# Biplots of fuzzy coded data

Zerrin Aşan[1] and Michael Greenacre[2]

[1]Department of Statistics, Anadolu University, Eskişehir, Turkey
Email: zasan@anadolu.edu.tr

[2]Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain
Email: michael@upf.es

## Abstract

A biplot, which is the multivariate generalization of the two-variable scatterplot, can be used to visualize the results of many multivariate techniques, especially those that are based on the singular value decomposition. We consider data sets consisting of continuous-scale measurements, their fuzzy coding and the biplots that visualize them, using a fuzzy version of multiple correspondence analysis. Of special interest is the way quality of fit of the biplot is measured, since it is well-known that regular (i.e., crisp) multiple correspondence analysis seriously under-estimates this measure. We show how the results of fuzzy multiple correspondence analysis can be defuzzified to obtain estimated values of the original data, and prove that this implies an orthogonal decomposition of variance. This permits a measure of fit to be calculated in the familiar form of a percentage of explained variance, which is directly comparable to the corresponding fit measure used in principal component analysis of the original data. The approach is motivated initially by its application to a simulated data set, showing how the fuzzy approach can lead to diagnosing nonlinear relationships, and finally it is applied to a real set of meteorological data.

**Key words**: defuzzification, fuzzy coding, indicator matrix, measure of fit, multivariate data, multiple correspondence analysis, principal component analysis.

## 1. Introduction

The term "biplot" was introduced by Gabriel [7] in the context of principal component analysis (PCA) as the representation of the rows and columns of a data table as points in a joint plot in which scalar products between row and column points optimally approximate the corresponding data elements. The idea of the biplot, a generalization of a two-variable scatterplot to many variables, has found its way into many other multivariate techniques having results that can be visualized in this way: for example, linear regression, generalized linear models, multidimensional scaling, log-ratio analysis, various types of correspondence analysis and discriminant analysis – see [11] and, for a recent account, [15]. Several papers showing this style of graphical representation have appeared in applications of fuzzy data analysis: some examples are [4, 21, 18].

By fuzzy coded data we will mean data on several variables on continuous measurement scales that have been recoded into categories in a fuzzy way, as opposed to crisp coded data where the coding is made into sets of dummy variables with only ones and zeros. The object of this paper on the biplots of such fuzzy coded data is twofold. First, we want to compare the original biplot of Gabriel with the biplot of fuzzy coded data – the former displays linear relationships only between variables, whereas the latter can display more general inter-variable relationships, leading to a richer interpretation. To analyze the fuzzy coded data we will use correspondence analysis (CA) since it is well adapted to nonnegative data on categorical scales [14]. The application of CA to multivariate categorical data is called multiple correspondence analysis (MCA – see [16] for a comprehensive overview), and since our application is a generalization of MCA to fuzzy coded data, one could call our analytical approach "fuzzy MCA". Second, we focus on the measure of quality of the biplot display in the case of fuzzy MCA. We argue that it is not the quality of display of the fuzzy coded data that should be measured, but rather that of the original data. Defuzzification of the biplot display allows us to reconstruct estimates of the original data, which leads to quality measures directly comparable to PCA's linear approach.

After a summary of the methodology (Section 2) we shall illustrate the benefit of the fuzzy approach using some simulated data, for which the structure is known (Section 3). We shall then prove some new theoretical results about defuzzification of the fuzzy MCA solution to establish correct measures of fit (Section 4) and discuss the scaling properties of fuzzy MCA (Section 5). Finally, we shall apply the methodology to a real data set from meteorology (Section 6) and conclude with a discussion.

## 2. Fuzzy coding and multiple correspondence analysis

CA [1, 2] is a method which graphically displays the rows and columns of a matrix of nonnegative data as points in a biplot-type spatial representation – for technical and practical details see [14], for example. It is a variation on PCA that is suited to ratio-scale data such as counts and proportions, in fact, to any nonnegative tabular data as long as all the data are measured in the same units, including zero/one observations such as presence-absence data. Multiple correspondence analysis (MCA) is the CA of multivariate categorical data, coded as sets of dummy variables in a crisp zero/one form. For example, suppose that an observed data set consists of four categorical variables, each with three categories, then Table 1 shows an example of some of the original multivariate data and their recoding in the form of three dummy variables for each categorical variable. MCA can be defined as the application of CA to the matrix of dummy variables.

To analyze continuous data on heterogeneous scales in the MCA framework, these data can be recoded into categories: for example, if three categories are used, these would represent "low", "medium" and "high" values of the variables. This discrete assignment of a continuous value to a category obviously loses a substantial part of the original information, which can be alleviated by using fuzzy coding. Table 2 shows an example, where instead of three crisp dummy variables coded as 0 or 1 exclusively, there are three fuzzy variables coded with values between 0 and 1 while still adding up to 1 for each variable. Fuzzy coding in the context of correspondence analysis (CA) appeared in French literature in the 1970s – van Rijckevorsel [28] attributes the idea originally to the doctoral thesis of Bordet [3], while the first published papers, to our knowledge, were by Guitonneau and Roux [17] and Gallego [8]. The CA of fuzzy categorical variables, i.e., fuzzy MCA, has not been directly compared to regular dimension-reduction approaches for visualizing continuous variables such as PCA, neither has the issue of measure-of-fit been addressed: these are the motivations for this article.

The basic idea of the data coding is simple. Given a typical cases-by-variables $N \times P$ matrix of continuous data, where variables can be measured on different scales, assign the values of each variable in a fuzzy way into $J$ categories, where the number $J$ is typically 3, 4 or 5 depending on the number of cases in the data and how much detail is required in the results. This is called *fuzzification* of the data – see [21, 22], for example. We have chosen the system of so-called "three-point triangular membership functions", also called piecewise linear functions, or second order B-splines – see [28] for a theoretical account of this topic. In Figure 1 it is shown how the

original values are mapped, via the triangular membership functions, to a five-category recoding ($J = 5$), using the minimum, quartiles and maximum as so-called "hinge points", with the first and last functions not being "shouldered". It is important for our future arguments that the fuzzification be linear and also be invertible, hence our choice of this simple form of membership function. Alternatives to triangular membership functions can be trapezoidal, Gaussian and generalized bell (or Cauchy) membership functions, which have various other theoretical advantages – see [19, 26], for example. These coding aspects and the choice of membership functions have been dealt with extensively in the literature [30, 29, 25].

In our fuzzification scheme each continuous value generates at most two positive nonzero fuzzy values in adjacent categories that add up to 1 – there can be a single positive value of 1 for data that fall exactly on the hinge points. The exact choice of hinge points is not so critical, thanks to the principle of distributional equivalence in CA (see, for example, [14]: pages 37–38). Using triangular membership functions as in Figure 1, the mathematical definition of the fuzzy values $z_1, z_2,\ldots,z_5$, for a five-category fuzzy coding is as follows, where $x$ is the original value on the continuous scale and the hinge points are $m_1,m_2,\ldots,m_5$:

$$z_1(x) = \begin{cases} \dfrac{m_2 - x}{m_2 - m_1}, \text{for } x \le m_2 \\ 0 \text{ otherwise} \end{cases} \quad z_2(x) = \begin{cases} \dfrac{x - m_1}{m_2 - m_1}, \text{for } x \le m_2 \\ \dfrac{m_3 - x}{m_3 - m_2}, \text{for } m_2 < x \le m_3 \\ 0 \text{ otherwise} \end{cases} \quad z_3(x) = \begin{cases} \dfrac{x - m_2}{m_3 - m_2}, \text{for } m_2 < x \le m_3 \\ \dfrac{m_4 - x}{m_4 - m_3}, \text{for } m_3 < x \le m_4 \\ 0 \text{ otherwise} \end{cases}$$

$$z_4(x) = \begin{cases} \dfrac{x - m_3}{m_4 - m_3}, \text{for } m_3 < x \le m_4 \\ \dfrac{m_5 - x}{m_5 - m_4}, \text{for } x > m_4 \\ 0 \text{ otherwise} \end{cases} \qquad z_5(x) = \begin{cases} \dfrac{x - m_4}{m_5 - m_4}, \text{for } x > m_4 \\ 0 \text{ otherwise} \end{cases} \tag{1}$$

The crisp form of such a recoding scheme, that is where each data value is coded strictly into a set of $J$ zero/one dummy variables, leads to what is called an *indicator matrix*, which is the matrix analyzed by regular MCA. The fuzzy coded data can then be referred to as a *fuzzy indicator matrix*

The algorithm for performing fuzzy MCA, that is the CA of a fuzzy indicator matrix, follows that of regular MCA (see, for example, [16: Chapter 2]):

1. Fuzzy code each of the $P$ variables into $J$ fuzzy variables, leading to $PJ$ fuzzy categories; for example, for $J = 5$, use the transformations in (1). The recoded data matrix, denoted by $\mathbf{Z}$, is thus $N \times PJ$, and has grand total $NP$, since each of the $N$ rows has $P$ sets of fuzzy values that each add up to 1.

2. Compute the matrix $\mathbf{P}$ as $\mathbf{Z}$ divided by its grand total: $\mathbf{P} = \mathbf{Z}/(NP)$, with row and column sums of $\mathbf{P}$ defined by $\mathbf{r}$ and $\mathbf{c}$: $\mathbf{r} = \mathbf{P1}$, $\mathbf{c}^{\mathsf{T}} = \mathbf{1}^{\mathsf{T}}\mathbf{P}$, where $\mathbf{1}$ denotes a (column) vector of 1s of appropriate order, and $^{\mathsf{T}}$ denotes vector and matrix transpose. Note in this special case that the elements of $\mathbf{r}$ are constants equal to $1/N$. The elements of $\mathbf{r}$ and $\mathbf{c}$ are called row and column *masses* in CA, and serve as weights in the analysis. $\mathbf{D}_r$ ($N \times N$) and $\mathbf{D}_c$ ($PJ \times PJ$) denote diagonal matrices of the respective masses.

3. Compute the matrix of standardized residuals, $\mathbf{S}$:

$$\mathbf{S} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^{\mathsf{T}})\mathbf{D}_c^{-1/2} = N^{1/2}(\mathbf{P} - (1/N)\mathbf{1}\mathbf{c}^{\mathsf{T}})\mathbf{D}_c^{-1/2}$$

4. Compute the singular-value decomposition (SVD) of $\mathbf{S}$:

$$\mathbf{S} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^{\mathsf{T}}$$

where the singular vectors in $\mathbf{U}$ and $\mathbf{V}$ are normalized as $\mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{I}$, and $\mathbf{D}_\alpha$ is the diagonal matrix of the *singular values*, which are positive and in descending order: $\alpha_1 \geq \alpha_2 \geq \ldots > 0$.

5. Compute the biplot coordinates of the row and column points:

$$\text{rows: } \mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\alpha \quad \text{columns: } \mathbf{\Gamma} = \mathbf{D}_c^{-1/2}\mathbf{V}$$

and use the first two columns, for example, of $\mathbf{F}$ and $\mathbf{\Gamma}$ to make a two-dimensional biplot. In the terminology of CA (see glossary of terminology in Appendix D of [14]), $\mathbf{F}$ contains the *principal coordinates* of the rows and $\mathbf{\Gamma}$ the *standard coordinates* of the columns. The joint plot of $\mathbf{F}$ and $\mathbf{\Gamma}$ constitutes a well-defined biplot (see Chapter 8 of [15]).

Fuzzy MCA shares many properties of regular MCA:

- Each category point receives a weight proportional to its marginal total across all the cases: thus the extreme categories (1 and 5 in the "unshouldered" five-category scheme) receive less weight because the values in the corresponding columns sum to less than the others, which is clear from Figure 1.

- Each set of *J* categories will be centred at the origin of the eventual biplot, where centring is in the weighted average sense, using the weights assigned to the categories.

- Each row point will be at the weighted average position of the category points according to its set of fuzzy values used as weights.

- The solutions for the coordinates are optimal scales (see Section 5); that is, the variance of the case points is maximized along principal axes of the solution, subject to a quadratic identification constraint on the column categories. Compared to an unstandardized PCA of the fuzzy matrix, the main difference that distinguishes the fuzzy MCA approach is that its quadratic constraint involves the weights assigned to the category points.

An important aspect that has not been treated in the literature is how to measure the quality of the fit, or alternatively the error, in biplots of such fuzzy coded data. This will be dealt with in a separate section (Section 4).

## 3. Application to simulated data

In order to demonstrate the difference between the PCA biplot and the fuzzy MCA biplot using data with a known structure, we constructed a data set of $N = 200$ cases and $P = 6$ variables *A*, *B*, *C*, *D*, *E* and *F* as follows:

– First 200 values of two uncorrelated random normal variables were generated, *X1* and *X2*, with mean 0 and variance 1, as well as of four uncorrelated uniform random variables, *U1*, *U2*, *U3*, *U4*, on the interval [0,1].

– $A = 20 + 3*X1 + 5*X2 + 2*(U1 - ½)$

– $B = 20 + 5*X1 - X2 + 2*(U2 - ½)$

– $C = U3$

– $D = \log(A)$

– $E = U4$

– $F = (A - 20)^2$

Thus $A$ and $B$ have been generated as linear combinations of the same two normal variables, onto which uniform noise has been added. Their theoretical correlation can be calculated as 0.336. $D$ is the logarithm of $A$ and $F$ is a quadratic function of $A$. $C$ and $E$ are random variables that have theoretically zero correlations with all of the variables. The correlation between $A$ and its log-transformed value $D$ is expected to be very high, while that between $A$ and its quadratically transformed value $F$, which has a minimum at the mean value of $A$, is expected to be low. The sample correlation matrix is given in Table 3.

Figure 2 shows the PCA biplot of these data, where the 200 cases are indicated by dots and the six variables by vectors. The data are standardized, as is customary in PCA, because of the widely differing scales of the variables. In such a biplot the projections of the row points onto the directions defined by the variable vectors gives an approximation, up to a scaling factor, of the original standardized data. Vectors that point in the same direction are thus positively correlated (e.g., $A$ and $D$) while those pointing in opposite directions are negatively correlated (e.g., $C$ and $E$). The impression given by the biplot agrees only partially with the way the data were constructed. The high correlation between $A$ and $D$ is apparent, as well as the lower, but nevertheless positive, correlation between $B$ and $A$ and between $B$ and $D$. But the two random "noise" variables $C$ and $E$ define a second dimension where they appear to be negatively correlated and the quadratic variable $F$ appears to be correlated with them.

On the other hand, consider the fuzzy MCA in Figure 3, which shows one point for each of the fuzzy categories (computations are performed using the **ca** package in R [23, 24]). In Figure 4, the five categories of each variable have been connected to show their trajectories in Figure 3 more clearly. The categories of $A$ and $D$ follow almost identical paths through the display, showing a curved pattern called the arch effect, which is typical in CA owing to its simplex geometry (see, for example, [14]: chap. 2). The category values of these two variables are almost identical because $D$ is a monotone function of $A$. In CA terminology we would use "association", rather than "correlation", to describe the relationship between $A$ and $D$, because the nature of the association can take many different forms, not only the linear form inherent in

the PCA approach. Thus the association between *A* and *D* is very strong, and it is positive because the association of the categories, from category 1 to category 5, is almost identical. The categories of *B* also follow the same curved path, but not as closely as *A* and *D*, with the same low to high (1 to 5) trajectory, hence we would conclude that *B* is positively associated with *A* and with *D*, but not as strongly as that between *A* and *D*. The quadratic variable *F* takes on a completely different pattern, with low values associated with the middle categories of *A*, *B* and *D*, and high values when these variables are either low or high – this is exactly the nature of the quadratic relationship of the constructed variable. Finally, the two "noise" variables *C* and *E* make small erratic trajectories near the origin of the display, agreeing with the fact that they have no association with any of the variables.

This example shows clearly the difference between the two approaches, and how the fuzzy coded version can come to a conclusion which is practically identical to the way the data were constructed, whereas the PCA, which can only visualize linear relationships, leads to several incorrect conclusions.

## 4. Defuzzification and the measure of fit

So far we did not comment on the goodness-of-fit of the two displays to the data. In the case of PCA it is customary to give percentages of variance explained by each dimension, and their sum for the two-dimensional solution. For Figure 2 it is 37.1% and 17.6% respectively, totalling 54.8% explained. The value of 54.8% has the same interpretation as in regression analysis – of the variance of the six variables analyzed, 54.8% of their variance has been explained by the two dimensions, or principal axes, of the PCA, and 45.3% is unexplained residual or "error" variance. It is clear from Figure 2 that too much prominence is being given to the two "noise" variables *C* and *E* – we shall return to this point later in this section.

In CA the idea is the same, namely to measure how much variance in the data, called "inertia" in CA, is explained by the solution. But the fuzzy MCA gives very low percentages of explained variance: 16.4% and 14.5%, totalling 30.9%. It is well-known that regular MCA gives very pessimistic estimates of explained variance. For example, Lebart [20] states that in MCA the "percentages of variance are misleading measures of information", and the same is true for fuzzy MCA. It might be thought that this is because fuzzy MCA embeds the data in a much higher-dimensional space (24 dimensional – the space of 30 fuzzy dummy variables that have 6

linear restrictions, each set of five summing to 1) compared to PCA (only 6-dimensional), so the chances of good reconstruction of the data in a two-dimensional solution are better for PCA. But this is only one reason, and a more important reason is that the rationale for the measure of fit is wrong. We are not interested in reconstructing the fuzzy coded data, which is what the 30.9% measures, but rather in reconstructing the original data. Fortunately, this can be done through a process of *defuzzification* of the solution.

Thanks to the form of the triangular membership functions, the fuzzy coded data can be transformed back to the original data by taking a linear combination of the hinge points, using the fuzzy values $z_1,z_2,\ldots,z_5$ as coefficients. Since these coefficients are nonnegative and add up to 1, this inverse transformation, called defuzzification, can be thought of as weighted averaging:

$$x = z_1 m_1 + z_2 m_2 + z_3 m_3 + z_4 m_4 + z_5 m_5 \qquad (2)$$

Now defuzzification can also be applied to the five numbers $\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_5$ which are estimated from the fuzzy MCA solution, which have the same property that they add up to 1, leading to estimates $\hat{x} = \hat{z}_1 m_1 + \hat{z}_2 m_2 + \cdots \hat{z}_5 m_5$ of the original data. These defuzzified estimates have favourable properties, proved in the Appendix, which we summarize here.

The matrix $\hat{\mathbf{X}}$ containing the estimates $\hat{x}$ can be written as $\hat{\mathbf{X}} = \hat{\mathbf{Z}}\mathbf{M}$, where $\hat{\mathbf{Z}}$ is the larger matrix of the estimates $\hat{z}$, and $\mathbf{M}$ is a full-rank matrix (see Appendix) – hence the rank of $\hat{\mathbf{X}}$ is equal to that of $\hat{\mathbf{Z}}$, which would be equal to 2 for a 2-dimensional solution. It is these estimates $\hat{x}$ that will be compared with the original data values $x$ in order to measure goodness of fit (or lack thereof).

A further property of the defuzzification is that the means of the estimates $\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_5$ are the same as the means of the original $z_1, z_2, \ldots, z_5$ – thus the defuzzified means recover the means of the original variables (see Appendix, result 1). For example, the set of five means for the fuzzy variables corresponding to the first variable *A* is [0.0723 0.2878 0.2725 0.2938 0.0737] and *A*´s hinge points are [4.68 15.64 20.37 24.80 34.06]; thus the computation 0.0723×4.68 + 0.2878×15.64 + ⋯ + 0.0737×34.06 = 20.18, the mean of *A*.

Using defuzzified estimates of the data is a feasible way to obtain a measure of fit because it can be proven that the reconstructed data $\hat{x}$ and the residuals $x - \hat{x}$ lie in orthogonal subspaces (see

Appendix, result 2), just as in PCA, showing that the dimensions are nested. Hence the decomposition of total variance into explained plus residual variance as well as the summation of percentages over single dimensions are valid (see Appendix, result 3). One of our objectives is to compare the fuzzy MCA approach with PCA, so this property of the defuzzified solution is crucial because it allows quantification of the success of each dimension of the fuzzy MCA, in parallel with the classical PCA approach.

After defuzzifying the estimations from the two-dimensional fuzzy MCA of Figure 3, the percentage of variance explained turns out to be 42.4%. This is more than the 30.9% explained variance of the fuzzy data, but still less than the 54.8% explained variance for the PCA. However, this is not surprising, since the fuzzy MCA is trying to account for nonlinearities in the data whereas the PCA only explains the linear part. It is interesting to see this measure of fit broken down in terms of how much variance of the individual variables is being explained in the two approaches:

| | *A* | *B* | *C* | *D* | *E* | *F* | overall |
|---|---|---|---|---|---|---|---|
| PCA | 90.4% | 39.1% | 56.6% | 92.8% | 14.4% | 34.8% | 54.8% |
| Fuzzy MCA | 87.8% | 28.3% | 0.5% | 80.6% | 4.0% | 53.5% | 42.4% |

It is clear that the fuzzy approach does much better in ignoring the "noise" variables, and that the better fit in the PCA is almost principally due to improved reconstruction of the random part of the data. If the explained variances are averaged over four variables, omitting those for *C* and *E*, the percentages are 64.3% and 63.8% respectively – there is no longer a big difference, and coupled with the fact that the fuzzy MCA gave an interpretation in line with the way the data were simulated, it is clearly a superior approach.

Because of the orthogonality of the dimensions of the defuzzified solution, percentages of variance can be computed for individual axes and summed – these individual percentages, given in Figure 3, turn out to be 7.3% and 32.6% for the first and second axes respectively. The second axis, which shows the close ordinal relationship between variables *A*, *B* and *D*, explains more variance in the original variables than the first, which accounts for the nonlinear relationship of variable *F* with variable *A*, and thus in turn with variables *B* and *D*. As far as the within-variable associations are concerned, at the fuzzy category level, the first axis is slightly more important than the second (16.4% compared to 14.5%) since it accounts for the category-level relationships within four variables (*A*, *B*, *D* and *F*).

## 5. Optimal scaling properties

Both PCA and fuzzy MCA optimize the variance explained on successive dimensions in their own way, but as shown above PCA optimizes the explanation of linear relationships while fuzzy MCA optimizes that of linear and nonlinear relationships. The positions of the cases on each dimension, called *scores*, are made of components due to each of the original variables, and each of these components can be correlated with the overall scores to quantify how well the dimension is agreeing with the component variables. We illustrate this idea, reminiscent of item analysis, in the case of the first dimension of each approach.

In the case of PCA, the score is a linear combination of the original variables (normalized), with the coefficients being those of the first eigenvector of the analysis. Each component is then just a constant times the variable, so the component–score correlations are just those between the columns of the original data matrix and the score vector. In the case of fuzzy MCA, the set of $J$ categories have scale values on the first dimension and the component of the score is the corresponding linear combination of the scale values. For example, the five scale values on the first dimension for variable *A* of the simulation are [ 1.594 0.415 –2.006 0.619 1.765 ] (i.e., the coordinates on the first dimension of category points *A1*,*A2*,…,*A5* in Figure 3 and first graph in Figure 4), and the first case has fuzzy data [ 0 0 0 0.771 0.229 ] on *A*; hence *A*'s component value to the score of case 1 is $0.771 \times 0.619 + 0.229 \times 1.765 = 0.881$. Once the components of all six variables are computed in the same way, the score (i.e., the position of case 1 on the first axis of Figure 3), is the average of these six values.

Table 4 shows the correlations between the six components of dimension 1 and the score on that dimension for PCA and fuzzy MCA. Here it is clear that fuzzy MCA is doing better overall than PCA, especially with respect to the nonlinearly related variable *F*. Overall performance is measured by the average squared correlation as well as Cronbach's $\alpha$ reliability coefficient. Since the performance of the fuzzy MCA appears to be slightly enhanced by the slightly higher correlations with the "noise" variables, we repeated the whole exercise without these variables, also shown in Table 4, and fuzzy MCA still performs better than the PCA.

## 6. Application to real data

Table 5 contains average values of five meteorological variables for 40 cities of Turkey, based on measurements taken in 2004 [27] – note that we use this example as an illustration of our approach to real data rather than a substantive meteorological application.

Figure 5 shows the PCA biplot of Table 5, where the data have been standardized, explaining 75.6% of the variance, and Figure 6 shows the fuzzy MCA of the same data set. Because there are only 40 cases, we reduced the number of fuzzy categories to three, with hinges at the minimum, median and maximum of each variable. The explained variance after defuzzification is 69.4%, not far behind that of PCA seeing that the data are embedded by the fuzzy coding into twice as many dimensions (10 in the case of the fuzzy MCA, 5 in the case of PCA). As in the simulated data the coding scheme allows the "low", "middle" and "high" categories of the five variables, labelled 1, 2 and 3, to associate with one another according to the inter-variable associations, as opposed to being constrained to be linear as in PCA. In Figure 5, for example, humidity and sunshine look strongly negatively correlated, while precipitation appears to correlate weakly positive with humidity and weakly negative with sunshine. Figure 6 tells a more interesting story: low sunshine, high humidity and high rainfall actually associate strongly, with Rize being the archetypal example and then Samsun, Trabzon, Zonguldak and Göztepe, etc. But at the other end low rainfall, low humidity and high sunshine fan out in different directions, high sunshine and low humidity negative on the vertical dimension for Diyarbakır, Siirt and Gaziantep, for example, and low precipitation positive on the first dimension for cities like Konya, Kırşehir, Erzincan, Afyon and Van, which are otherwise "middle" on sunshine and humidity.

## 7. Discussion and conclusions

Both PCA and fuzzy MCA operate on the same data in different forms, but it is the fuzzy coding that takes the data into a higher-dimensional space in which higher-order associations can be explored, whereas PCA is only capable of explaining linear relationships. Even though there are more parameters in the fuzzy MCA for a fixed dimensionality of the solution, it appears to perform slightly worse than PCA in reconstructing the original data, but this is again because it has more to explain.

Another way to compare the PCA and fuzzy MCA on a more equal footing is to use only two membership functions that code just the linear information, also known as doubling in CA [12, 22]. The two endpoints $m_1$ (minimum) and $m_2$ (maximum) are used as hinges, and the membership functions for the "positive" and "negative" doubled variables are simply:

$$z_+(x) = (x - m_1)/(m_2 - m_1) \qquad z_-(x) = (m_2 - x)/(m_2 - m_1) = 1 - z_+(x) \qquad (3)$$

These two values sum to one and code how close the data are to the respective endpoints; defuzzification is achieved as before by weighted averaging: $x = z_-(x)\, m_1 + z_+(x)\, m_2$. In the meteorological example the fuzzy coded data are now 5-dimensional, as for PCA, and after defuzzification of the 2-dimensional solution, the measure of fit is 75.0%, just fractionally lower than the optimal PCA fit – this illustrates that the two methodologies essentially coincide in their quest for linear associations. This slight difference in explained variance is due to the fact that the fuzzy MCA standardizes the data differently from PCA as a result of the chi-square metric – Greenacre [12], pp.175–179, calls this standardizing by "polarization" rather than by the variance.

Our examples have consisted of continuous variables only, but in the French literature the justification for fuzzy coding has mostly been to permit continuous variables to be analysed jointly with categorical ones – see, for example, [9, 17]. The situation of mixed discrete-continuous data presents the particular problem for defining measures of fit which take into account in an equitable way the different characteristics of logical and fuzzy coding. Various approaches are possible. For example, Gower [10] defines a distance function which attempts to equalize the contributions of the different variables to the total variance. Escofier and Pagès [6] define a doubling transformation of continuous data, different from (3) above, which is more suitable for analysing continuous data jointly with dichotomous categorical data. They consider groups of homogeneous variables, for example the group of continuous variables (in original form or fuzzified) and the group of categorical variables, they then standardize them internally using the first eigenvalue as a surrogate for the table variance, and then proceed to joint analysis. Most of these approaches can be reduced to a type of reweighting of the variables to equalize in some sense their contributions to the joint analysis.

The main and novel contribution of this paper is to show how the solution of the fuzzy analysis using CA, which is essentially a nonlinear treatment of the data, can be defuzzified to give results that can be directly compared to those of the linear approach in PCA. The results proved

in the Appendix underpin the use of the defuzzified solution to measure the fit of the result to the original data, since this solution gives an orthogonal decomposition of variance, just as in PCA. These new results about the defuzzification and consequent convenient measure of fit are a consequence of the particular triangular membership functions used, which are linear and invertible. One can use other membership functions for the fuzzification of the data, of course, but for nonlinear membership functions the favourable defuzzification properties will not hold. We can also not allow "shoulders" in the triangular membership functions, where the end categories are a constant value of 1 below and above chosen extreme values, because this would make the coding non-invertible.

An important aspect of the recoding of the data into fuzzy categories, demonstrated clearly by the simulated data set but also in the real one, is that the method of fuzzy MCA can visualize nonlinear relationships between variables – this property holds for all forms of membership function. Since one of our objectives has been to compare the biplots of fuzzy coded data with the standard PCA biplot, this flexibility in the type of relationship that can be diagnosed in the data is a distinct advantage over linear PCA.

## Acknowledgments

The second author thanks the BBVA Foundation for financial support in this research, as well as the Ministry of Science and Innovation grants MTM2008-00642 and MTM2009-09063. The reports of two referees on the first version of this paper led to significant improvements and are hereby acknowledged with thanks.

## References


[1] J.-P. Benzécri, L'Analyse des Données. Tôme II : L'Analyse des Correspondances, Dunod, Paris, 1973.

[2] J.-P. Benzécri, Correspondence Analysis Handbook, Marcel Dekker, Amsterdam, 1992.

[3] C. Bordet, Etudes de Données Geophysiques, Doctoral thesis (3ème cycle), Université de Paris VI, France, 1973.

[4] F. Chevenet, S. Dolédec, D. Chessel, A fuzzy coding for the analysis of long term ecological data, Freshwater Biology 31 (1994) 295-209.

[5] B. Escofier, Traitement simultané de variables qualitatives et quantitatives en analyse factorielle, Les Cahiers de l'Analyse des Données 4 (1979) 137–146.

[6] B. Escofier, J. Pagès, Le traitement des variables qualitatives et des tableaux mixtes par analyse factorielle multiple, in : E. Diday (Ed.), Data Analysis and Informatics IV, North Holland, Amsterdam, 1986, pp. 179–191.

[7] K.R. Gabriel, The biplot graphic display of matrices with application to principal component analysis, Biometrika 58 (1971) 453–467.

[8] F.J. Gallego, Codage flou en analyse des correspondances, Les Cahiers de l'Analyse des Données 7 (1982) 413–430.

[9] B.M. Ghermani, C. Roux, M. Roux , Sur le codage logique des données héterogènes, Les Cahiers de l'Analyse des Données 1 (1977) 115–118.

[10] J.C. Gower. A general coefficient of similarity and some of its properties, Biometrics 27 (1971) 857–871.

[11] J.C. Gower, D. Hand, Biplots, Chapman and Hall, London, 1996.

[12] M.J. Greenacre, Theory and Applications of Correspondence Analysis, Academic Press, London, 1984.

[13] M.J. Greenacre, From simple to multiple correspondence analysis, in: M.J. Greenacre, J. Blasius (Eds.), Multiple Correspondence Analysis and Related Methods, Chapman & Hall / CRC, London, 2006, pp. 41–76.

[14] M.J. Greenacre, Correspondence Analysis in Practice. Second Edition. Chapman & Hall / CRC, London, 2007.

[15] M.J. Greenacre, Biplots in Practice, BBVA Foundation, Madrid, 2010. Freely downloadable from http://multivariatestatistics.org.

[16] M.J. Greenacre, J. Blasius, (Eds.), Multiple Correspondence Analysis and Related Methods, Chapman & Hall / CRC, London, 2006.

[17] A. Guitonneau, M. Roux, Sur la taxinomie de genre Erodium, Les Cahiers de l'Analyse des Données 1 (1977) 97–113.

[18] E. Hadjicharalambous, K.L. Kalburtji, A.P. Mamolos, A. P., Fuzzy set analysis and canonical correspondence analysis of soil arthropods (Coleoptera, Isopoda) in organic and conventional agroecosystems, in: W. J. Horst et al. (Eds.), Plant Nutrition – Food Security and Sustainability of Agro-Ecosystems, Kluwer, Amsterdam, 2001, pp. 1020–1021.

[19] J.S.R. Jang, C.T. Sun, E. Mizutani, Neuro Fuzzy and Soft Computing: a Computational Approach to Learning and Machine Intelligence, Prentice Hall, USA, 1997.

[20] L. Lebart, Validation techniques in multiple correspondence analysis, in : M.J. Greenacre, J. Blasius (Eds.), Multiple Correspondence Analysis and Related Methods, Chapman & Hall / CRC, London, 2006, pp. 179–195.

[21] P. Loslever, S. Bouilland, Marriage of fuzzy sets and multiple correspondence analysis: examples with subjective interval and biomedical signals, Fuzzy Sets and Systems 107 (1999) 255–275.

[22] F. Murtagh, Correspondence Analysis and Data Coding with R and Java, Chapman & Hall / CRC., London, 1999.

[23] O. Nenadić, M.J. Greenacre, Computation of multiple correspondence analysis, in: M.J. Greenacre, J. Blasius (Eds.), Multiple Correspondence Analysis and Related Methods, Chapman & Hall / CRC, London, 2006, pp. 523–551..

[24] O. Nenadić, M.J. Greenacre, Correspondence analysis in R, with two- and three-dimensional graphics: the ca package, Journal of Statistical Software 20 (2007), URL http://www.jstatsoft.org/v20/i03/, last accessed August 18 2010.

[25] S. Şentürk, Fuzzy Logic Approach In Experimental Design, Doctoral thesis, Department of Statistics, Faculty of Science, Anadolu University, Eskişehir, Turkey, 2006.

[26] M. Smithson, J. Verkuilen, J., Fuzzy Set Theory, Sage, California, 2006.

[27] Turkey's Statistical Yearbook, 2004, Turkish Statistical Institute, Ankara. URL www.hermesprojekt.de/downloads/e-book/jahrbuch_tuerkei.pdf, last accessed August 18 2010.

[28] J.L.A. van Rijckevorsel, Fuzzy coding and B-splines, in: J.L.A. van Rijckevorsel, J. de Leeuw (Eds.), Component and Correspondence Analysis, Wiley, Chichester, UK, 1988, 33–54.

[29] J. Verkuilen, Assigning membership in a fuzzy set analysis, Sociological Methods and Research 33 (2005) 462–496.

[30] Q. Zhou, M.K. Purvis, N.K. Kazabov, A membership selection function method for fuzzy neural networks, Information Science Discussion Paper Series, 97/15, University of Otago, New Zealand, 1997.

# APPENDIX

## Some theoretical results about CA of fuzzy coded data

Consider the $N \times P$ data matrix $\mathbf{X}$ and corresponding fuzzy coded matrix $\mathbf{Z}$, using the triangular membership functions (1). In the CA of $\mathbf{Z}$ the row masses are all equal to $1/N$ and the column masses $c_j$ in vector $\mathbf{c}$ are the column averages divided by $P$; $\mathbf{D}_c$ is the diagonal matrix of the column masses.

From the definition of the triangular membership functions (1) we can write the relationship between $\mathbf{X}$ and $\mathbf{Z}$ as the following linear defuzzification formula:

$$\mathbf{X} = \mathbf{ZM} \tag{4}$$

where:

$$\mathbf{M} = \begin{bmatrix} \mathbf{m}_1 & 0 & \cdots & \cdots & 0 \\ 0 & \mathbf{m}_2 & 0 & \cdots & 0 \\ 0 & 0 & \mathbf{m}_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \mathbf{m}_P \end{bmatrix}$$

where $\mathbf{m}_j$ is a vector of the hinge parameters of the membership function for the $j$-th variable (in our example these are the minimum, three quartiles and maximum )

The defuzzified approximate values, obtained from the reconstructed values in $\hat{\mathbf{Z}}$, are similarly obtained as $\hat{\mathbf{X}} = \hat{\mathbf{Z}}\mathbf{M}$.

The CA of $\mathbf{Z}$, defined in Section 2, implies the same type of decomposition as for the crisp equivalent $\mathbf{Z}$ (see [13: chapter 2]), which can be written as:

$$\mathbf{Z} = P(\mathbf{1}\mathbf{1}^{\mathsf{T}} + \sqrt{N}\mathbf{U}\mathbf{D}_{\alpha}\mathbf{V}^{\mathsf{T}}\mathbf{D}_c^{-1/2})\mathbf{D}_c \quad \text{where } \mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{I} \tag{5}$$

To estimate the data from a $K$-dimensional approximation, we use the first $K$ columns of $\mathbf{U}$ and of $\mathbf{V}$, denoted by $\mathbf{U}_{[K]}$ and $\mathbf{V}_{[K]}$ respectively, and the first $K$ singular values in diagonal matrix $\mathbf{D}_{\alpha[K]}$:

$$\hat{\mathbf{Z}} = P(\mathbf{1}\mathbf{1}^{\mathsf{T}} + \sqrt{N}\mathbf{U}_{[K]}\mathbf{D}_{\alpha[K]}\mathbf{V}_{[K]}^{\mathsf{T}}\mathbf{D}_c^{-1/2})\mathbf{D}_c \tag{6}$$

The following results can then be proved.

1. The means $\bar{\mathbf{x}}$ of $\mathbf{X}$ are the same as those of $\hat{\mathbf{X}}$ and are equal to the defuzzified averages of the columns of $\mathbf{Z}$ (or of $\hat{\mathbf{Z}}$)

   *Proof:* $\bar{\mathbf{x}}^{\mathsf{T}} = \frac{1}{N}\mathbf{1}^{\mathsf{T}}\mathbf{X} = \frac{1}{N}\mathbf{1}^{\mathsf{T}}\mathbf{ZM}$

   Hence the means $\bar{\mathbf{x}}$ are the defuzzified column means of $\mathbf{Z}$.

   Since $\mathbf{Z}$ and $\hat{\mathbf{Z}}$ have the same column means (this is a standard property of CA approximations), the above proof can be reversed to show that the defuzzified means of $\hat{\mathbf{Z}}$, i.e., $\frac{1}{N}\mathbf{1}^{\mathsf{T}}\hat{\mathbf{Z}}\mathbf{M} = \frac{1}{N}\mathbf{1}^{\mathsf{T}}\hat{\mathbf{X}}$, is also $\bar{\mathbf{x}}^{\mathsf{T}}$, that is the means of $\mathbf{X}$ and $\hat{\mathbf{X}}$ are the same.

2. The deviations between the data and the data estimated by defuzzifying the reconstructed data from fuzzy MCA are orthogonal to these estimates.

*Proof:* The result holds first for the fuzzy coded matrix and its estimated values from CA. Suppose that the subindex $_{[-K]}$ indicates the remaining singular components from the ($K$+1)-th onwards, so that for example $\mathbf{U} = [\ \mathbf{U}_{[K]}\ \mathbf{U}_{[-K]}\ ]$. Then from (5) and (6)

$$(\mathbf{Z}-\hat{\mathbf{Z}})^{\mathsf{T}}\hat{\mathbf{Z}} = P^2\sqrt{N}\mathbf{D}^{1/2}\mathbf{V}_{[-K]}\mathbf{D}_{\alpha[-K]}\mathbf{U}_{[-K]}^{\mathsf{T}}(\mathbf{1}\mathbf{1}^{\mathsf{T}} + \sqrt{N}\mathbf{U}_{[K]}\mathbf{D}_{\alpha[K]}\mathbf{V}_{[K]}^{\mathsf{T}}\mathbf{D}_c^{-1/2})\mathbf{D}_c = \mathbf{0}$$

because $\mathbf{U}_{[-K]}^{\mathsf{T}}\mathbf{1} = \mathbf{0}$ (the rows have equal masses, so their coordinates have arithmetic mean zero) and $\mathbf{U}_{[-K]}^{\mathsf{T}}\mathbf{U}_{[K]} = \mathbf{0}$ (orthogonality of the singular vectors). The linear operation of defuzzifying does not change this property:

$$(\mathbf{X}-\hat{\mathbf{X}})^{\mathsf{T}}\hat{\mathbf{X}} = \mathbf{M}^{\mathsf{T}}(\mathbf{Z}-\hat{\mathbf{Z}})^{\mathsf{T}}\hat{\mathbf{Z}}\mathbf{M} = \mathbf{0}$$

3. As a result of 2. the sum-of-squares of $\mathbf{X}$ decomposes into two components:

$$\text{trace}[\mathbf{X}\mathbf{X}^{\mathsf{T}}] = \text{trace}[(\mathbf{X}-\hat{\mathbf{X}})(\mathbf{X}-\hat{\mathbf{X}})^{\mathsf{T}}] + \text{trace}[\hat{\mathbf{X}}\hat{\mathbf{X}}^{\mathsf{T}}]$$

and this property is maintained for any common centring and standardization of $\mathbf{X}$ – in our application centring is with respect to the common means and standardization with respect to the standard deviations of the original variables.

The above results generalize to any fuzzy coding for which the defuzzification transformation is linear, as in (4).

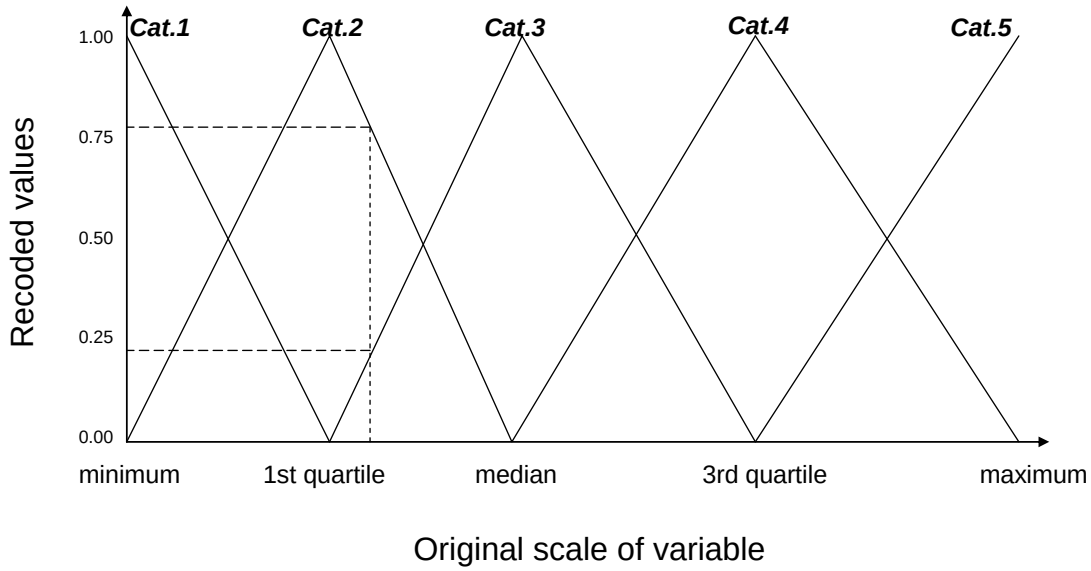


*Figure 1:* Triangular membership functions to code a continuous variable (horizontal axis) into five fuzzy categorical variables.

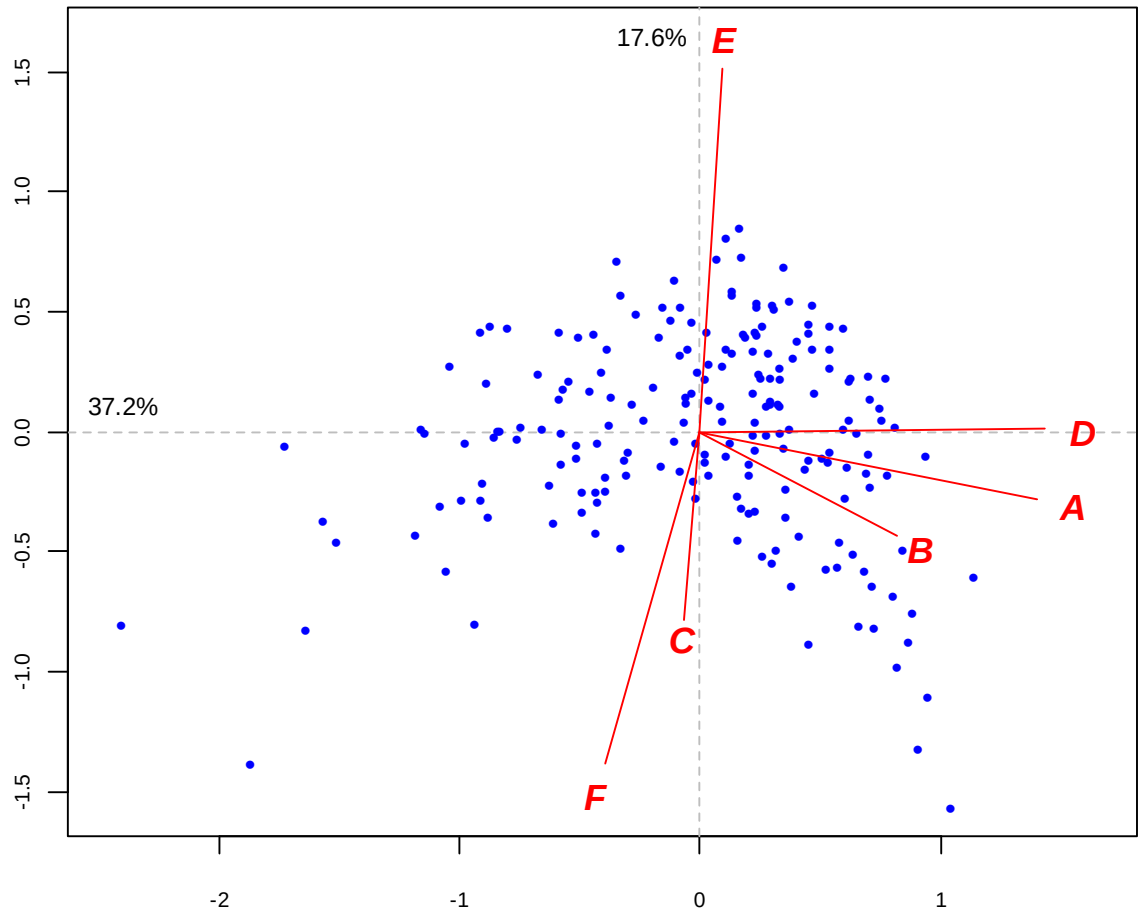


*Figure 2:* PCA biplot of the simulated data. The two "noise" variables *C* and *E* play prominent roles on the second dimension and variable *F*, which is a quadratic function of *A*, appears to be correlated with the second dimension.

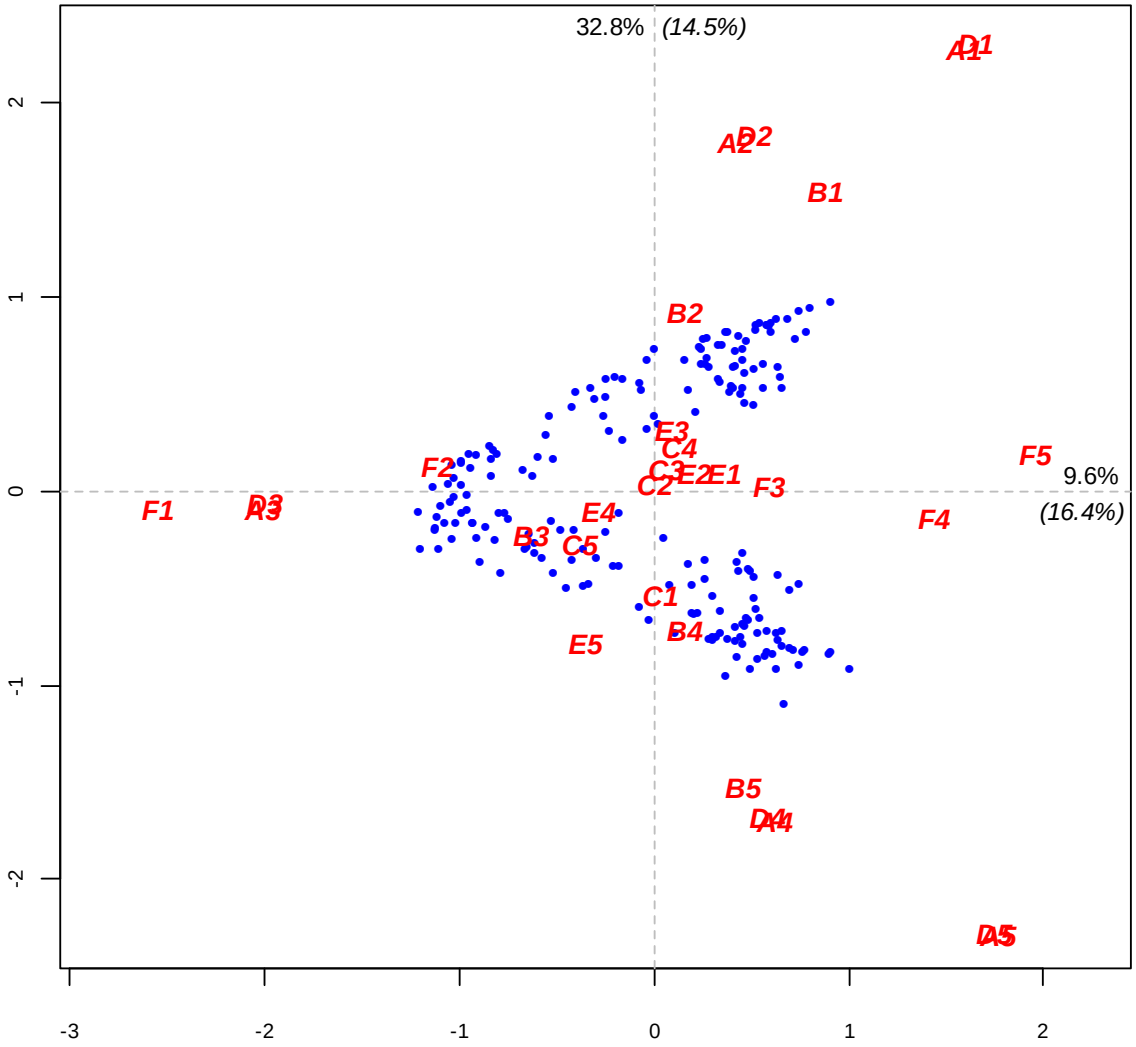


*Figure 3:* Fuzzy MCA biplot of the simulated data. Each variable is represented by five points: for example, *A1*, *A2*, *A3*, *A4* and *A5* are categories 1 to 5 of variable *A*. The percentages of variance in parentheses are those obtained for the fuzzy coded data on each dimension, totalling 30.9%. The other percentages (32.8% and 9.6%, totalling 42.4%) are for the defuzzified solution – described in Section 4 – where the second axis turns out to explain more variance than the first.

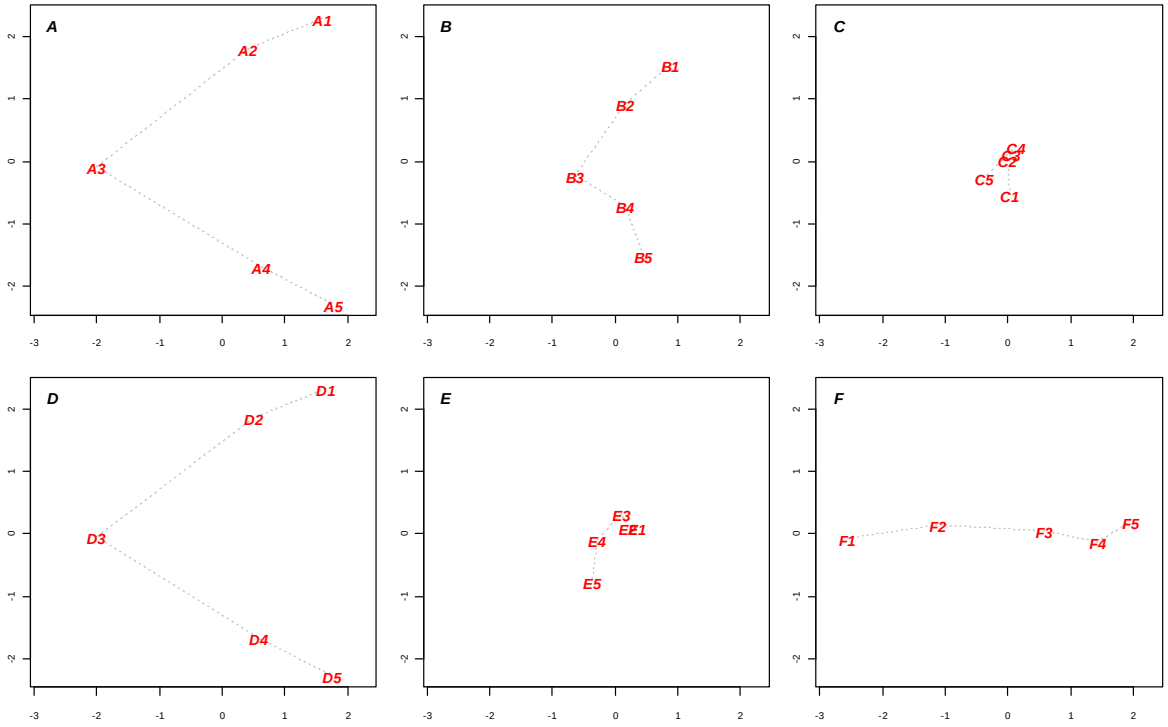


*Figure 4:* The trajectories of the six variables in Figure 3, linking the categories 1 to 5 of each variable separately for comparison.

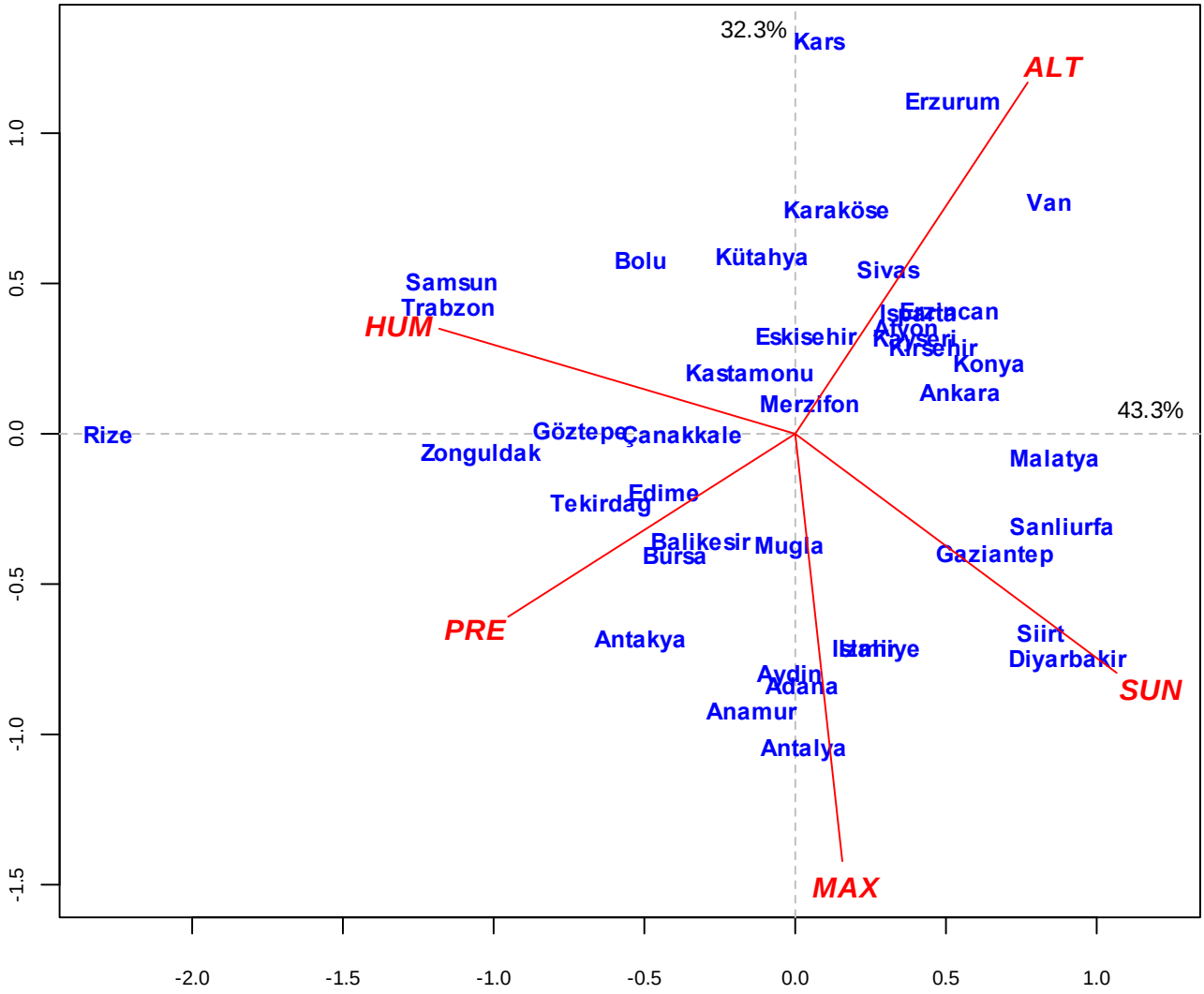


*Figure 5:* PCA biplot of Table 5, where variables have been standardized. 75.6% of the variance is explained by the first two dimensions.

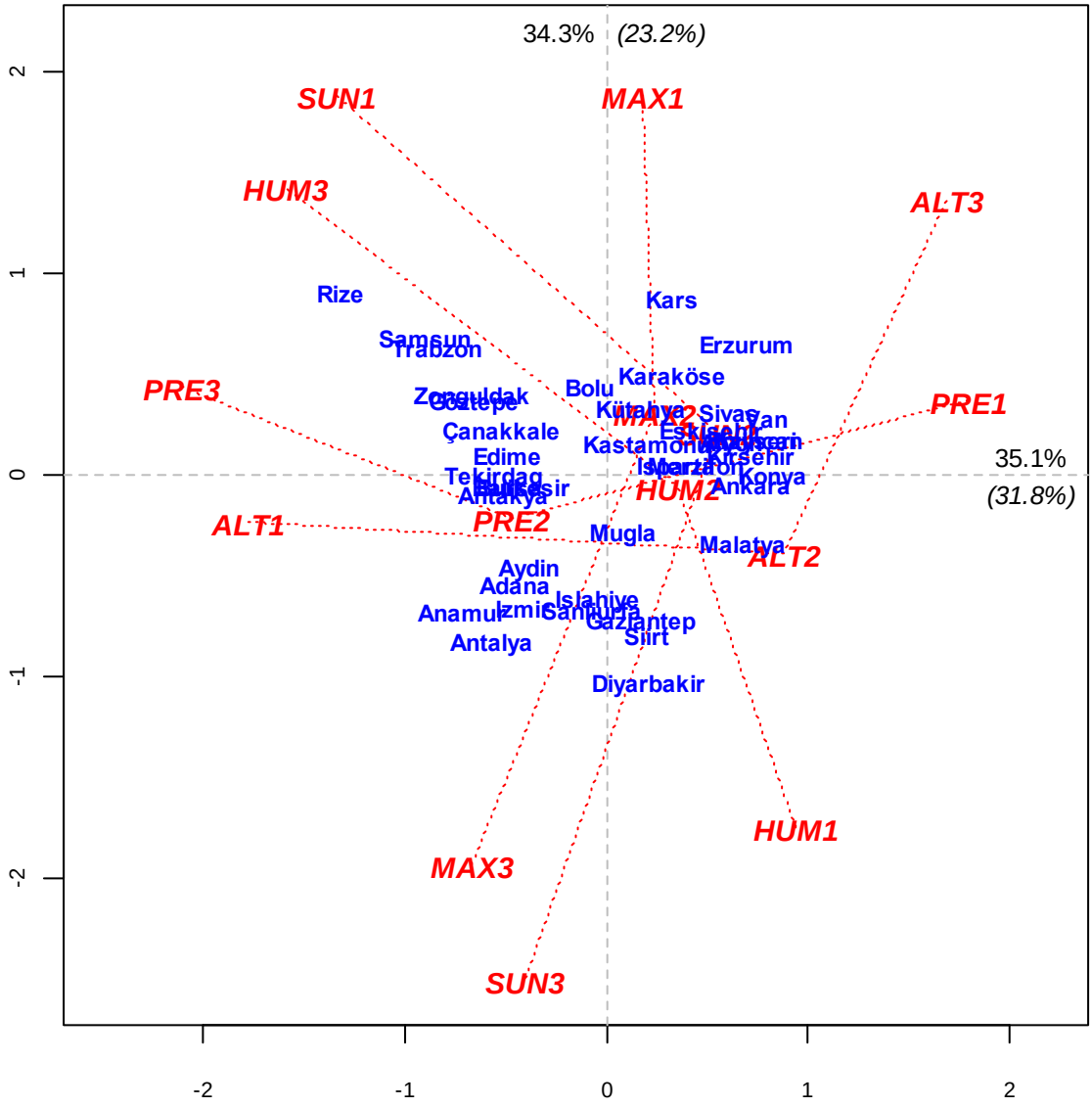


*Figure 6:* Fuzzy MCA biplot of Table 5, where each variable has been fuzzy coded into three categories. After defuzzification 69.4% of the variance of the original standardized data is explained in two dimensions, whereas 55% of the fuzzy coded data is explained (percentages in parentheses),

| *A* | *B* | *C* | *D* | *A1* | *A2* | *A3* | *B1* | *B2* | *B3* | *C1* | *C2* | *C3* | *D1* | *D2* | *D3* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 2 | 1 | 2 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 1 | 2 | 1 | 3 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 3 | 1 | 2 | 2 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |

*Table 1:* On the left, in each row, some observations on four categorical variables, *A* to *D*, with three categories each, and on the right, their coding into three dummy variables for each variable (crisp coding).

| *A* | *B* | *C* | *D* | *A1* | *A2* | *A3* | *B1* | *B2* | *B3* | *C1* | *C2* | *C3* | *D1* | *D2* | *D3* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.1 | 21.5 | 5.6 | 9.6 | 0.83 | 0.17 | 0.00 | 0.00 | 0.12 | 0.88 | 0.21 | 0.79 | 0.00 | 0.39 | 0.61 | 0.00 |
| 3.7 | 15.0 | 5.8 | 8.5 | 0.00 | 0.78 | 0.22 | 0.33 | 0.67 | 0.00 | 0.16 | 0.84 | 0.00 | 0.77 | 0.23 | 0.00 |
| 2.6 | 16.1 | 6.3 | 13.2 | 0.94 | 0.06 | 0.00 | 0.04 | 0.96 | 0.00 | 0.00 | 0.79 | 0.21 | 0.00 | 0.15 | 0.85 |
| . | . | . | . | . | . | . | . | . | . | . | . | . | | | |
| . | . | . | . | . | . | . | . | . | . | . | . | . | | | |
| . | . | . | . | . | . | . | . | . | . | . | . | . | | | |

*Table 2:* On the left, in each row, some observations on four continuous variables, *A* to *D*, and on the right, their fuzzy coding into three categories.

| | *A* | *B* | *C* | *D* | *E* | *F* |
|---|---|---|---|---|---|---|
| *A* | 1.0000 | 0.3623 | -0.0179 | 0.9742 | 0.0194 | -0.0636 |
| *B* | 0.3623 | 1.0000 | -0.0251 | 0.3500 | 0.0413 | -0.0077 |
| *C* | -0.0179 | -0.0251 | 1.0000 | -0.0200 | -0.0378 | -0.0031 |
| *D* | 0.9742 | 0.3500 | -0.0200 | 1.0000 | 0.0204 | -0.2738 |
| *E* | 0.0194 | 0.0413 | -0.0378 | 0.0204 | 1.0000 | -0.0593 |
| *F* | -0.0636 | -0.0077 | -0.0031 | -0.2738 | -0.0593 | 1.0000 |

*Table 3:* (Linear) sample correlations between the six variables used in the simulation study.

| | PCA | fuzzy MCA |
|---|---|---|
| all six variables | | |
| *A* | 0.949 | 0.992 |
| *B* | 0.557 | 0.346 |
| *C* | 0.044 | 0.123 |
| *D* | 0.971 | 0.991 |
| *E* | 0.064 | 0.167 |
| *F* | 0.265 | 0.988 |
| average squared correlation | 0.371 | 0.517 |
| Cronbach's α | 0.662 | 0.814 |
| without "noise" variables | | |
| *A* | 0.950 | 0.992 |
| *B* | 0.556 | 0.343 |
| *D* | 0.972 | 0.992 |
| *F* | 0.263 | 0.990 |
| average squared correlation | 0.557 | 0.766 |
| Cronbach's α | 0.734 | 0.898 |

*Table 4:* Correlations between variable components and the overall score on the first dimension, for the simulation study, shown first for all six variables and then for the four variables without the "noise" variables *C* and *E*. The average squared correlation and Cronbach's α reliability coefficient is shown in each case.

| | SUN | HUM | PRE | ALT | MAX |
|---|---|---|---|---|---|
| Adana | 7.55 | 66 | 647.1 | 27 | 45.6 |
| Afyon | 7.09 | 64 | 434.4 | 1034 | 39.8 |
| Anamur | 8.33 | 69 | 993.5 | 5 | 44.2 |
| Ankara | 7.19 | 60 | 377.7 | 891 | 40.8 |
| Antakya | 7.15 | 70 | 1124.1 | 100 | 43.9 |
| Antalya | 8.28 | 64 | 1052.3 | 54 | 45.0 |
| Aydın | 7.42 | 63 | 857.7 | 57 | 44.6 |
| Balıkesir | 6.56 | 70 | 5885.0 | 147 | 43.7 |
| Bolu | 5.49 | 73 | 536.4 | 742 | 39.4 |
| Bursa | 6.35 | 69 | 696.3 | 100 | 43.8 |
| Çanakkale | 7.31 | 73 | 615.4 | 6 | 38.8 |
| Diyarbakır | 8.00 | 54 | 491.4 | 677 | 46.2 |
| Edime | 6.24 | 70 | 585.9 | 51 | 42.2 |
| Erzincan | 6.57 | 60 | 366.8 | 1218 | 40.6 |
| Erzurum | 7.05 | 64 | 447.0 | 1758 | 35.6 |
| Eskişehir | 6.46 | 68 | 373.9 | 801 | 40.6 |
| Gaziantep | 8.00 | 60 | 548.8 | 855 | 44.0 |
| Göztepe | 6.23 | 75 | 677.2 | 33 | 40.5 |
| Isparta | 7.29 | 61 | 581.0 | 997 | 38.0 |
| İslâhiye | 7.46 | 60 | 842.0 | 518 | 45.4 |
| İzmir | 8.06 | 62 | 691.1 | 29 | 43.0 |
| Karaköse | 6.24 | 68 | 533.3 | 1631 | 39.9 |
| Kars | 6.27 | 70 | 501.2 | 1775 | 35.4 |
| Kastamonu | 6.12 | 70 | 461.6 | 800 | 42.2 |
| Kayseri | 7.11 | 65 | 374.6 | 1093 | 40.7 |
| Kırşehir | 7.17 | 63 | 377.8 | 1007 | 40.2 |
| Konya | 7.29 | 60 | 325.9 | 1031 | 40.6 |
| Kütahya | 6.02 | 67 | 564.7 | 969 | 38.8 |
| Malatya | 7.40 | 54 | 387.5 | 948 | 42.2 |
| Merzifon | 6.35 | 67 | 392.4 | 755 | 42.6 |
| Muğla | 7.48 | 62 | 1196.3 | 646 | 41.6 |
| Rize | 4.14 | 77 | 2300.4 | 9 | 38.2 |
| Samsun | 4.46 | 75 | 650.3 | 4 | 38.4 |
| Siirt | 7.43 | 51 | 726.5 | 896 | 46.0 |
| Sivas | 6.43 | 64 | 417.0 | 1285 | 40.0 |
| Tekirdağ | 5.40 | 76 | 575.4 | 549 | 46.8 |
| Trabzon | 4.36 | 72 | 833.8 | 3 | 38.4 |
| Şanlıurfa | 8.28 | 49 | 463.1 | 30 | 38.2 |
| Van | 7.43 | 59 | 380.4 | 1661 | 37.5 |
| Zonguldak | 5.54 | 72 | 1220.2 | 137 | 40.5 |

*Table 5:* Averages of five meteorological variables observed in 40 cities of Turkey during 2004: *SUN* – daily hours of sunshine; *HUM* – humidity (%); *PRE* – annual precipitation (mm); *ALT* – altitude (m); *MAX* – maximum temperature (°C).