15

Data Doubling and Fuzzy Coding

Michael Greenacre

CONTENTS

An important aspect of Benzécri's French school of data analysis is the recoding of data prior to visualization by correspondence analysis (CA), a theme treated in detail in the book by Murtagh (2005). The method of CA is seen as a universal tool for visualizing data of any kind, once recoded into a suitable form. To understand what makes a data set suitable for CA, one has to consider the elements of a frequency table, which is the primary object on which CA is applicable, and which needs no pretransformation:

- Each cell of a frequency table is a count.
- The row or column frequencies are expressed relative to their marginal totals as profile vectors.
- The marginal frequencies of the table provide masses that weight the row and column profiles in the measure of variance and in the dimension reduction.
- The chi-square distance between profiles inherently standardizes the profile elements.

The idea of recoding data prior to CA is illustrated nicely in the original definition of multiple correspondence analysis (MCA), where all categorical variables are recoded as dummy variables (see Chapter 3 by Lebart and

Copyright © 2014. Chapman and Hall/CRC. All rights reserved.

Visualization and Verbalization of Data, edited by Jorg Blasius, and Michael Greenacre, Chapman and Hall/CRC, 2014. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/utd/detail.action?docID=1414098.
Created from utd on 2018-10-15 14:53:55.

Saporta in this book). Variants of this idea are the so-called *doubling* of data and *fuzzy coding*, which are the topics of this chapter.

Doubling (*dédoublement* in French) can be considered a simple generalization of the MCA of dichotomous variables, that is, when each categorical variable has only two categories. But, instead of the coding [0 1] or [1 0] used in MCA, there is a pair of nonnegative values between 0 and an upper bound, for example, 10, and the two values have a sum equal to the upper bound: for example, [3 7] or [9 1], or even [0 10] where the bounds are included. This type of coding arises in the analysis of ratings, preferences, or paired comparisons, as I will show later.

Doubling can be considered a special case of fuzzy coding (*codage flou* in French), for which a set of fuzzy categories is defined, often between two and five in number. In fuzzy coding the data are nonnegative values between 0 and 1 that sum to 1, like a small profile vector: for example, examples of fuzzy coding with three categories are [0.2 0.3 0.5] and [0.012 0.813 0.175]. Since the doubling described previously could just as well be coded in CA between 0 and the upper bound of 1 rather than between 0 and 10 in that case, by dividing by the upper bound, the three examples of doubling mentioned above can equivalently be considered two-category fuzzy coding, with values [0.3 0.7], [0.9 0.1], and [0 1], respectively. In what follows the idea of doubling and fuzzy coding will be further developed in the specific contexts in which they are applied.

15.1 Doubling of Ratings

Rating scales are ubiquitous in the social sciences, for example, the five-point Likert scale of agreement: 1 = strongly agree, 2 = somewhat agree, 3 = neither agree nor disagree, 4 = somewhat disagree, and 5 = strongly disagree. As an example of such data, I reconsider the Spanish data subset from the International Social Survey Program (ISSP, 2002) on attitudes to working women, previously analysed by Greenacre (2010, chap. 9) and available at http://zacat.gesis.org. This data set includes eight statements to which 2,107 respondents answered on this 5-point scale (cases with missing data have been excluded). The statements are

A: A working mother can establish a warm relationship with her child.

B: A preschool child suffers if his or her mother works.

C: When a woman works the family life suffers.

D: What women really want is a home and kids.

E: Running a household is just as satisfying as a paid job.

Copyright © 2014. Chapman and Hall/CRC. All rights reserved.

Visualization and Verbalization of Data, edited by Jorg Blasius, and Michael Greenacre, Chapman and Hall/CRC, 2014. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/utd/detail.action?docID=1414098.
Created from utd on 2018-10-15 14:53:55.

F: Work is best for a woman's independence.

G: A man's job is to work; a woman's job is the household.

H: Working women should get paid maternity leave.

In order to recode the data into its doubled form, the 1–5 scale has first to be transformed to a 0–4 scale by subtracting 1, so that the upper bound is 4 in this case. This variable is then given its label and a negative sign, for example, ***A–***, since it quantifies the level of disagreement. The doubled version of the data is then computed by subtracting that value from 4, creating a variable with label ***A+***, which codes the level of agreement. Table 15.1 shows the doubled data for the first respondent. The value of 2 ('somewhat agree') for statement ***A*** becomes 3 and 1 in the doubled coding for ***A+*** and ***A–***, while the 4 ('somewhat disagree') for statement ***B*** becomes 1 and 3 for ***B+*** and ***B–***.

There is a connection between the doubled coding and the counting paradigm inherent in CA. If one thinks of the original scale points 1 to 5 written from right to left, so that strong agreement (= 1 on the original scale) is to the right and strong disagreement (= 5) to the left, a response of 2 ('somewhat agree') has 3 scale points to the left of it and 1 scale point to the right, hence the coding of [3 1], indicating more weight toward the positive pole of agreement. A response of 4 has 1 scale point to the left and 3 points to the right, hence [1 3]. A middle response of 3 has 2 scale points on either side; hence, it is balanced and coded [2 2], giving an equal count to both poles. The 'strongly agree' response to statement ***F*** is coded as a count of 4 for agreement and a count of 0 to disagreement.

With this doubling of the variables there is a perfect symmetry between the agreement and disagreement poles of the attitude scale. Since there are some statements worded positively and some worded negatively toward women working (statements ***A*** and ***B*** are good examples), one would expect these to be answered in a reversed way, as indeed respondent 1 does, agreeing to ***A*** and disagreeing to ***B***. The use of integer values from 0 to 4 is, in fact, arbitrary, since each row of data will sum to a constant row margin, $8 \times 4 = 32$ in this case. So values from 0 to 1 could have been used to give the coding a probabilistic flavour: a [3 1] would thus be [¾ ¼] and the row sums

TABLE 15.1

Original 5-Point Scale Data for Respondent 1, and the Coding in Doubled Form

Original Responses

A	***B***	***C***	***D***	***E***	***F***	***G***	***H***
2	4	3	3	4	1	4	1

Doubled Data

A+	***A–***	***B+***	***B–***	***C+***	***C–***	***D+***	***D–***	***E+***	***E–***	***F+***	***F–***	***G+***	***G–***	***H+***	***H–***
3	1	1	3	2	2	2	2	1	3	4	0	1	3	4	0

Note: For each variable the negative pole of the doubled value is 1 less than the original response code, and the positive pole is 4 minus the negative pole.

Copyright © 2014. Chapman and Hall/CRC. All rights reserved.

Visualization and Verbalization of Data, edited by Jorg Blasius, and Michael Greenacre, Chapman and Hall/CRC, 2014. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/utd/detail.action?docID=1414098.
Created from utd on 2018-10-15 14:53:55.

would sum to a constant 8 (the number of variables) for each respondent. The coding then looks very much like the coding in MCA for dichotomous variables, but in MCA only the extreme points [0 1] and [1 0] are observed.

The doubled data matrix, with 16 columns in this example, has a dimensionality of 8, exactly the same as if one had analysed the original response data by principal component analysis (PCA). The CA of doubled data has a very close similarity to the PCA of the original (undoubled) data, and the only methodological difference is in the metric used to measure distances between case points. In PCA, because all variables are on the same five-point rating scale, no standardization would be performed and the metric would be the regular unweighted Euclidean distance. Greenacre (1984) showed that the chi-square metric for a matrix of doubled data implied a standardization of the original variables similar to that of a Bernoulli variable, which has variance $p(1 - p)$, where p and $(1 - p)$ are the probabilities of 'success' (coded as 1) and 'failure' (coded as 0). In the present example, the means across the 2,107 cases of the doubled variables are given in the first row of Table 15.2 (each pair also adds up to 4). The second row gives the same means divided by 4, so that each pair adds up to 1. The third row computes the product of the pair of values of each variable, an estimation of the Bernoulli variance, while the fourth and last row is the square root of the variance estimate, in other words, an estimate of standard deviation.

The result shown by Greenacre (1984, p. 183) is that the chi-square distance between cases, based on the doubled matrix, is—up to an overall scaling constant—the same as the Euclidean distance between cases, based on the original undoubled matrix, after standardizing the variables using the above estimates of standard deviation. Hence, a PCA using this standardization would yield the same result. Compared to the regular unstandardized PCA that would usually be applied to these data, CA, using the chi-square distance, would boost the contribution of variable ***H*** slightly, because its standard deviation is somewhat smaller than the others. In other words, compared to the Euclidean distance, the chi-square distance increases the contributions of questions with attitudes toward the extremes of the scale, where the variance is lower.

The CA asymmetric map of these doubled data is shown in Figure 15.1. The geometry of the joint map is very similar to that of an MCA for dichotomous data. Figure 15.2 shows examples of two cases, the first of which (on the left-hand side of the map) has only extreme responses to the eight statements: ***A*** = 1, ***B*** = 5, ***C*** = 5, ***D*** = 5, ***E*** = 5, ***F*** = 1, ***G*** = 5, ***H*** = 1, strongly agreeing to the statements in favour of working women, and strongly disagreeing to those against. This case lies at the average position of the corresponding endpoints of the variables and is consequently the leftmost case in the map. The case on the right, on the other hand, has some intermediate response categories, with the response pattern ***A*** = 4, ***B*** = 1, ***C*** = 1, ***D*** = 2, ***E*** = 3, ***F*** = 4, ***G*** = 2, ***H*** = 1. Some of the response categories thus lie at intermediate positions on the line segments of the variables. For example, the response ***D*** = 2 is indicated by a

Copyright © 2014. Chapman and Hall/CRC. All rights reserved.

Visualization and Verbalization of Data, edited by Jorg Blasius, and Michael Greenacre, Chapman and Hall/CRC, 2014. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/utd/detail.action?docID=1414098.
Created from utd on 2018-10-15 14:53:55.

Copyright © 2014. Chapman and Hall/CRC. All rights reserved.

TABLE 15.2

Means of Original Doubled Data, Their Rescaled Values, and Estimates of Variance and Standard Deviation for Each Variable

A+	*A–*	*B+*	*B–*	*C+*	*C–*	*D+*	*D–*	*E+*	*E–*	*F+*	*F–*	*G+*	*G–*	*H+*	*H–*
2.45	1.55	2.14	1.86	2.24	1.76	1.90	2.10	1.90	2.10	2.91	1.09	1.38	2.62	3.47	0.53
0.613	0.387	0.534	0.466	0.560	0.440	0.474	0.526	0.476	0.524	0.728	0.272	0.344	0.656	0.867	0.133
0.237		0.249		0.246		0.249		0.249		0.198		0.226		0.115	
0.487		0.499		0.496		0.499		0.499		0.445		0.475		0.340	

Note: Row 1: The means of the eight doubled variables on a 0–4 scale. Row 2: The means of doubled variables on a 0–1 scale (first row divided by 4), denoted by p_j and $1 - p_j$, $j = 1, \ldots, 8$. Row 3: Values of the product $p_j(1 - p_j)$, an estimate of variance. Row 4: Values of $\sqrt{p_j(1-p_j)}$, an estimate of standard deviation.

Visualization and Verbalization of Data, edited by Jorg Blasius, and Michael Greenacre, Chapman and Hall/CRC, 2014. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/utd/detail.action?docID=1414098.
Created from utd on 2018-10-15 14:53:55.

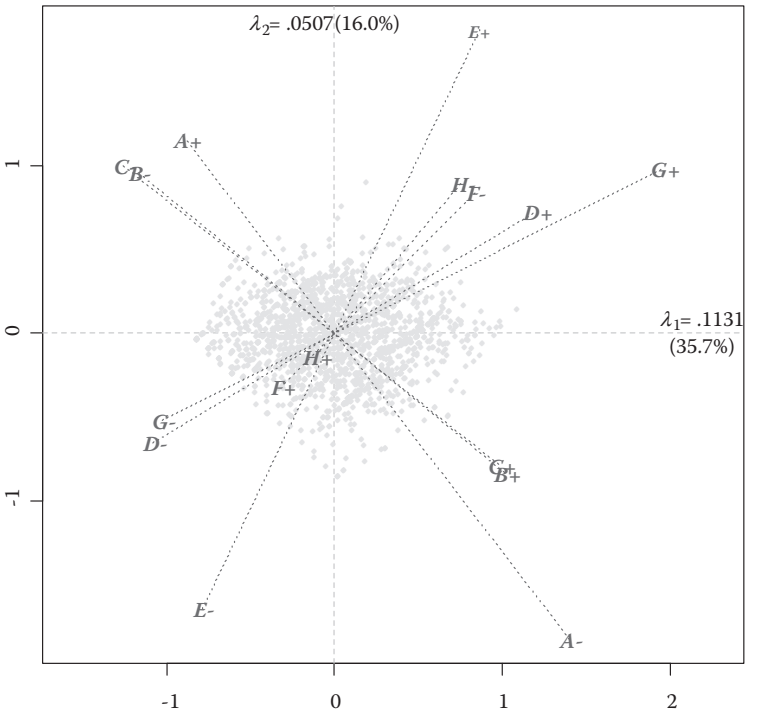

FIGURE 15.1
CA map of the doubled matrix of women working data, showing the agreement (+) and disagreement (–) poles of the rating scales and the 2,107 respondents. The scaling is 'row-principal', that is, variables in standard coordinates and cases in principal coordinates. 51.7% of the total inertia is explained by the two-dimensional solution.

Copyright © 2014. Chapman and Hall/CRC. All rights reserved.

cross at a position on the ***D*** segment corresponding to the 'somewhat agree' category, as if the segment was cut into four equal intervals between ***D–*** and ***D+***. Similarly, the middle response ***E*** = 3 corresponds to the midpoint of the ***E*** segment. Given the positions on the eight segments of this case's responses, the case lies at the average position, as shown.

As for MCA, the category points for each variable have their centroid at the origin. This means that the average rating is exactly displayed by the position of the origin on each segment. For example, the average response to statement ***G*** is toward disagreement, because the centre is closer to the ***G–*** pole, while for ***F*** and ***H*** the average attitude is closer to agreement. If one calibrates each segment linearly from 1 at the + end to 5 at the – end, the corresponding mean can be read off exactly. Notice, however, that there is an alternative calibration possible for the biplot that allows prediction of the values of each case; see, for example, Greenacre (2010, chap. 3).

Visualization and Verbalization of Data, edited by Jorg Blasius, and Michael Greenacre, Chapman and Hall/CRC, 2014. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/utd/detail.action?docID=1414098.
Created from utd on 2018-10-15 14:53:55.

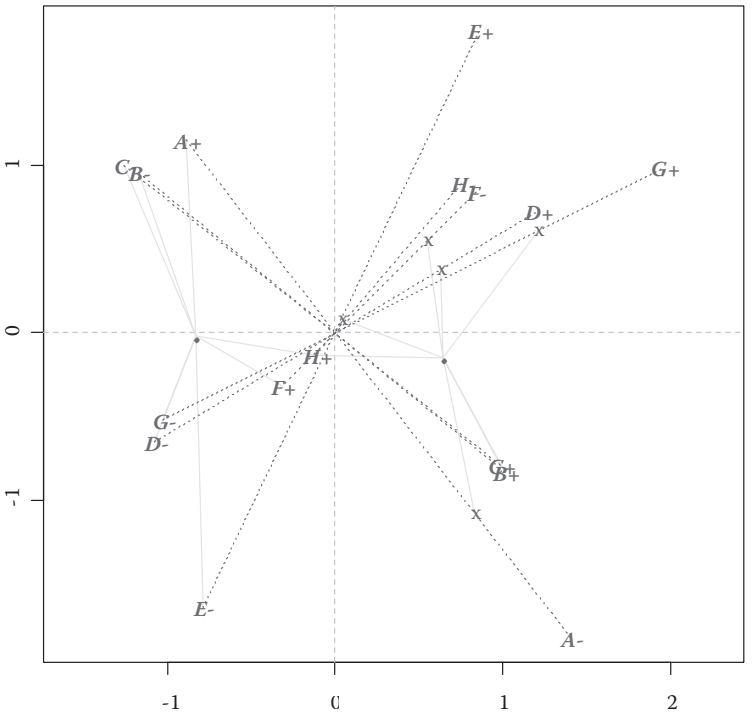

FIGURE 15.2
Examples of the barycentric relationship between cases and variables in the CA map of Figure 15.1. The case on the left has response pattern 1 5 5 5 5 1 5 1, while the case on the right has responses 4 1 1 2 3 4 2 1, where 1 ('strong agreement') corresponds to the + pole of the scale and 5 ('strong disagreement') to the – pole.

15.2 Doubling of Preferences and Paired Comparisons

The same doubling strategy can be used in the CA of preferences, or rank orderings, and paired comparison data. Rank orderings can be thought of as a special case of ratings: for example, if m objects are being ranked, then this is equivalent to an m-point rating scale where the respondent is forced to use each scale point only once. If six objects, ***A*** to ***F***, were ranked, from 1 (most preferred) to 6 (least preferred), then doubling would again entail reversing the scale and subtracting 1, giving the + pole and creating the – pole as 5 minus the + pole. Suppose object ***C***, for example, was ranked second; then its doubled coding would be ***C+*** = 4 and ***C–*** = 1, indicating

Copyright © 2014. Chapman and Hall/CRC. All rights reserved.

Visualization and Verbalization of Data, edited by Jorg Blasius, and Michael Greenacre, Chapman and Hall/CRC, 2014. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/utd/detail.action?docID=1414098.
Created from utd on 2018-10-15 14:53:55.

that the respondent preferred ***C*** over four other objects, but preferred one object to ***C***. For a sample of n cases ranking m objects the doubled matrix can be constructed as before as an $n \times 2m$ matrix, with each object doubled. But in this case there is an alternative way to set up the matrix: the doubled columns can be stacked on top of each other to form a $2n \times m$ matrix so that the cases are doubled, not the objects. The original data set (after subtracting 1) then receives case labels with – appended, while the doubled matrix has case labels with +. Torres and Greenacre (2002) showed that the CA of this latter data structure is equivalent to Nishisato's dual scaling of preferences (Nishisato, 1994). In this CA map each object appears as a single point, while each case appears as two points, a + case point representing the case's preferences with respect to the six objects and a – point exactly opposite at the same distance from the origin representing the case's 'dispreferences'. Clearly, the – points are redundant to the display, but are nevertheless displayed by analogy with the doubled rating scales described previously.

Paired comparisons can be coded in a similar doubled fashion. Suppose respondents were asked to compare the six objects pairwise. There are 15 unique pairs (in general, $\frac{1}{2}m(m-1)$ pairs for m objects), and a typical response might be ***A*** > ***B***, ***C*** > ***A***, ***A*** > ***D***, ***A*** > ***E***, ***F*** > ***A***, ***B*** > ***C***, ***B*** > ***D***, . . . , and so on, where > stands for 'is preferred to'. The doubled variables are now real counts of how many times an object is preferred and dispreferred, respectively. For example, in the above list ***A*** is preferred to three others and two are preferred to ***A***, so the data pair would be ***A+*** = 3 and ***A–*** = 2. This type of coding can also be used for incomplete, but balanced, paired comparison designs. Notice that doubled values being coded here are the margins of an objects-by-objects preference matrix for each respondent (6×6 in this illustration)—these margins are not necessarily equivalent to the paired comparisons, when there are inconsistencies in the judgements, whereas in the case of a ranking these margins are equivalent to the ranking information.

Before explaining fuzzy coding let me point out a nonparametric analysis of a table of continuous data, which is related to the CA of preferences. Suppose that a data matrix consists of observations by n cases on p continuous variables. Rank order each variable in ascending order, so that the lowest value is replaced by 1, the second lowest by 2, and so on, up to n for the highest value. Notice that these 'preference orderings' are down the columns of the data matrix, not across the rows as discussed above where objects are ordered. Then double the variables as before, defining the positive pole as the rank minus 1, and the negative pole as $(n-1)$ minus the positive pole. The CA of this matrix gives identical results, up to a scaling factor, to the PCA of the (undoubled) matrix of ranks (Greenacre, 1984, p. 132).

Copyright © 2014. Chapman and Hall/CRC. All rights reserved.

Visualization and Verbalization of Data, edited by Jorg Blasius, and Michael Greenacre, Chapman and Hall/CRC, 2014. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/utd/detail.action?docID=1414098.
Created from utd on 2018-10-15 14:53:55.

15.3 Fuzzy Coding of Continuous Variables

The idea in fuzzy coding is to convert a continuous variable into a pseudo-categorical (i.e., fuzzy) variable using so-called *membership functions*. I illustrate this with the simplest example of triangular membership functions, shown in Figure 15.3, defining a fuzzy variable with three categories. On the horizontal axis is the scale of the original variable and three *hinge points*, chosen as the minimum, median, and maximum values of the variable. The three functions shown are used for the recoding, and on the left a particular value of the original variable is shown to be recoded as 0.3 for category 1, 0.7 for category 2, and 0 for category 3. This coding is invertible: given the fuzzy observation [0.3 0.7 0.0] the value of the original variable is

$$\text{Original variable} = 0.3 \times \text{minimum} + 0.7 \times \text{median} + 0.0 \times \text{maximum} \quad (15.1)$$

Fuzzy coding in CA first appeared in the thesis of Bordet (1973) and was subsequently used by Guittonneau and Roux (1977) in a taxonomic study where continuous variables characterizing plants were coded into fuzzy categories, thus enabling them to be analysed jointly with several categorical variables. The data set in question consists of 82 examples from a taxonomic study of 75 different species, which can be combined into 11 groups of the plant genus *Erodium* (see Figure 15.4): for example, *Erodium plumosa* and *Erodium romana*. Apart from the botanical objective of seeing how these groups differ in terms of their characteristics, the interest from a data analytic point of view is the mix of 33 categorical and five continuous variables, and the challenge of analysing them jointly. Like the original authors, I will apply fuzzy coding to the continuous variables in order to analyse them jointly with the categorical variables, recoding each variable into four categories using four hinge points: minimum, first tercile, second tercile, and maximum.

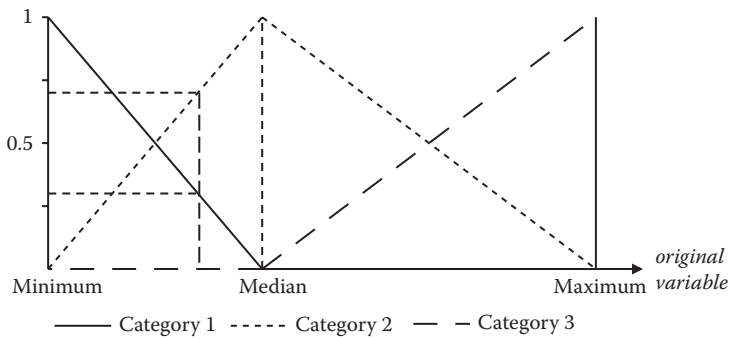

FIGURE 15.3
Triangular membership functions that convert a continuous variable into a fuzzy variable with three categories. The minimum, median, and maximum are used as hinge points.

Copyright © 2014. Chapman and Hall/CRC. All rights reserved.

Visualization and Verbalization of Data, edited by Jorg Blasius, and Michael Greenacre, Chapman and Hall/CRC, 2014. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/utd/detail.action?docID=1414098.
Created from utd on 2018-10-15 14:53:55.

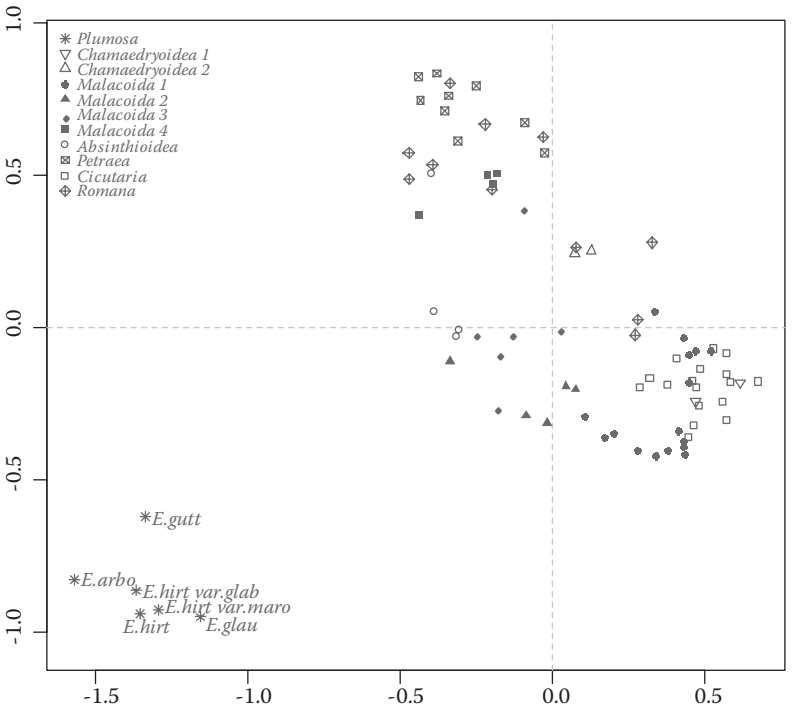

FIGURE 15.4
Correspondence analysis of recoded *Erodium* data, showing species only in principal coordinates, with clear separation of the six species forming the subgenus *Plumosa*. For species names, see Guittonneau and Roux (1977).

Of the categorical variables 15 are dichotomous, 13 are trichotomous, four have four categories, and one has six. Guittonneau and Roux (1977), who published the data, recoded the 13 trichotomous variables into two categories by assigning values [½ ½] for the second category when it occurred, arguing that it was always an intermediate category between categories 1 and 3. In my reanalysis, I prefer to keep all three categories. The 33 categorical variables thus generate $15 \times 2 + 13 \times 3 + 4 \times 4 + 1 \times 6 = 91$ dummy variables. Each of the five continuous variables is transformed to four fuzzy categories, using triangular membership functions as described above. Thus, $5 \times 4 = 20$ additional variables are generated, so that the grand total is 111 variables instead of the original 38. (It is interesting to notice how times have changed. In their data recoding, which extends the number of variables to be analysed by introducing many dummy variables and fuzzy categories, Guitonneau and Roux (1977) say that an inconvenience of the recoding is that 'after transformation, the table requires much more memory in the computer; moreover, the computation time is increased by an important proportion'. Their table, after transformation, was only of size 82×98, very small by today's standards.)

Copyright © 2014. Chapman and Hall/CRC. All rights reserved.

Visualization and Verbalization of Data, edited by Jorg Blasius, and Michael Greenacre, Chapman and Hall/CRC, 2014. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/utd/detail.action?docID=1414098.
Created from utd on 2018-10-15 14:53:55.

Before passing these data through CA, there is an issue of weighting of the dummy and fuzzy variables, which authors Guittonneau and Roux also considered. The argument was, and still is, that the more categories a variable has, the higher the variance created in the generated dummy variables. Also, crisp (0/1) and fuzzy coded variables with the same number of categories do not have the same inherent variances. In their analysis Guittonneau and Roux (1977) applied an algorithm proposed by Benzécri (1977a) and implemented by Hamrouni (1977) to reweight the groups of variables, each considered to be homogeneous internally (i.e., on the same measurement scale), but each having different number of categories. In my reanalysis, I have followed this idea, reweighting the four groups of variables—15 dichotomous, 13 trichotomous, five polychotomous, and five fuzzy—so that their variances (i.e., inertias) are equalized in the analysis. This amounted to multiplying the four sets of columns by the following respective coefficients: 0.228, 0.132, 0.201, and 0.439. Figure 15.4 shows the display of the samples in the CA of the full 82×111 recoded and reweighted matrix, with symbols indicating the 11 different groups of species. There is a clear separation of one group of six species at the bottom left, forming the subgenus *Plumosa*. Figure 15.5 shows

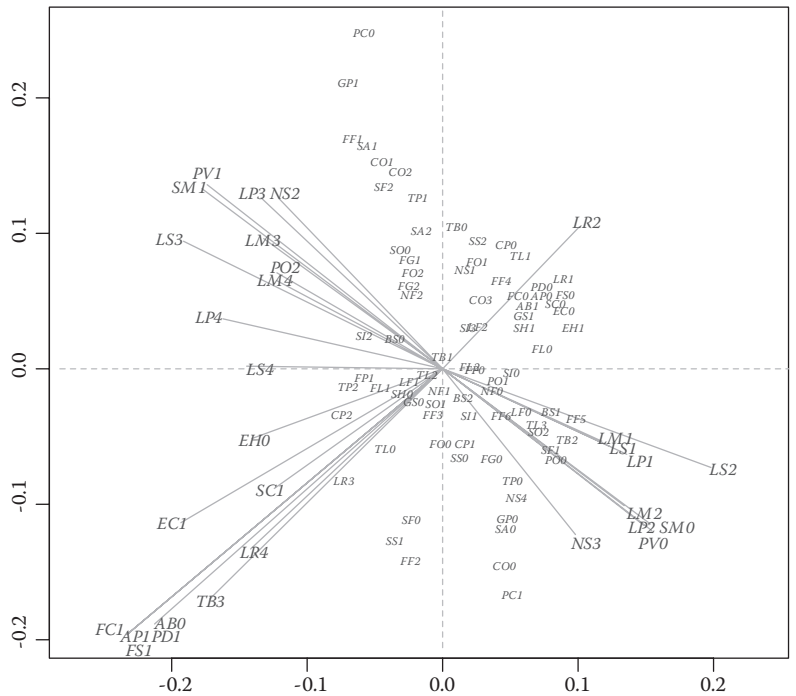

FIGURE 15.5
Correspondence analysis of recoded *Erodium* data, showing in standard coordinates the categories of the variables that contribute more than average to the two-dimensional solution. For a description of the variable names, see Guittonneau and Roux (1977).

Copyright © 2014. Chapman and Hall/CRC. All rights reserved.

Visualization and Verbalization of Data, edited by Jorg Blasius, and Michael Greenacre, Chapman and Hall/CRC, 2014. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/utd/detail.action?docID=1414098.
Created from utd on 2018-10-15 14:53:55.

the corresponding plot of the categories in the contribution biplot scaling (see Greenacre, 2013), with variable categories that contribute more than average to the axes of the two-dimensional solution shown in larger font. The categories ***AB0***, ***AP1***, ***FS1***, ***PD1***, and ***FC1***, for example, are associated exclusively, or almost exclusively, with subgenus *Plumosa* and are its main botanical indicators. Five out of these six species also fall into category ***TB3***, for example, the highest category of the four-category variable ***TB***, which is also stretching far in the direction of these species.

In order to visualize the variation within the other species, either the *Plumosa* species should be removed, or the variables associated with them, but this is unsatisfactory in both cases. If the species are removed, then some variables have zero observations in the remaining species and have to be removed anyway; if we remove the variables, then some information is lost where there are some observations of these variables in other species. A good compromise here is to use subset CA (Greenacre and Pardo, 2006a), retaining all the variables but excluding the six species of subgenus *Plumosa* (subset CA has been used in multiple correspondence analysis as well—see Greenacre and Pardo (2006b)—to exclude missing value categories, where

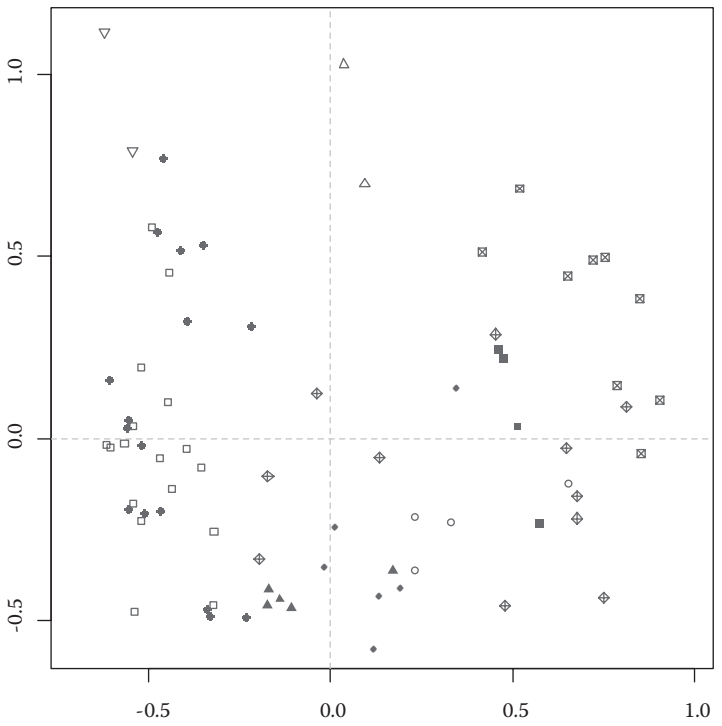

FIGURE 15.6
Subset correspondence analysis of recoded *Erodium* data, omitting species in subgenus *Plumosa*. See Figure 15.4 for key to group symbols.

Copyright © 2014. Chapman and Hall/CRC. All rights reserved.

Visualization and Verbalization of Data, edited by Jorg Blasius, and Michael Greenacre, Chapman and Hall/CRC, 2014. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/utd/detail.action?docID=1414098.
Created from utd on 2018-10-15 14:53:55.

it is also called missing data passive (Gifi, 1980) or specific multiple correspondence analysis; see Chapter 12 by Le Roux in this book). In subset analysis the column margins of the original table are maintained; hence, the centre of the space as well as the metric from the original analysis are preserved. Figure 15.6 shows the species in the subset excluding those in subgenus *Plumosa* (see Figure 15.4), but maintaining exactly the same space (i.e., same origin, same distance metric, and same point masses). Figure 15.7 shows the categories, in standard coordinates, of the most contributing variables—this is different from Figure 15.5, which showed the highly contributing categories, in contribution coordinates with lines from the origin. In Figure 15.7 the contributions of each variable were aggregated, and the variables with higher than average contribution are plotted, showing all their categories, connected to show their trajectories. The second principal (vertical) axis shows the trajectories of the variables ***LP***, ***LS***, ***LM***, ***LR***, and ***NS***, all lining up from low to high in their expected order. However, there is an interesting mirror-image pattern of these trajectories along the first principal (horizontal) axis, contrasting low values of variable ***NS*** and high

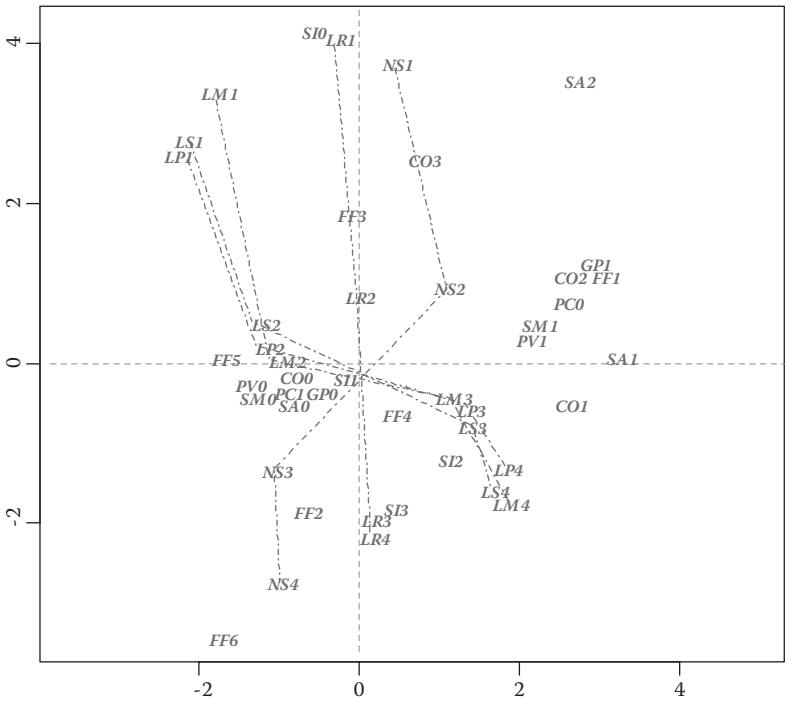

FIGURE 15.7
Subset correspondence analysis of recoded *Erodium* data, showing in standard coordinates the categories of the variables that contribute more than average to the two-dimensional solution. The fuzzy-coded variables, among the top contributors, are shown with their consecutive categories connected.

Copyright © 2014. Chapman and Hall/CRC. All rights reserved.

Visualization and Verbalization of Data, edited by Jorg Blasius, and Michael Greenacre, Chapman and Hall/CRC, 2014. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/utd/detail.action?docID=1414098.
Created from utd on 2018-10-15 14:53:55.

values of ***LM***, ***LP***, and ***LS*** against the high values of ***NS*** and low values of the other three, with variable ***LR*** not following this contrast. While the pattern is clear, it has no obvious interpretation.

15.4 Explained Inertia

Up to now I have generally refrained from reporting inertias and percentages of explained inertia. With respect to the data matrix entering the respective analyses, the relevant results are as follows:

- Analysis of reweighted data, Figures 15.4 and 15.5: Total inertia = 1.642, of which 27.7% is explained by the two-dimensional solution.
- When doing a subset analysis excluding subgenus *Plumosa*, the inertia of 1.642 splits into two parts: 0.0256 due to *Plumosa* and 1.386 due to the rest, the latter being decomposed in the subset analysis.
- Subset analysis excluding *Plumosa*, Figures 15.6 and 15.7: Total inertia = 1.386, of which 26.0% is explained by the two-dimensional solution.

It is well known that in the CA of data coded in an indicator matrix, for example, in a multiple correspondence analysis (MCA), the percentages of inertia seriously underestimate the explained variance. Greenacre (1995) demonstrated an easy way to adjust the total inertia as well as the parts of inertia to obtain improved estimates of the overall quality of the solution (for a recent account, see Greenacre, 2007, chap. 19). The same phenomenon appears in the analysis of fuzzy-coded data, and Aşan and Greenacre (2011) showed how estimates from a CA of fuzzy data can be transformed back to the original continuous scales of the data, a process called defuzzification, in order to obtain improved parts of variance explained. In the present example, however, there are two additional aspects that make this strategy more complicated: first, there is a mixture of measurement scales, including crisp (zero/one) and fuzzy coded data; and second, the groups of variables have been reweighted. Estimating more realistic overall percentages of explained inertia in this situation remains an open problem.

For the present situation, therefore, it is not possible to obtain a single adjusted estimate of the global quality of the solution, but one can make separate computations of the adjusted inertia explained for the crisply coded categorical data (variables 1 to 33) and fuzzy coded data (variables 34 to 38), in both the analysis reported in Figures 15.4 and 15.5 and the subset analysis of Figures 15.6 and 15.7. The adjusted percentages of each group as well as their original percentages are reported in Table 15.3. In all cases the adjusted

Copyright © 2014. Chapman and Hall/CRC. All rights reserved.

Visualization and Verbalization of Data, edited by Jorg Blasius, and Michael Greenacre, Chapman and Hall/CRC, 2014. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/utd/detail.action?docID=1414098.
Created from utd on 2018-10-15 14:53:55.

TABLE 15.3

Adjusted Percentages of Inertia Explained by Crisp and Fuzzy Variables in the Two Analyses of the *Erodium* Data, Before and After Adjustment

	Crisp Variables		Fuzzy Variables	
	(Original)	**Adjusted**	**(Original)**	**Adjusted**
Global analysis, 2D solution	(27.4%)	71.2%	(28.5%)	43.8%
Subset analysis, 2D solution	(19.0%)	35.8%	(44.9%)	67.5%

percentages are higher, especially for the categorical data coded crisply in indicator form.

15.5 Discussion

Doubling and fuzzy coding are part of the philosophy of Benzécri's French school of data analysis, enabling ratings, preferences, paired comparisons, and continuous data to be coded in a way that fits the correspondence analysis paradigm. Correspondence analysis is thus considered to be a versatile dimension reduction and visualization technique, where the data are pretransformed in various ways prior to analysis. Fuzzy coding transforms continuous data to a form that is comparable to categorical data, and so enables analysis of mixed measurement scales. Doubling can be considered a special case of fuzzy coding when there are only two fuzzy categories. Reweighting the variables can help to compensate for different numbers of categories in the various measurement scales. Explained variance is still an open issue: here it was necessary to consider separately the parts of explained variance for the crisply coded (zero/one) variables from that of the fuzzy coded ones, and it would be preferable to develop a global measure.

Copyright © 2014. Chapman and Hall/CRC. All rights reserved.

Visualization and Verbalization of Data, edited by Jorg Blasius, and Michael Greenacre, Chapman and Hall/CRC, 2014. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/utd/detail.action?docID=1414098.
Created from utd on 2018-10-15 14:53:55.

Copyright © 2014. Chapman and Hall/CRC. All rights reserved.

Visualization and Verbalization of Data, edited by Jorg Blasius, and Michael Greenacre, Chapman and Hall/CRC, 2014. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/utd/detail.action?docID=1414098.
Created from utd on 2018-10-15 14:53:55.