Fuzzy coding in constrained ordinations

Michael Greenacre

Department of Economics and Business
Universitat Pompeu Fabra
08005 Barcelona
Spain
E-mail: michael.greenacre@upf.edu

Faculty of Biological Sciences, Fisheries & Economics
University of Tromsø
N-9037 Tromsø
Norway

Abstract: Canonical correspondence analysis and redundancy analysis are two methods of constrained ordination regularly used in the analysis of ecological data when several response variables (for example, species abundances) are related linearly to several explanatory variables (for example, environmental variables, spatial positions of samples). In this report I demonstrate the advantages of the fuzzy coding of explanatory variables: first, nonlinear relationships can be diagnosed; second, more variance in the responses can be explained; and third, in the presence of categorical explanatory variables (for example, years, regions) the interpretation of the resulting triplot ordination is unified because all explanatory variables are measured at a categorical level.

Keywords: Canonical correspondence analysis, crisp coding, dummy variables, fuzzy coding, redundancy analysis.

Introduction

Constrained ordination of species abundances (or biomasses) in a set of samples, in the presence of environmental covariates, is routinely performed in ecological studies using canonical correspondence analysis (CCA) and redundancy analysis (RDA) (for example, ter Braak and Verdonschot, 1995, McCune, 1997, Wagner, 2004, see also Greenacre, 2010: chaps 12 and 15). Environmental variables are generally continuous in nature, for example latitude and longitude of the sampling points, temperature, depth, concentration of a pollutant, etc., but can also be categorical, for example sediment type, season, region, etc. Continuous variables are usually included linearly or in a transformed form, for example logarithmically transformed. Latitude and longitude are often included along with their squares and even cubic terms, to explain more flexibly the spatial component of variance in the biological data (Borcard, Legendre and Drapeau, 1992). Makarenkov and Legendre (2002) made a general proposal for adding polynomial terms of the explanatory variables in CCA and RDA to capture nonlinear effects. In this report we demonstrate an alternative approach of coding continuous variables as fuzzy categorical variables. This approach has several benefits: it leads to (i) a natural accounting for non-linear relationships between the biological and environmental variables, (ii) improved explained variance and (iii) a unified interpretation of the triplots in constrained ordinations.

Fuzzy coding

A continuous variable such as temperature can be recoded into k categories by cutting up the range of the variable into k intervals, using k–1 cutpoints, and then assigning the values of the variable to one of the categories. For example, a temperature range of –4°C to 5°C can be cut into $k = 3$ intervals, using cutpoints –1°C and 2°C, and an observed value of 2.5°C would fall into the third category. Such a categorical variable then generates three dummy (zero/one) variables

and this observed value in the third category would be coded as [0 0 1]. This type of coding is called crisp coding because it assigns the value totally to one category. Clearly, with this type of coding, a value of 3.3°C, for example, is also coded as [0 0 1] and is thus indistinguishable from the value of 2.5°C, leading to a substantial loss of information as a result of the recoding.

By contrast, fuzzy coding converts the original value into k "pseudo-categorical" values that represent the value of the continuous variable uniquely and exactly. Rather than cutpoints we use k membership functions, for example the triangular membership functions depicted in Figure 1. To define these functions we need k "hinge points" – for example, in Figure 1 the example of $k = 3$ is illustrated and the hinge points are the minimum value, the median (taken as 1°C) and the maximum value. As an example of fuzzy coding a given temperature value of 2.5°C, which is above the median, this value corresponds to 0 on the first membership function coding the "low" category, 0.625 on the second "middle" category and 0.375 on the third "high" category, giving a fuzzy coding of [0 0.625 0.375]. The value 3.3°C, slightly higher than 2.5°C, has a coding of [0 0.425 0.575]. The three values add up to 1, like the crisp coding, but – unlike the crisp coding – can be reverse transformed to recover the original values, as weighted averages of the hinge points:

$$(0 \times -4) + (0.625 \times 1) + (0.375 \times 5) = 2.5 \qquad (0 \times -4) + (0.425 \times 1) + (0.575 \times 5) = 3.3$$

In the following we will also fuzzy code the spatial position of samples, using a fuzzy longitude and a fuzzy latitude variable. For example, suppose the area under consideration lies between longitudes 20′ and 50′, and between latitudes 70′ and 75′ (Figure 2). Using the extreme values and their midpoints, longitude 35′ and latitude 72′30″ respectively, as hinge points, we can fuzzy code the longitude and latitude of a sample, say 41′00″ (=41.0) and 71′48″ (=71.8), as [0 0.6 0.4] and [0.28 0.72 0] respectively, as shown in Figure 2. Reversing the values for latitude

because latitude varies from bottom up, and then computing the outer product of these two vectors (the 3×3 matrix of all pairwise products), we obtain values for the 8 points of the compass and a central category:

$$\begin{bmatrix} 0 \\ 0.72 \\ 0.28 \end{bmatrix} \begin{bmatrix} 0 & 0.6 & 0.4 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0.432 & 0.288 \\ 0 & 0.168 & 0.112 \end{bmatrix}$$

Thus nine fuzzy values are equivalent to the position [41′00″, 71′48″], and the pattern of the values indicates that the sample is south-east of the centre, closer to the centre than to the eastern, south-eastern and southern points – see the position of the sample in Figure 2, marked by a cross. This position can be recovered exactly by taking a weighted average of the nine points shown in Figure 2 that combine the three hinges for longitude and the three for latitude (only four of them are nonzero):

$$0.168 \times \begin{bmatrix} 35 \\ 70 \end{bmatrix} + 0.112 \times \begin{bmatrix} 50 \\ 70 \end{bmatrix} + 0.432 \times \begin{bmatrix} 35 \\ 72.5 \end{bmatrix} + 0.288 \times \begin{bmatrix} 50 \\ 72.5 \end{bmatrix} = \begin{bmatrix} 41 \\ 71.8 \end{bmatrix}$$

In the following sections we shall implement the above system of coding first in an artificial example to demonstrate the ability of the fuzzy coding to capture nonlinear effects, and second applied to a real data set.

Artificial data set with known gradients

An artificial data set is used first to illustrate the idea, where the gradients are known. Two uncorrelated gradients were created, denoted by X (ranging from 0 to 10) and Y (ranging from 51 to 100), for a sample size of 300. Then five "species abundances", denoted by A to E, were generated from these two gradients with the following characteristics: A and B both have quadratic relationships with gradient X, with a minimum at X=6, and no relationship with Y; B

has less variance than A; C has the opposite relationship, a quadratic relationship with a maximum at X=6, and a variance approximately the same as B; D has a positive linear relationship with both X and Y; E has a weaker positive relationship with X and a weaker negative relationship with Y. The data of all five "species" were perturbed by random noise to partially disguise these relationships. The inertias of the five "species" in the CA, which is a measure of its variance in the analysis, are in the following ordering: A > B > E > C > D – thus A and B are in this sense the most important variables of the data set, and happen to also have a nonlinear relationship with X.

Figure 3 shows the CCA of this data set, where the inertia in the constrained space is 25.7% of the total for the species. The constrained space is two-dimensional, and thus this solution explains 100% of the constrained inertia. The two gradients are plotted using their (weighted) correlation coefficients with the two dimensions of the solution to define triplot vectors. The interpretation of the Y vector is in accordance with the way the data were generated, since only D and E were linearly related to Y, D positively and E negatively. The interpretation in Figure 3 of the X vector would be that A and B are negatively related to X, while C is positively related. In particular, notice that only linear relationships can be inferred in such a triplot, so there is no possibility to diagnose the nonlinear relationships of species A, B and C with gradient X.

Now the variables X and Y were fuzzy coded into five categories each, and the fuzzy categories included as constraining variables in a new CCA. Figure 4 shows a different configuration of the species (approximately a 135 degree rotation of the species configuration of Figure 3), with the categories of the gradients joined together by lines from low (1) to high (5). The sequence of categories for Y shows an approximate straight pattern, oriented similarly to the Y vector with respect to the five species as in Figure 3, with the lower categories Y1 and Y2 following

increasing abundance of species E and the higher categories Y4 and Y5 bending towards increased abundance of species D, again in accordance with the way Y was constructed as a linear gradient affecting D and E in opposite senses. The sequence of X categories, however, now forms a curved pattern, with the lowest and highest categories on the right hand side in the direction of species A and B, and the middle categories on the left hand side. This reflects exactly the three quadratic relationships of A, B and C with this variable, with C having the reverse relationship compared to A and B. Because the lower categories of X (X1 and X2) are on the upper side of the vertical axis and the higher categories (X4 and X5) on the lower side, this implies that E has a negative relationship with X, while D has a positive one, again exactly how the data were constructed. Clearly Figure 4 is more informative than Figure 3 about the true structure of the data. The variance of the species in the constrained space is now 73.4%, of which 98.0% is actually contained in the two-dimensional solution of Figure 4. Thus 72.0% of the species inertia is explained in the two-dimensional solution of Figure 4, compared to 25.7% in the two-dimensional solution of Figure 3, giving a huge gain in explained inertia as well as in interpretability of the solution.

Application to constrained ordination of real data

The real data set considered here consists of the abundances of 30 fish species at 89 sampling stations from the shrimp survey in the Barents Sea in the period April-May 1997, each based on a 20-minute bottom trawl. The spatial position, latitude and longitude, as well as depth and temperature of each station are used as environmental covariates. The spatial position is coded into nine fuzzy categories as described previously, and depth and temperature are coded into five categories each.

The canonical correspondence analysis (CCA) of these data is shown in Figure 5, as well as the geographical positions of the stations. The contribution biplot scaling (Greenacre, 2012) is shown in the map of stations and species, with species having coordinates related to their contributions to the respective dimensions. The solution is dominated by five species: *Se_me* (*Sebastes mentella*, deepwater redfish) *Bo_sa* (*Boreogadus saida*, polar cod), *Mi_po* (*Micromesistius*, blue whiting), *Me_ae* (*Melanogrammus*, haddock) and *Tr_es* (*Triosopterus esmarkii*, Norway pout). The fuzzy categories are displayed at the weighted averages of the stations, given as a separate display, slightly enlarged for legibility. Temperature shows a pattern of low values in the north and north-east, associated with abundance of polar cod, middle values in the west, associated with deepwater redfish, and higher values in the south and especially south-west, associated with blue whiting, haddock and Norway pout. Obviously, showing temperature as a single vector in the display would not be able to show this pattern. Depth follows a more "straight" trajectory in the solution, from lower depths in the south and south-west to higher depths in the north-west and west, with a tendency to medium depths in the north.

The total inertia of the data set is 2.781, of which 1.783 (64.1%) is explained by the environmental variables. By contrast, if the four environmental variables are included in their original continuous form, only 1.085 (39.0%) of the inertia is explained. Of course, the four variables imply only four parameters in the latter case (i.e., four dimensions in the constrained space), whereas the fuzzy categories imply more free parameters: four for the spatial variables, and four each for depth and temperature (i.e., 12 dimensions in the constrained space). But a comparison of the two alternatives with the same number of free parameters still shows that the fuzzy coding has a benefit in terms of inertia explained, at least in this application. For example, using only the spatial information to constrain the solution, if the positions are coded by latitude and longitude as well as their squared terms, as proposed by Borcart, Legendre and Drapeau

(1992), this implies four parameters and the inertia explained is 40.1%, whereas it is 50.5% for the fuzzy coded spatial variables, which also absorb four free parameters.

Discussion

Fuzzy coding of continuous variables was first introduced into the ordination context in the French literature, originally in a doctoral thesis by Bordet (1973) and subsequently used by Ghermani, Roux and Roux (1977) and Guitonneau and Roux (1977) to facilitate the joint analysis of continuous and discrete variables. Various other applications have appeared, for example Loslever and Bouilland (1999) and Loslever and Lepoutre (2004). Aşan and Greenacre (2010) demonstrated the ability of fuzzy coding to capture nonlinear relationships amongst continuous variables in biplots. They also showed how estimates in correspondence analysis of the fuzzy-coded categories of continuous variables can be back-transformed (i.e., defuzzified) to estimates of the original variables to obtain explained variances for each dimension of the solution.

In this report the benefit of fuzzy coding of continuous environmental variables in constrained ordinations has been demonstrated. Categorical variables have more flexibility to explain the relationships of the environmental variables with the pattern of species abundances in the ordination, as shown by the application in Figure 5. Fuzzy coding transforms continuous variables into fuzzy categories with no loss of information, since a fuzzy-coded variable can be back-transformed to its original value (Aşan and Greenacre, 2010). This is an improvement over the strategy of coding a categorical variable crisply as a set of dummy variables according to a slicing up of the variable into intervals, where the information about the value of the variable within each interval is lost. Another advantage is that the interpretation of the constraining environmental variables in the ordination is unified over continuous and categorical variables. An environmental variable that is truly categorical, for example sediment type, would be coded

crisply as a set of dummy variables and would be displayed in exactly the same way as a fuzzy variable – in Figure 5, for example, a crisp category would be at the average of the stations corresponding to it, just like the fuzzy categories are at the weighted average of their corresponding stations. Thus there is only one rule of interpretation, instead of different ones for continuous and categorical variables.

Significance testing in this framework, by permutation tests, can be conducted in the usual way (see, for example, Oksanen 2011). To test an explanatory variable with several fuzzy categories, all categories should be included in the test of the relationship of the ordination solution with the variable. A comparison of models can also be made, for example to test the linear model against one with more than two fuzzy categories – notice that the linear model is equivalent to coding the variable using two triangular membership functions, i.e., two fuzzy categories. The number of fuzzy categories to use depends on the nature of the relationship being investigated: three fuzzy categories, which imply two free parameters, would allow the possibility of one turning point in the relationship, four fuzzy categories would permit two possible turning points, and so on.

Software and supplementary material

All computations were made using the **ca** package (Nenadić and Greenacre, 2007) and own scripts in the R language (R Development Core Team, 2011). An R function `fuzzy.tri` for fuzzy coding into any number of categories using triangular membership functions is provided as supplementary material, as well as the artificial and real data sets used as examples.

Acknowledgments

This research has been supported by the BBVA Foundation, Madrid, Spain. Partial support of Spanish Ministry of Science and Innovation grants MTM2008-00642 and MTM2009-09063 are also acknowledged. Data in this article are a subset of the fish species abundance data collected during the former annual shrimp surveys by the Norwegian Institute of Fisheries and Aquaculture (NIFA) and the Institute of Marine Research (IMR) in the Barents Sea (Fossheim, Nilssen and Aschan, 2006). This subset of data, called "BarentsFish", has been used in the annual multivariate statistics course at the Faculty of Biosciences, Fisheries and Economics at the University of Tromsø, given by Michael Greenacre and Raul Primicerio.

References

Aşan, Z. and Greenacre, M. 2011. Biplots of fuzzy coded data. Fuzzy Sets and Systems, 183: 57–71.

Borcard, D., Legendre, P. and Drapeau, P. 1992. Partialling out the spatial component of ecological variation. Ecology 73: 1045–1055.

Bordet, C. 1973. Études de Données Géophysiques, Doctoral thesis (3ème cycle), Université de Paris VI, France.

Fossheim, M., Nilssen, E.M. and Aschan, M. 2006. Fish assemblages in the Barents Sea. Marine Biology Research 2: 260–269.

Ghermani, B.M., Roux, C. and Roux, M. 1977. Sur le codage logique des données héterogènes, Les Cahiers de l'Analyse des Données 1 : 115–118.

Greenacre, M. 2010. Biplots in practice. BBVA Foundation, Madrid. Freely downloadable from www.multivariatestatistics.org.

Greenacre, M. 2012. Contribution biplots. Journal of Computational and Graphical Statistics, accepted for publication, in press.

Guitonneau, A. and Roux, M. 1977. Sur la taxinomie de genre Erodium. Les Cahiers de l'Analyse des Données 1 : 97–113.

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

Loslever, P. and Lepoutre, F.X. 2004. Analysis of objective and subjective data using fuzzy coding and multiple correspondence analysis: principle and example in a sitting posture study. Theoretical Issues in Ergonomics Science 5: 425–443.

Makarenkov, V. and Legendre, P. 2002. Nonlinear redundancy analysis and canonical correspondence analysis based on polynomial regression. Ecology 83: 1146–1161.

McCune, B. 1997. The influence of noisy environment data on canonical correspondence analysis. Ecology 78: 2617–2623.

Nenadić, O. and Greenacre, M. 2007. Correspondence analysis in R, with two- and three-dimensional graphics: the ca package, Journal of Statistical Software 20. URL http://www.jstatsoft.org/v20/i03/, last accessed June 13 2012.

Oksanen, J. 2011. Multivariate analysis of ecological communities in R: vegan tutorial. Publication of University of Oulu Computer Centre, volume 83, issue 403. URL: http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf, last accessed June 13 2012.

R Development Core Team (2011). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `http://www.R-project.org`

ter Braak, C.J.F. and Verdonschot, P.F.M. 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. Aquatic Sciences, 57: 255-289.

Wagner, H.H. 2004. Direct multi-scale ordination with canonical correspondence analysis. Ecology 85: 342–351.

Figure 1: Coding of a continuous variable as three fuzzy categories, showing two examples: the temperature 2.5°C is coded as [0 0.625 0.375] and 3.3°C as [0 0.425 0.575].

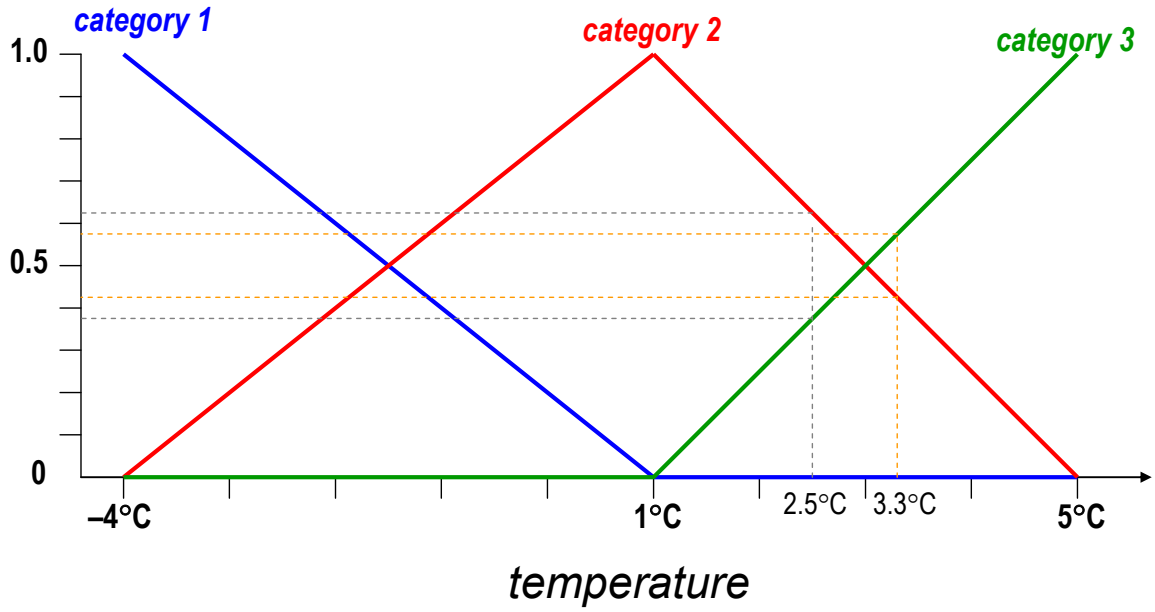

Figure 2: Fuzzy coding of a spatial variable, using a fuzzy coding of longitude and latitude. Three-category coding on each axis is illustrated, leading to 9 categories for the two-dimensional position.

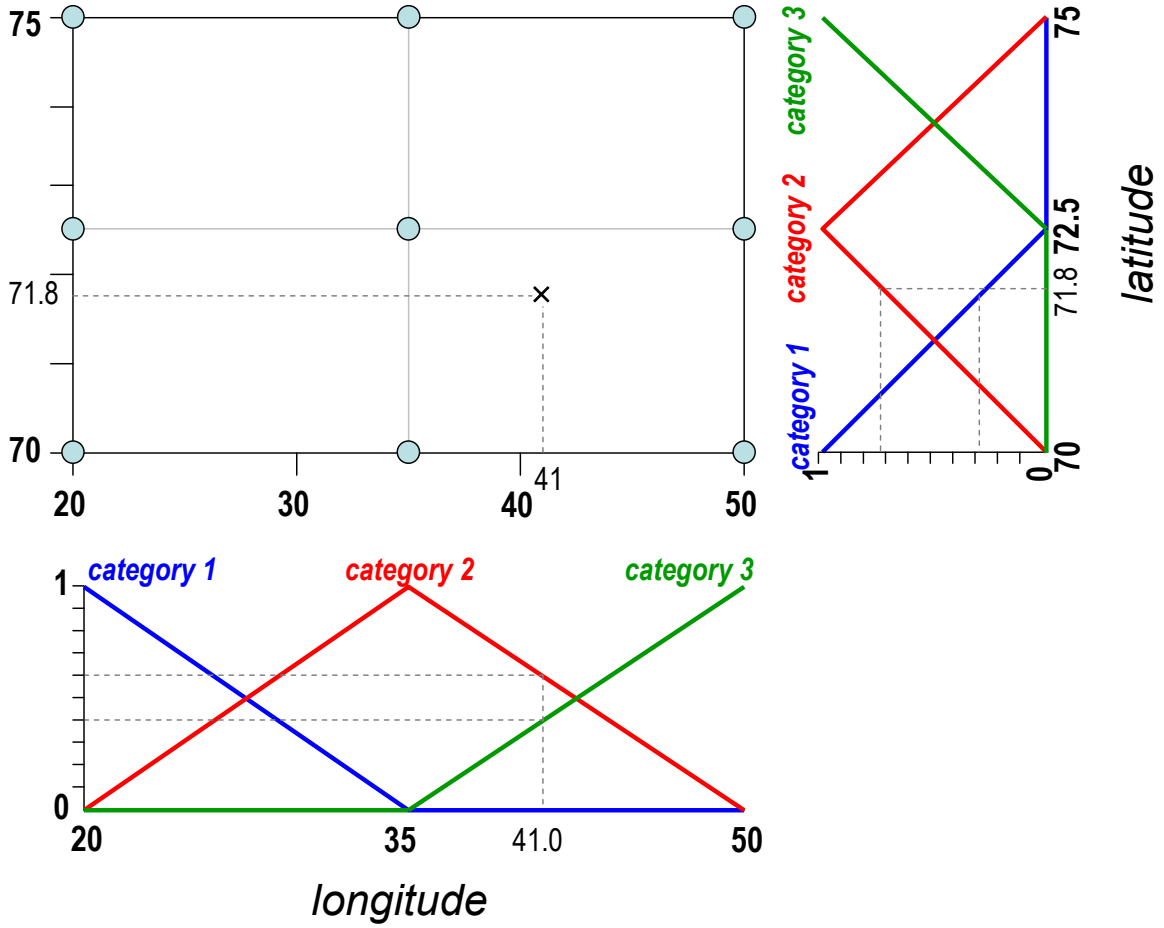

Figure 3: Canonical correspondence analysis of artificial species data set, with linear constraints defined by two environmental gradients X and Y. The gradients account for 25.7% of the inertia of the species data, all of which is contained in this solution.

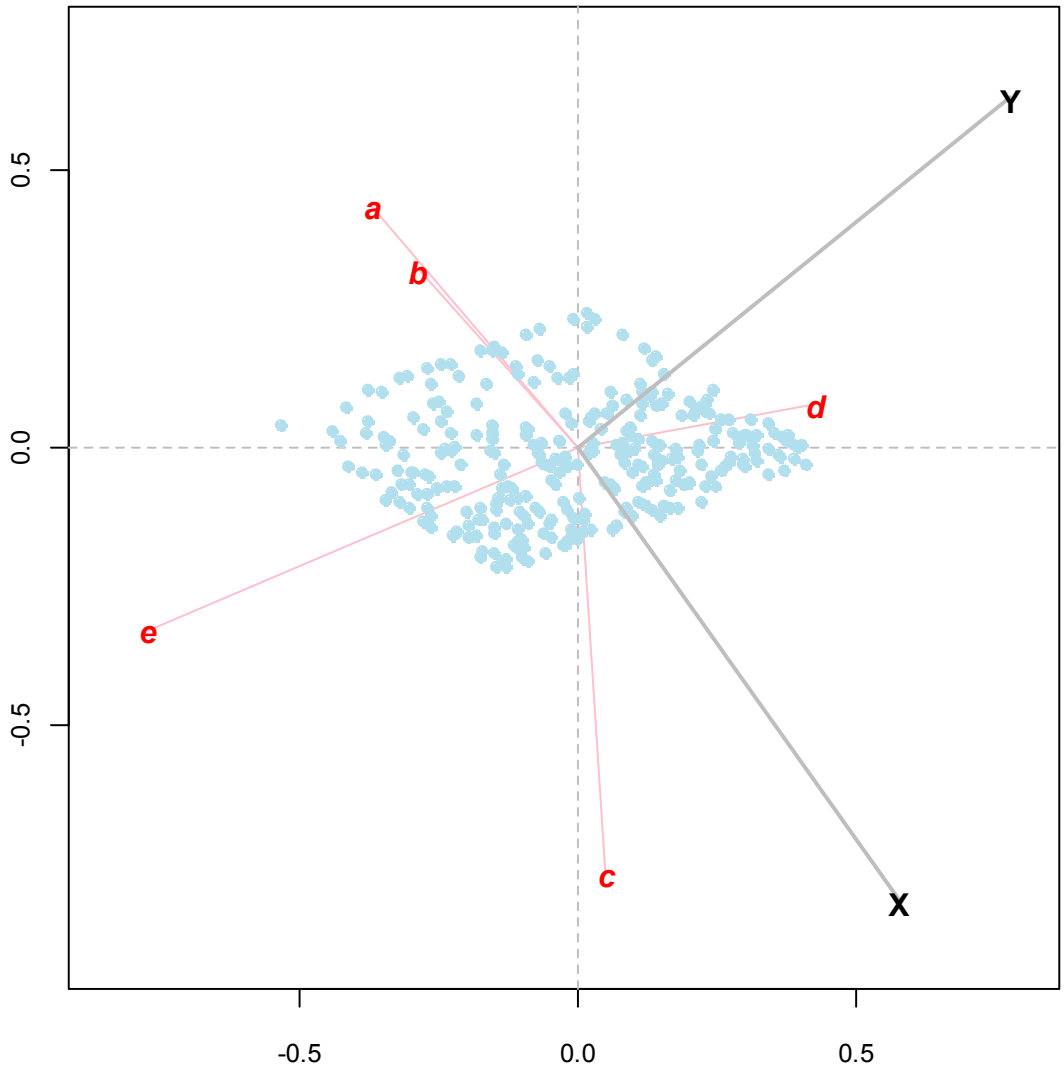

Figure 4: Canonical correspondence analysis of the same artificial species data as in Figure 3, but constrained by the gradients each coded into five fuzzy categories. The species inertia now accounted for in the constrained space is 73.4%, almost all of which is contained in this solution (72.0% of the species inertia), and the patterns of the fuzzy categories are in agreement with the way the data were constructed.

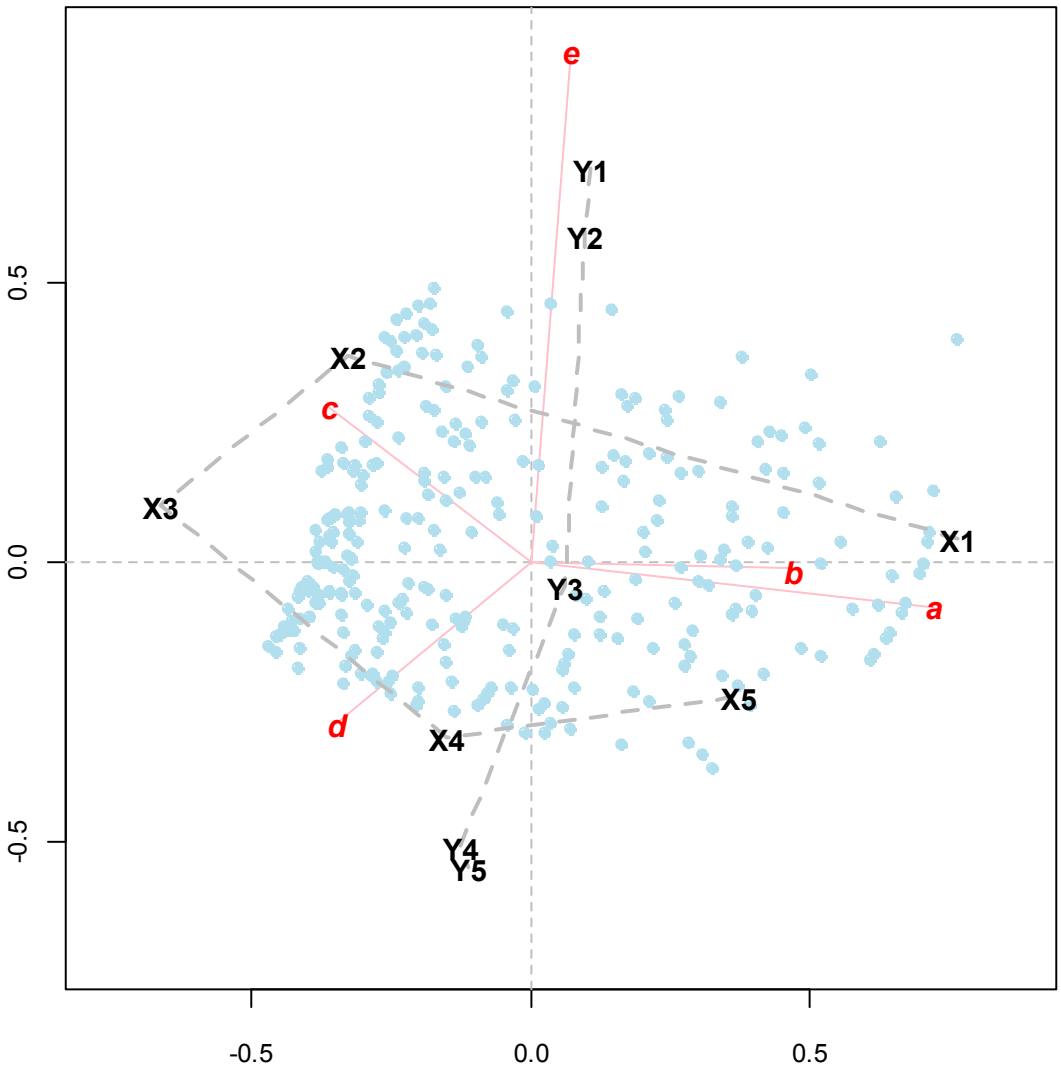

Figure 5: Canonical correspondence analysis of the fish data set, with spatial (longitude, latitude) and environmental variables (depth, temperature) coded fuzzily. The fuzzy categories, contained in the frame on the left, are shown enlarged on the right. The spatial map of the stations in the Barents Sea is shown at bottom right.