: rpardoa@fbbva.es

Acknowledgments: The authors acknowledge research support by the BBVA Foundation (Fundación BBVA).

Subset Correspondence Analysis: Visualizing Relationships Among a Selected Set of Response Categories from a Questionnaire Survey

Abstract: It is shown how correspondence analysis may be applied to a subset of response categories from a questionnaire survey, for example the subset of undecided responses or the subset of responses for a particular category. The idea is to maintain the original relative frequencies of the categories and not re-express them relative to totals within the subset, as would normally be done in a regular correspondence analysis of the subset. Furthermore, the masses and chi-square metric assigned to the data subset are the same as those in the correspondence analysis of the whole data set. This variant of the method, called Subset Correspondence Analysis, is illustrated on data from the ISSP survey on Family and Changing Gender Roles.

Keywords: categorical data, correspondence analysis, questionnaire survey.

1. Introduction

Simple correspondence analysis (CA) and multiple correspondence analysis (MCA) are statistical methods which are particularly useful for exploring relations amongst a large set of categorical variables. The associations between the categories of the variables are visualized in a spatial map, allowing interpretations of their similarities and differences in order to provide empirical indicators which can lead to formulations of hypotheses, more formal parametric analyses or even the redesign of the data-gathering instruments (for example, questionnaires).

It is common practice in CA to include all the categories of the variables under consideration in the analysis, since this gives the most comprehensive and global view of their interrelationships. However, for substantive as well as methodological reasons it is likely that after a global view of the data, it would be interesting to focus attention on a reduced set of response categories. Suppose, for example, we have a response scale with categories "agree", "neither agree nor disagree", "disagree" and "don't know", with the possibility of missing data as well. Then we might want to analyse the categories of "agreement" only, or all "substantive" response categories, that is excluding the "non-substantive responses" (NSRs, such as "don't know", the missing response and also, in some cases, the neutral "neither agree nor disagree"). It may even be interesting to study the NSRs by themselves, that is excluding all the "substantive" response categories.

Another reason for restricting the categories analysed to a subset is when there are many variables and thus many category points in the graphical displays. It is true that in situations where the variables of interest and their categories are few, the maps resulting from CA and MCA and all the associated numerical diagnostics maps are easily interpretable. On the contrary, when it is a question of exploring many variables, where we wish to preserve their individual character rather than collapse them into summated scales, and where we furthermore wish to explore their associations with a large number of socio-demographic variables, the resulting maps are saturated with points and are difficult to interpret. In this situation we can seldom proceed beyond an interpretation of broad generalities which are usually expected anyway and not surprising, and even if we try to interpret more dimensions of the solution there are so many points contributing to each dimension that only vague conclusions are possible. So for this practical reason too, it would be interesting to be able to

analyse reduced sets of categories, to facilitate the interpretation as well as to make the conclusions substantively richer and more interesting.

Since CA and MCA inherently assume that we are analysing the full set of categories for each variable, we shall demonstrate how these methods should be adapted to cope with the analysis of a subset of categories. As an application of our proposal, we consider the following data taken from the International Social Survey Programme (ISSP) survey of family and changing gender roles in 1994, involving a total sample of 33,123 respondents and conducted in 24 countries (former East and West Germany are still considered separately here). We shall focus on four questions, listed in Table 1, which capture attitudes towards a crucial facet of women's role at the end of the 20th century, namely their participation in work outside the home environment. Even though the number of questions is quite small, the benefits of looking at subsets of response categories are still evident. The questions specifically ask whether married women should work or stay at home at four different points of time in their married lives: (1) before having a child, (2) with a preschool child, (3) when the youngest child is at school, and (4) when the children have left home. The possible responses in each case are "work full-time", "work part-time", "stay at home", or "unsure/don't know". A few non-responses are also observed, which we grouped with the category of "unsure" to form a category of "non-substantive response" (NSR). In addition to the responses to these four questions, we have data on several categorical variables for each respondent: sex, age, marital status, education, social class and country – Table 1 lists all six of these exogenous variables and their respective categories which we wish to relate to the attitudes about women's participation in the labour market. The raw data of interest are thus of the form given in Table 2(a), while Table 2(b) shows the equivalent data where the response categories are coded as zero-one dummy variables in the columns of an indicator matrix.

Insert Tables 1 and 2 about here

There are two possible strategies to visualize the relationship amongst the attitudinal categories and how these are related to the exogenous variables: firstly, multiple correspondence analysis (MCA), that is CA of the indicator matrix of dummy variables, with the categories of the exogenous categories displayed as so-called "supplementary points" in the map (see, for example, Greenacre, 1993, chapter 11); or secondly, CA of the crosstabulations of the variables with the exogenous variables, where these crosstabulations are concatenated in a super-matrix (Greenacre, 1994). In the former case, the associations

amongst the attitudes are displayed and then the exogenous variables are related *a posteriori* to these associations. We have chosen to illustrate our approach using the latter strategy, which more directly relates the attitudes to the exogenous variables. We shall comment further on the MCA approach in the discussion in Section 5.

In Section 2 we describe the CA of the complete set of data, showing the problems that result and motivating the need for a adaptation of the method for a subset of points. As we shall demonstrate, it is not appropriate to simply apply CA to the submatrix of data on which we want to concentrate. In Section 3 we outline the technical features and properties of the new methodology, which we call *subset correspondence analysis.* Its application to the subset of "stay at home" responses is given in Section 4, and Section 5 concludes with a discussion.

2. Correspondence analysis of the complete matrix

Table 3 shows part of the data matrix of interest, the crosstabulations of the attitudes with the first two exogenous variables concatenated row-wise and column-wise. CA applied to this super-matrix of tables leads to a map where the row and column categories are depicted as points (Figure 1). This map is typical of the results obtained when analysing survey data such as these, and also for MCA for that matter: the response categories form a horseshoe curve from left to right, with the "work full-time" categories towards top left, the "work part-time" categories in the lower central part of the map, and the "stay at home" categories towards top right. The NSRs are close to the origin, slightly to the left, and in fact differentiate themselves more on the third dimension of the solution, not shown here. Thus the first dimension is an overall dimension which ordinates the respondents and their various biographical categories from "liberal" on the left to "conservative/traditional" on the right in terms of this issue, with groups on the left favourable to working women and those on the right unfavourable. The second dimension would line up the groups in terms of their "polarization" on this issue, for example Spain is the highest positive on the second dimension because it has higher than average frequencies of response both in favour of women working full-time and staying at home, while West Germany is very low down on this axis showing less than average of these extreme responses and more than average in favour of women working part-time.

Insert Table 3 and Figure 1 about here

None of these results are particularly surprising: the first dimension which is similar to the summated rating scale, the quadratic effect on the second dimension and the categories of missings coming together along the third dimension. It is difficult to make more specific interpretations in maps such as Figure 1. For example, consider the responses "stay at home" for the four questions. In Figure 1 we can see that "stay at home before first child" and "stay at home when children have left home" are at the extreme conservative end of the scale, whereas "stay at home with preschool child" and "stay at home when youngest child at school" are not seen as so conservative, particularly the latter. This ordering is very much what one would expect and is not very surprising. The association between these four categories and the countries, for example, is only shown within the broad dimensions of the map, which has been determined by all the response categories. We can see the more traditional countries on the right and the more liberal countries on the left, but the specific relationship of the countries with the "stay at home" responses is not so clear. As we shall show later (Section 4), there is a very interesting variation of "stay at home" responses within the conservative countries which is not represented in the above map. Neither will it help us to look at further axes, from the fourth axis onwards, since there are so many points contributing and being mapped to each axis that the results are really quite confused.

Therefore, having seen the overall spread of the countries and other groups in terms of all these categories, the question is how we can focus on one type of response, for example the traditional attitude of "stay at home", and compare the countries, age groups, and so on, just on these responses. Thus we would like to construct maps restricted to a subset of responses, whatever that subset might be. Using as an example the subset of "stay at home" responses, we might be tempted simply to omit all other columns in the data matrix of Table 3 and apply CA to the subset of "stay at home" columns. This would be undesirable for two important reasons. First, and more importantly, CA would express the frequencies of "stay at home" across the four questions in each row relative to the total frequency of "stay at home" in that row. This means that if Swedes generally give low frequencies of "stay at home" compared to Poles (which is actually the case), this fact would be lost in the calculation of the "stay at home" profiles for Sweden and Poland. What we visualize in the analysis would be the "shape" of the pattern of responses, that is where the peaks and troughs are in the profile, losing the effect of "size" or overall level of "stay at home" response. Second, the masses allocated to the row categories would be proportional to the frequency of "stay at home" responses across all four questions, not to the sample sizes associated with the categories: for

example, Swedes would be weighted much less than Poles rather than proportional to their actual sample sizes in the survey.

We shall show that it is possible to avoid both of these drawbacks by introducing a simple variant into the CA procedure. In fact, the novelty is to suppress the automatic feature of CA to calculate profiles of the given data. Specifically, we use the profiles and masses of the full data matrix and select the subset of response categories from the profile matrix, without further re-expression of the profiles with respect to their new totals within the subset. With this simple modification of the CA algorithm, the procedure is completely satisfactory and resolves both difficulties described above.

3. Correspondence analysis of a subset of a data matrix

CA 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, 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).

The most general solution is as follows. Suppose that we have a data matrix $\mathbf{Y}$ ($n \times m$), usually pre-centred with respect to rows or columns or both. We assume that the rows represent sampling units such as respondents or groups of 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 $\mathbf{Y}$ are thus presumed to be points with varying masses, given in $\mathbf{D}_r$, 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:

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

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 $\mathbf{Y}$**).** 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($\mathbf{Y}$).

As is well-known (see, for example, Greenacre, 1984, Appendix), the solution can be obtained neatly using the generalized singular value decomposition (GSVD) of the matrix $\mathbf{Y}$. 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 $\mathbf{Y}$ 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. $\mathbf{S} = \mathbf{D}_r^{1/2} \mathbf{Y} \mathbf{D}_w^{1/2}$ (2)

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

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

4. Principal axes of rows: $\mathbf{D}_w^{-1/2} \mathbf{V}$ (5)

5. Standard coordinates of columns: $\mathbf{G} = \mathbf{D}_w^{1/2} \mathbf{V}$ (6)

Notice that the row coordinates are scaled such that $\mathbf{F}^{\mathsf{T}} \mathbf{D}_r \mathbf{F} = \mathbf{D}_a^2$, that is the weighted sum-of-squares of the row points (i.e., their inertia) on the *k*-th dimension is equal to a_k^2, secondly that the principal axes are orthonormal basis vectors in the metric $\mathbf{D}_w$:

$(\mathbf{D}_w^{-1/2}\mathbf{V})^{\mathsf{T}}\mathbf{D}_w(\mathbf{D}_w^{-1/2}\mathbf{V}) = \mathbf{I}$, and thirdly that the standard coordinates are the projections of the unit vectors in the row space (i.e., the original basis vectors in the identity matrix **I**) onto the principal axes, where the projections are in the metric $\mathbf{D}_w$: $\mathbf{G} = \mathbf{I}\mathbf{D}_w\mathbf{D}_w^{-1/2}\mathbf{V} = \mathbf{D}_w^{1/2}\mathbf{V}$.

A two-dimensional solution, say, would use the first two columns of **F** and **G**. As mentioned before, the columns (variables) are conventionally depicted by arrows and the rows (respondents or groups of respondents) by points. Furthermore, the different scales of the two sets of coordinates often necessitate having a different scale for the row and column points, but respecting the *aspect ratio* in each case, i.e. a scale unit on the horizontal axis should be equal to a scale unit on the vertical axis. The total inertia is the sum of squares of the singular values $\alpha_1^2+\alpha_2^2+\ldots$, the inertia accounted for in a two-dimensional solution is the sum of the first two terms $\alpha_1^2+\alpha_2^2$, while the inertia not accounted for (minimized in formula (1)) is the remainder of the sum: $\alpha_3^2+\alpha_4^2+\ldots$.

Now ordinary CA is the above procedure applied to a table of frequencies, usually a two-way contingency table **N**, as follows. First divide **N** by its grand total n to obtain the correspondence matrix $\mathbf{P} = (1/n)\,\mathbf{N}$. Let the row and column marginal totals of **P** be the vectors **r** and **c** respectively, and $\mathbf{D}_r$ and $\mathbf{D}_c$ be the diagonal matrices of these vectors. Thinking of the table as a set of rows (an identical argument applies if we think of it as columns), calculate the row profiles by dividing the rows of **P** by their row totals: $\mathbf{D}_r^{-1}\mathbf{P}$. Then CA is a weighted principal components analysis of the row profiles in $\mathbf{D}_r^{-1}\mathbf{P}$, where distances between profiles are measured by the so-called chi-squared metric defined by $\mathbf{D}_c^{-1}$ and the profiles are weighted by the row masses in $\mathbf{D}_r$. Since the centroid of the row profiles turns out to be exactly the vector $\mathbf{c}^{\mathsf{T}}$ of marginal column totals, the solution is given by (1)-(5) above with the centred **Y** equal to $\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}^{\mathsf{T}}$, $\mathbf{D}_r$ equal to the present $\mathbf{D}_r$ and $\mathbf{D}_w$ equal to $\mathbf{D}_c^{-1}$.

We now wish to apply the above theory to a subset of the table, maintaining the same row and column weighting as in classical CA described above, but applied to a *subset of the profiles* rather than a subset of the original frequency table, hence avoiding the recalculation of profiles for the selected subset. That is, suppose that **H** is a selected subset of the columns of $\mathbf{D}_r^{-1}\mathbf{P}$ and that the corresponding subset of the column totals **c** is denoted by **h**, that is (as

in ordinary CA) $\mathbf{h}$ is the weighted average of the rows of $\mathbf{H}$: $\mathbf{H}^\mathsf{T}\mathbf{r} = \mathbf{h}$. Then subset correspondence analysis (abbreviated as s-CA) is defined as the weighted 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 s-CA solution is obtained using (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) which is decomposed is thus:

$$\mathbf{S} = \mathbf{D}_r^{1/2}(\mathbf{I} - \mathbf{1}\mathbf{r}^\mathsf{T})\mathbf{H}\mathbf{D}_h^{-1/2} \tag{7}$$

and the biplot decomposition using the row principal and column standard coordinates from (4) and (6) is thus:

$$\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_a \qquad\qquad \mathbf{G} = \mathbf{D}_h^{-1/2}\mathbf{V} \tag{8}$$

In the special case of CA and MCA the alternative biplot scaling $\mathbf{D}_h\mathbf{G} = \mathbf{D}_h^{1/2}\mathbf{V}$ for the columns, proposed by Gabriel and Odoroff (1990), is particularly useful since it reduces the length of the column vectors proportionally to their respective masses, giving shorter vectors for the rare categories which are usually in outlying positions in the CA map. All the usual numerical diagnostics (or contributions) of ordinary CA apply as before, since the total inertia, equal to the sum-of-squares of (7), can be broken down into parts corresponding to points and to principal axes, thanks to the SVD decomposition (see Greenacre, 2004).

4. Application to attitudes about women in labour market

Table 4 shows the subset of the profile matrix which we wish to analyze, the profile values for just the "stay at home" response categories for the four questions. Also shown are the masses assigned to each row point and the (weighted) average profile values used in the chi-square distance between rows. Figure 2 shows the s-CA map of the data of Table 4.

Insert Table 4 and Figure 2 about here

The first feature to notice here and the main difference to ordinary CA is that the column points, shown here as vectors, are not necessarily centred. In fact, the first (horizontal) dimension has all the vectors pointing to the positive side, indicating a dimension of overall "size" or level of response of the "stay at home" categories to the four questions.

The lining up of the row points, that is the countries, age groups, etc., along this axis will be quite similar to that of the first axis of Figure 1, since they reflect the same liberal–traditional dimension. The directions of the four arrows in Figure 2, however, give us a much better appreciation of the differences between the four questions. The "stay at home" responses to questions 1 and 4 are highly correlated, but both much less correlated with the response to question 3, with response to question 2 between these extremes. This shows features which we could not see in Figure 1: for example, the Philippines is clearly the most conservative overall, being most positive on dimension 1 (average percentage "stay at home" response to the four questions = 34.5%) but in terms of question 3 (for women with preschool child) they are not by any means the most conservative – Philippines has a "stay at home" response percentage for this question of 39.0% whereas Poland, Hungary, Bulgaria and Russia have percentages of 69.7%, 62.7%, 56.7% and 56.1% respectively (these latter countries have average percentages of "stay at home" response over the four questions of 33.3%, 27.2%, 24.3%, and 25.7% respectively – see Table 4). In fact, the average percentage response over all countries for question 3 is 48.6%, so the Philippines is actually below average. In Figure 2 Poland, Hungary, Bulgaria and Russia all project more positively than Philippines onto the direction defined by the vector VAR3, verifying the data. The map thus shows differences in "shape" for the "stay at home" responses which were impossible to see in Figure 1, since that map was dominated by the broad dimensions of size and polarisation mentioned previously (Section 2). A similar feature at the liberal end of dimension 1 is observed for Israel, for example. Israel is the third most liberal country (see Table 4) but on questions 1 and 4 it is much closer to average than one might expect: for these questions its percentages are 3.2% and 5.3% respectively, which are close to median values in both cases, with 12 countries having smaller percentages than 3.2% for question 1, and 13 countries having percentages lower than 5.3% for question 4.

The countries show much more dispersion than the other exogenous variables, but these variables also show interesting patterns. For example, the male-female difference is, as might be expected, that males have a more traditional attitude than females, but as we see in Figure 2 the differences are greater in the direction of variables 1 and 4 than for the other two variables. This can also be verified in the data: for example, the ratio of differences between males and females for question 1 is 0.0802/0.0467 = 1.71, whereas it is 0.2060/0.1568 = 1.31 for question 2.

4. Discussion and conclusions

Correspondence analysis is primarily applicable for the analysis of contingency tables or other frequency tables where relative frequencies in rows or columns are visualized as points, while using the marginal frequencies as point masses or as estimates of variance to define a normalized distance function, the chi-square distance, between points. One of Benzécri's (1973) basic principles of *Analyse des Données* (Data Analysis) is that one should analyse the full extent of available information, a principle which implies that every possible category of response, including missing responses, be analysed together. When analysing several variables, however, it is almost always the case that the interpretation is obscured 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. For example, one might find that the categories of missing response generally separate from other categories along a particular dimension, but at the same time all the other response categories also contribute in varying amounts to this dimension so that it is not easy to make specific conclusions about the pattern of missing data. Once the broad picture is seen in the complete analysis, there is value in focusing on different subsets of categories, thus simplifying the maps and compartmentalizing the interpretation.

Given the interest to restrict our view to subsets of points, we have argued that CA should not simply be applied to the corresponding submatrix of data. We are interested in the same relative frequencies as in the complete analysis, that is the profiles, but want to map a subset of the profiles, which should not be re-expressed relative to their own subtotals. Similarly, the masses assigned to the points should also be the same as in the complete analysis, and not be determined by the masses in the subgroup. The simple variant of the method which we propose in the form of s-CA has these properties and notice that s-CA applied to the full set of categories is just regular CA.

The requirement here that a subset of relative frequencies should not be re-expressed relative to its subtotal is in contrast to the treatment of so-called "subcompositions" in compositional data analysis (Aitchison, 1986). Usually in a physical or chemical context, where samples are decomposed into components which are then measured and expressed in percentages by weight of the total sample, it is a basic principle that analytical methods have *subcompositional invariance*. That is, suppose a sample has organic and inorganic compounds, and we choose to restrict our attention to the organic compounds, then it would

be natural in this context to re-express the organic compounds relative to the total weight of the organic compounds. Subcompositional invariance would imply that any relationship we now find between the organic compounds should be identical to the relationship found if we had analysed the full set of data. As shown by Aitchison & Greenacre (2002), most multivariate methods including CA do not have subcompositional invariance, so special methods and transformations such as logratios are used in compositional data analysis which obey this principle. In our context of social and behavioural research, however, when we are dealing with multi-attribute data on several variables, it is clear that the notion of a subcomposition does not apply, since we are restricting our attention to selected categories from several variables, that is from several tables. There could be a situation, however, when we analyse a single contingency table and wish to restrict our map to a subset of categories of a single variable, where the subcompositional invariance principle would apply, and in this situation we would recommend biplots based on compositional data analysis as an alternative to CA.

Everything we have proposed here applies in exactly the same way to multiple correspondence analysis (MCA), which is the analysis of the individual responses in the form of an indicator matrix (see, for example, Table 2(b)). Technically, MCA is just CA applied to the indicator matrix, so if we were interested in a subset of categories, for example the "stay at home" categories, we would select this subset of the profile values and maintain the same masses and metric as in MCA. Notice that the row profiles in MCA consist of zeros with a value of $1/Q$ for every response (where Q is the number of variables or questions, in our example $Q = 4$). Hence in our example, performing MCA on a subset would involve picking out the subset of the profiles, consisting of zeros and values of ¼ as the case may be, with row masses all equal to $1/n$ (where n is the number of respondents) and the metric determined by the usual average values for each category. The exogenous variables, also coded as dummy variables for their categories are then depicted on the map in the usual way, as averages of the respondent points giving the respective responses.

References

Aitchison, J.A. (1986). *Compositional Data Analysis*. London: Chapman and Hall.

Aitchison, J.A. and Greenacre, M.J. (2002). Biplots of compositional data. *Applied Statistics,* **51**, 375–392.

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

Gabriel, K.R. and Odoroff, C.L. (1990). Biplots in biomedical research. *Statistics in Medecine*, **9**, 469–485.

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). London: Academic Press, pp. 141–161.

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.

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

Greenacre, M.J. and Blasius, J. (2005). *Multiple Correspondence Analysis and Related Methods.* In preparation.

ISSP (1994). *Family and Changing Gender Roles II, Survey ZA2620*. Zentralarchiv für Empirische Sozialforschung, University of Cologne, Germany.

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 in 1994 by the International Social Survey Program (ISSP).

Should married women work…

(1) … before first child?

(2) … with preschool child?

(3) … when youngest child at school?

(4) … when children have left home?

Response scales for each question: W (work full-time), w (work part-time), H (stay at home), ? (unsure/don't know)

Exogenous variables:

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)
Social class	7 groups: S0 (other), S1 (lower class), S2 (working class), S3 (upper working/lower middle), S4 (middle), S5 (upper middle), S6 (upper)
Country	24 countries: AUS (Australia), DW (West Germany), DE (East Germany) GB (Great Britain), NI (Northern Ireland), USA, A (Austria), H (Hungary), I (Italy), IRL (Ireland), NL (Netherlands), N (Norway), S (Sweden), CZ (Czechoslovakia), SLO (Slovenia), PL (Poland), BG (Bulgaria), RUS (Russia), NZ (New Zealand), CDN (Canada), RP (Phillipines), IL (Israel), J (Japan), E (Spain)

Table 2

Raw data in two different but equivalent forms: (a) the original response pattern data for questions 1 to 4 and the first two exogenous variables sex and age; (b) the indicator (dummy variable) form of coding. Response categories for questions 1 – 4 are: 1 = W *"work full-time"*, 2 = w *"work part-time"*, 3 = H *"stay at home"* and 4 = ? *"unsure/don't know"* and non-response; for sex: 1 = M *"male"*, 2 = F *"female"*; for age: six groups, 1 – 6. Data is shown only for first six respondents out of *n* = 33123.

(a)

1	2	3	4	Sex	Age	...
1	3	2	2	2	6	...
1	2	2	2	2	4	...
1	3	4	4	2	1	...
1	2	2	1	2	4	...
1	3	2	4	1	5	...
1	2	1	1	2	1	...
.	.	.	.	.	.	...
.	.	.	.	.	.	...
.	.	.	.	.	.	...
.	.	.	.	.	.	...

33123 cases

(b)

1				2				3				4				Sex		Age						...
W	w	P	?	W	w	P	?	W	w	P	?	W	w	P	?	M	F	1	2	3	4	5	6	
1	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	1	...
1	0	0	0	0	1	0	0	0	1	1	0	0	1	0	0	0	1	0	0	0	1	0	0	...
1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	1	1	0	0	0	0	0	...
1	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	...
1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	1	0	...
1	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	...
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	...
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	...
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	...
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	...

Table 3

Super-table of contingency tables crosstabulating the exogenous variables with the attitudinal questions on women working or staying at home, showing first two exogenous variables only (*cf*. Tables 1 and 2).

	1				2				3				4			
	W	w	H	?	W	w	H	?	W	w	H	?	W	w	H	?
Sex M	10309	2436	1207	1091	3220	7634	3099	1090	1360	4972	7732	979	9300	3030	1182	1531
Sex F	13712	2435	842	1056	4020	9999	2829	1197	1628	6926	8324	1167	12100	3408	893	1644
Age A1	3562	744	272	361	1621	2408	527	383	494	2095	1947	403	3448	741	261	489
Age A2	5192	885	337	459	1829	3659	864	521	804	2840	2710	519	4668	1151	334	720
Age A3	5173	879	332	486	1523	3795	1038	514	738	2603	3053	476	4516	1337	321	696
Age A4	4006	767	321	326	953	3065	1062	340	399	1833	2879	309	3436	1154	326	504
Age A5	3158	749	338	231	711	2406	1120	239	317	1397	2577	185	2691	1037	397	351
Age A6	2953	851	453	288	609	2320	1322	294	240	1139	2908	258	2660	1027	436	422
...	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
...	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
...	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Table 4

Subset of profiles analysed, for "stay at home" (H) responses only, showing masses allocated to each row and the subset of column averages used as inverse weights in the chi-square distance between rows. Since the masses sum to 1, multiply the masses by 6 (the number of variables) to obtain the sample proportions, for example the proportions of males and females are 0.4542 and 0.5448 respectively. See Table 1 for abbreviations.

	1H	2H	3H	4H	*mass*
M	0.0802	0.2060	0.5140	0.0786	*0.075693*
F	0.0467	0.1568	0.4613	0.0495	*0.090798*
A1	0.0551	0.1067	0.3942	0.0528	*0.024852*
A2	0.0490	0.1257	0.3943	0.0486	*0.034583*
A3	0.0483	0.1511	0.4444	0.0467	*0.034568*
A4	0.0592	0.1959	0.5312	0.0601	*0.027272*
A5	0.0755	0.2502	0.5757	0.0887	*0.022522*
A6	0.0997	0.2909	0.6398	0.0959	*0.022869*
ma	0.0647	0.1943	0.5016	0.0665	*0.107176*
wi	0.0943	0.2813	0.6077	0.0927	*0.012594*
di	0.0430	0.1274	0.4625	0.0384	*0.008650*
se	0.0507	0.1812	0.4589	0.0676	*0.002083*
si	0.0481	0.1109	0.4004	0.0466	*0.035650*
E0	0.1127	0.2077	0.5352	0.0986	*0.001429*
E1	0.1971	0.3869	0.5995	0.1998	*0.005540*
E2	0.0705	0.2498	0.5758	0.0677	*0.033542*
E3	0.0553	0.1695	0.5136	0.0491	*0.032057*
E4	0.0472	0.1507	0.4739	0.0476	*0.043303*
E5	0.0358	0.1278	0.4196	0.0415	*0.019257*
E6	0.0279	0.1043	0.3642	0.0274	*0.018195*
S0	0.0855	0.1765	0.4920	0.0703	*0.040646*
S1	0.1716	0.3201	0.5182	0.1542	*0.006068*
S2	0.0706	0.2183	0.5125	0.0729	*0.039046*
S3	0.0516	0.1889	0.5224	0.0673	*0.007407*
S4	0.0359	0.1503	0.4736	0.0455	*0.057830*
S5	0.0244	0.1151	0.3872	0.0297	*0.008655*
S6	0.0632	0.1154	0.3599	0.0714	*0.001832*
AUS	0.0189	0.1014	0.5969	0.0401	*0.008539*
DW	0.0304	0.2520	0.6201	0.0377	*0.011603*
DE	0.0082	0.0567	0.1967	0.0091	*0.005500*
GB	0.0157	0.0722	0.5649	0.0126	*0.004810*
NIRL	0.0302	0.1175	0.5587	0.0175	*0.003170*
USA	0.0234	0.0760	0.4915	0.0213	*0.007085*
A	0.0175	0.2082	0.5804	0.0443	*0.004881*
H	0.0673	0.2980	0.6273	0.0967	*0.007548*
I	0.0756	0.1640	0.3870	0.1483	*0.005122*
IRL	0.0390	0.2221	0.4518	0.0509	*0.004644*
NL	0.0093	0.0727	0.3889	0.0119	*0.009691*
N	0.0134	0.1008	0.4270	0.0124	*0.010134*
S	0.0024	0.0316	0.2751	0.0032	*0.006219*
CZ	0.0647	0.1931	0.5333	0.0186	*0.005132*
SLO	0.1088	0.2157	0.5238	0.0933	*0.005178*
PL	0.1879	0.3694	0.6970	0.0784	*0.007955*
BG	0.0573	0.2999	0.5667	0.0501	*0.005620*
RUS	0.0946	0.2753	0.5611	0.0976	*0.010053*
NZ	0.0120	0.0692	0.6309	0.0191	*0.005017*
CDN	0.0104	0.0750	0.3431	0.0153	*0.007246*
RP	0.3150	0.4033	0.3900	0.2733	*0.006038*
IL	0.0319	0.0715	0.1788	0.0529	*0.006471*
J	0.0660	0.2387	0.5741	0.0767	*0.006556*
ave	*0.0611*	*0.1788*	*0.4860*	*0.0615*	

Figure 1

CA map of Table 3, showing the four response categories for each question (W = "work full-time", w = "work part-time", H = stay at home", ? = "unsure/don't know/missing"). See Table 1 for abbreviations of row category points.

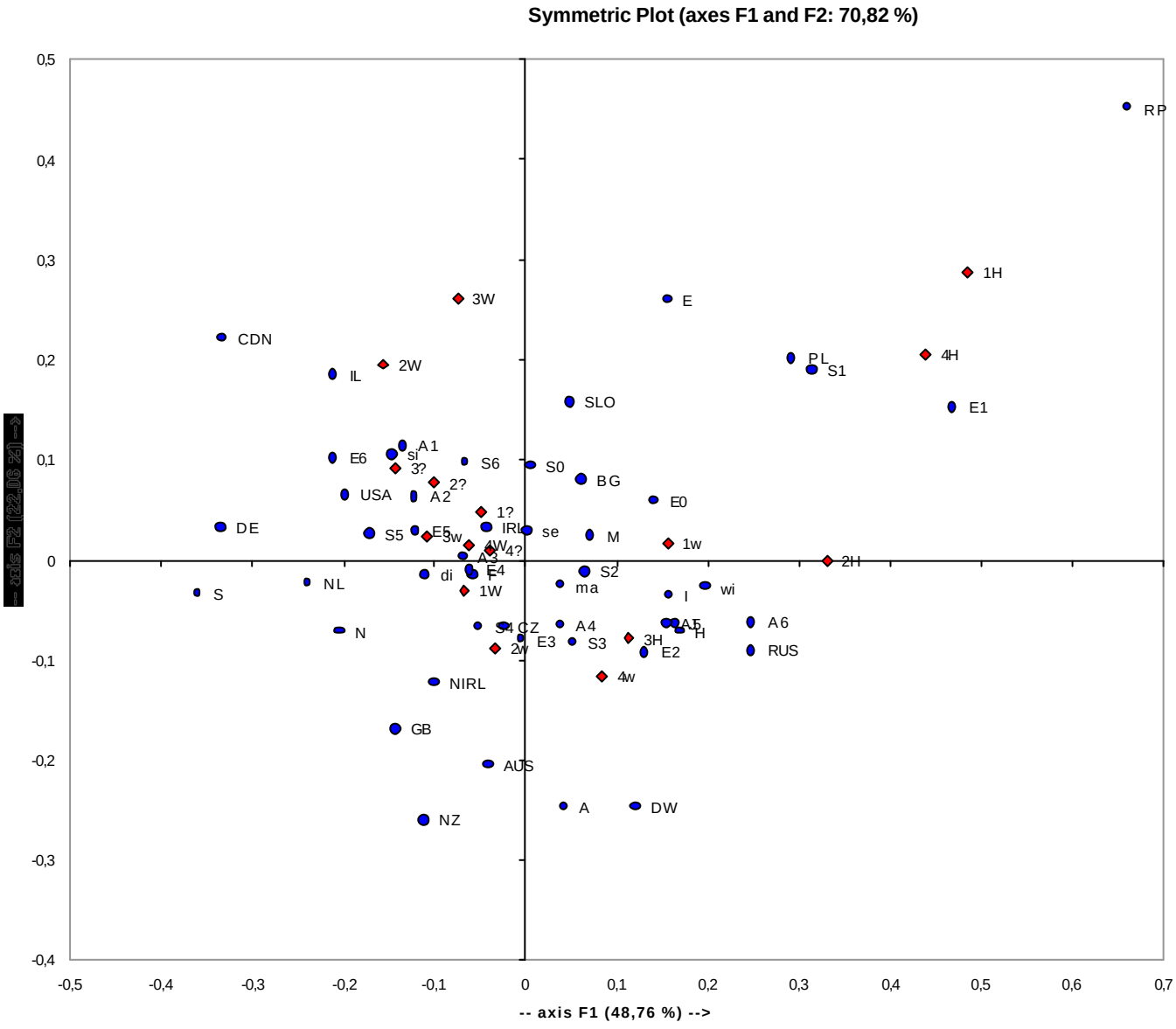

Figure 2

Subset CA map of "stay at home" categories of Table 4, showing dimension of size on axis 1. Inertias on the first axis: 0.0536 (78.0%), second axis: 0.0120 (17.5%); thus 95.5% of the inertia of the "stay-at-home" response values is displayed here.