32 Canonical correspondence analysis of lowland pasture vegetation in the humid tropics of Mexico

Lira-Noriega, A., Laborde, J., Guevara, S., Sánchez-Ríos, G., Zuur, A.F., Ieno, E.N. and Smith, G.M.

32.1 Introduction

The aim of this chapter is to provide an application of canonical correspondence analysis (CCA), and we will use a lowland tropical vegetation data set. It should be noted that the aim is not to provide a detailed statistical analysis of these data, as other methods may be more appropriate to answer the underlying questions. The reason for this is that the original data set contained a large number of zero abundance for most species, and therefore statistical techniques discussed in Chapters 10, 15, 26 and 28 (non-metric multidimensional scaling and the Mantel test) are more appropriate tools to analyse the original data. However, aggregating the data (using families instead of individual species and averages per pasture instead of individual sampling plots) to reduce the number of zeros gave a data set to which CCA can be applied.

It is well known that from the middle of the twentieth century to the present, vast areas of the tropical rain forest on the American continent have been destroyed, and at an alarming rate. Entire landscapes, previously covered by luxuriant tropical rain forest, are now occupied by extensive man-made pastures where cattle graze, mainly to satisfy the demand of the ever growing urban population of developing countries. Most people, particularly ecologists, think of these man-made pastures as the antithesis of life (or biodiversity). Not only do they destroy the most complex and diverse of vegetation communities, they are extremely poor in species and have become simplified systems. Furthermore, in the Americas, this is exacerbated by the fact that they are grazed by Asian-derived cattle and are commonly dominated by African grasses. However, even though pastures are currently the most common type of vegetation in the lowland humid tropics of the Americas, ecologists seem to have persistently avoided studying them.

The lack of the most basic information about these pastures in the Americas is remarkable when we compare this with the detailed knowledge that exists for the vegetation characteristics and ecology of man-made grasslands in temperate countries, mainly Europe and the U.S.A., and of South American savannas. Another more immediate and local contrast emerges when one considers that the vast body

of knowledge and published research on the vegetation structure, composition and dynamics of tropical rain forest has been carried out within biological research stations. The majority of these are in close contact with or surrounded by man-made pastures; however specialised literature, more often than not, regards these pastures as a 'non-habitat'. Even though the replacement of tropical rain forest by pastures is a serious modification of the natural world, we can say little about this anthropogenic system without engaging in a detailed study of its vegetation.

The current study was carried out at two localities adjacent to a tropical rain forest reserve in Mexico, where cattle are currently being raised, although the history of the establishment and management of the pastures differ between localities. Twenty active pastures were sampled and analysed in order to describe the spatial variation of vegetation characteristics and to evaluate whether differences in their management history and practices (i.e., grazing regime, herbicide use, etc.) were able to explain this spatial variation.

32.2 The study area

Vegetation sampling was done in the volcanic mountain range known as Los Tuxtlas, in the state of Veracruz, Mexico (Figure 32.1). The mountain range is 90 km long, oriented in a NW-SE direction, and 40–50 km across. It ranges from sea level up to 1680 m above sea level, emerging from the coastal plain in the southernmost part of the Gulf of Mexico. Los Tuxtlas is the most humid region along the coast of the Gulf of Mexico, with an annual precipitation of 4000 mm. Although it rains all year round, there is a three-month 'dry' season from March to May and a 'wet' season from June to February.

At the beginning of the twentieth century, Los Tuxtlas was covered by more than 300,000 hectares of tropical forest; in 1991 only 15–20% of the original area was still forested, the rest had been transformed into pastures, crops, roads and urban areas. In 1991, pastures covered 160,000 ha of previously forested areas and were mostly located in the lowlands, below 500 m. Currently, the largest tract of tropical rain forest in the lowlands of the region is found in the Los Tuxtlas tropical biology research station (640 ha) and has been managed by the National Autonomous University of Mexico (UNAM) since 1967.

The two localities studied are La Palma and Balzapote. These are adjacent to each other and to the UNAM station (Figure 32.1). La Palma was founded in the 1930s, and since that time, cattle raising has been the main economic activity of its inhabitants. Balzapote was founded 10 years later by farmers who had little or no cattle, but instead practiced 'slash and burn' agriculture, growing mainly maize, beans and squash. During the 1970s and the 1980s, socioeconomic factors discouraged subsistence agriculture, while cattle ranching was favoured. At the beginning of the 1990s, most of Balzapote was covered by pastures that were tended by reluctant ranchers who were previously successful farmers.

In the region there are two different ways of establishing a pasture. In one, cattle that have been grazing in another pasture (where grasses are seeding) are

introduced into a crop field with maize stalks. The grass seeds are deposited via the cow dung, and with the help of frequent weeding (by hand and using a machete, or by spraying selective herbicides), the growth of native grass species is favoured. In this case, grasses are not actively sown, but their establishment is induced. This type of pasture is locally known as 'grama pasture'. Alternatively, the grass is directly sown or planted in the ground, and currently the most commonly used species is African Star grass (*Cynodon plectostachyus*). This introduced and improved species does not produce viable seeds naturally in Mexico, and therefore, it is propagated vegetatively in the region; i.e., by planting 10–15 cm long segments of its stolon between crops. This type of pasture is known locally as 'star pasture'.

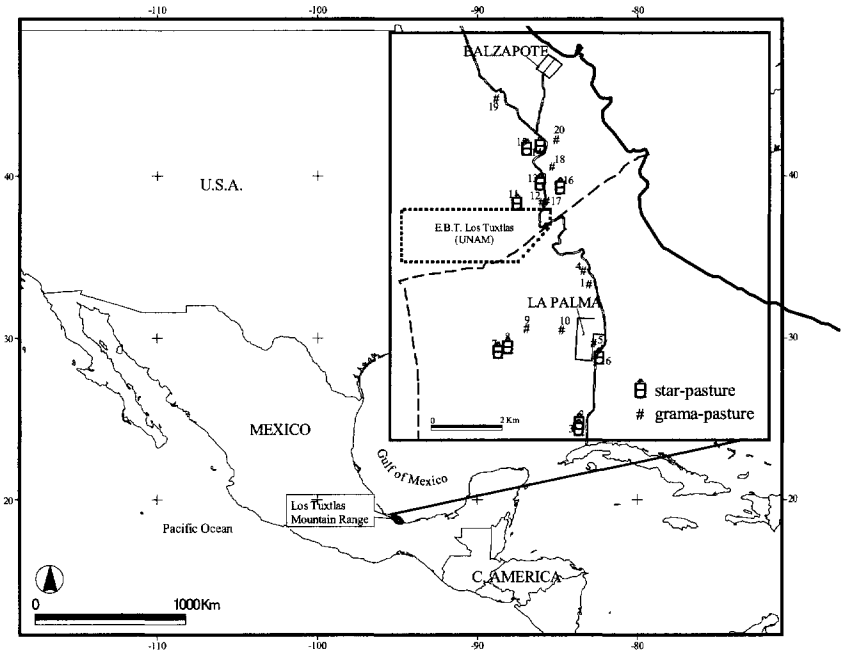

Figure 32.1. Map of study area, showing the 20 pastures sampled.

32.3 The data

Vegetation sampling in the 20 pastures was carried out during the dry season (April 10–May 12) and the rainy season (September 25–October 11) of 1992. However, as temporal changes in vegetation were only related to biomass and not to composition (Lira-Noriega 2003), we only used the dry season data set for the CCA. At each of the two localities, 10 actively grazed and well-tended pastures (hearty herbaceous cover without invading shrubs) were selected: Five were 'star

pastures', and five were 'grama pastures' (Figure 32.1). Selected pastures were those in which the owner allowed us access and agreed to be interviewed about pasture history and management details. At each pasture a visually homogeneous and undivided area (no fence subdivisions) larger than one hectare was selected, and within this, a 100 x 100m area was marked with pegs avoiding the shade of trees when possible. Within the marked hectare, ten 2 × 2 m plots were randomly selected for sampling (Figure 32.2). All plant species rooted within the plot were identified. The percent cover of each species within the plot was assigned to one of six possible categories: < 1%; 1–5%; 5–25%; 25–50%; 50–75% and >75%, represented by the numbers 1 to 6, respectively. For each plot the minimum and maximum height of the turf (foliage height) was measured to the nearest centimetre. For each pasture, the altitude above sea level and slope were recorded. In addition, several variables were obtained by interviewing the owners of each pasture. These included the dates the forest was felled, and the pasture was established, in addition to other features related to management practices (number of cows, frequency and method of weeding, etc.; see Table 32.1).

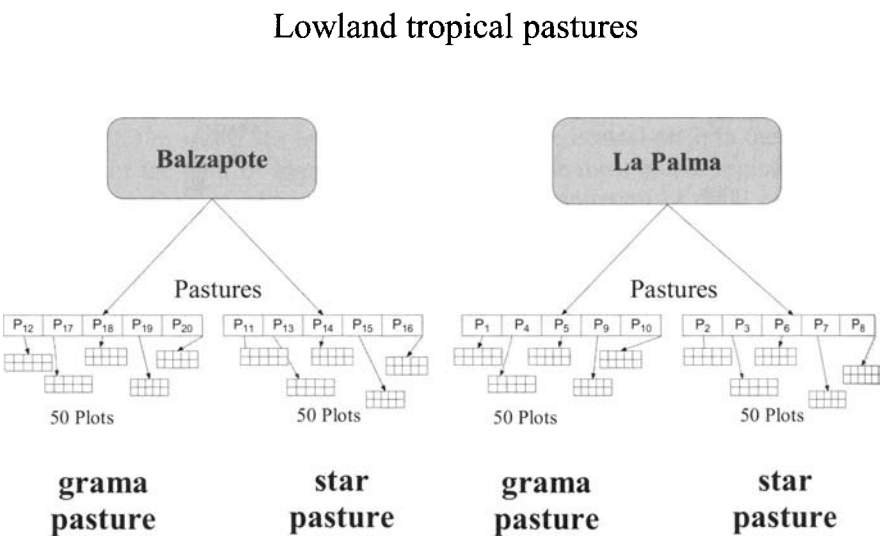

Figure 32.2. The experimental design. The two sampling localities (Balzapote and La Palma) each contain 10 pastures: 5 'grama pastures' with native grass and 5 'star pastures' with introduced grass (see Section 32.2). Each pasture has 10 plots.

The main underlying question in this study is whether there is a difference in species communities among the four groups of pastures in Figure 32.2. Each group is characterized by a different management intensity and history. To quantify these four groups, a nominal variable 'block' is introduced. It has four values

identifying the classes (i) grama pastures in Balzapote, (ii) star pastures in Balzapote, (iii) grama pastures in La Palma and (iv) star pastures in La Palma.

The original data contained 171 species, most of them are rare. To reduce the large number of zeros in the data matrix, field data per species were converted into plant cover per family for each plot. As the introduced African star grass (*C. plectostachyus*) is the only species that is directly sown by the ranchers, it is considered a separate family for this analysis. The remaining species of the grass family (Poaceae, formerly Gramineae) are pooled together as they all produce viable seed in Los Tuxtlas and are native to Mexico. We treat cover as a continuous variable because the difference between coverage of 0 and 1 is considered equally important as the differences between coverage of 5 and 6. It is in fact an ordinal variable with six classes (<1%, 1–5%, 5–25%, 25–50%, 50–75% and >75%). A list of all the families used in this analysis is available online.

Table 32.1. A summary of available explanatory variables. *Measured in the field; ** From interviews with the owner of the pasture.

Explanatory Variable	Remarks
Altitude above sea level*	Continuous variable (metres).
Field slope*	Continuous variable (degrees).
Bared soil*	Continuous variable measured in a similar way as vegetation cover (1–6).
Time since forest clearing**	Categorical variable with index values from 1–8, representing ages from approximately 6 to 40 years.
Cattle grazing intensity**	Continuous variable (head of cattle per hectare).
Weeding frequency**	Categorical variable: 1 = no weeding, 2 = weeding with a machete once per year, 3 = weeding with a machete more than once per year.
Herbicide spraying**	Categorical variable: 1 = no spraying, 2 = one spray per year, 3 = more than one spray per year.
Plague**	Nominal variable: 0 = no plague, 1 = plague in the pasture the year before sampling. This was an atypical insect herbivore attack on grass leaves in some pastures the year before sampling.
Minimum vegetation height*	Continuous variable (cm).
Maximum vegetation height*	Continuous variable (cm).

32.4 Data exploration

Observations equal to zero

The first point we look at is how many observations in the family data are equal to zero as this determines which multivariate method should be applied in the next step of the analysis. If there are lots of zeros in the data, then the correlation and covariance coefficients (used by principal component analysis and redundancy

analysis), and the Chi-square distance function (used by correspondence analysis and canonical correspondence analysis) are less suitable to define association (Chapters 12, 13, 26, 28). For such data the Jaccard, Sørensen or Bray–Curtis indices might be more appropriate, followed by non-metric multidimensional scaling.

There are different ways to get an idea of the number of zeros in a data set. We can either look at the spreadsheet, express the number of zeros as a percentage, or visualise the zeros in a figure. The last option is carried out here (Figure 32.3). Each symbol '-' means that a particular observation was equal to 0. Note that there is indeed a large number of observations equal to zero.

This indicates that it is not appropriate to apply principal component analysis (PCA) or redundancy analysis (RDA) on these data as two families who are jointly absent at sites are calculated as more similar than families who are jointly present. And correspondence analysis (CA) and canonical correspondence analysis (CCA) will be dominated by patchy families. In an initial analysis, we did apply a CCA on these data and it gave a strong arch effect (Chapters 12 and 13). Detrended CCA removes the arch effect, but the results may still be dominated by patchy species (or families in this case). It may be an option to apply a special data transformation so that Chord distances can be visualised in RDA.

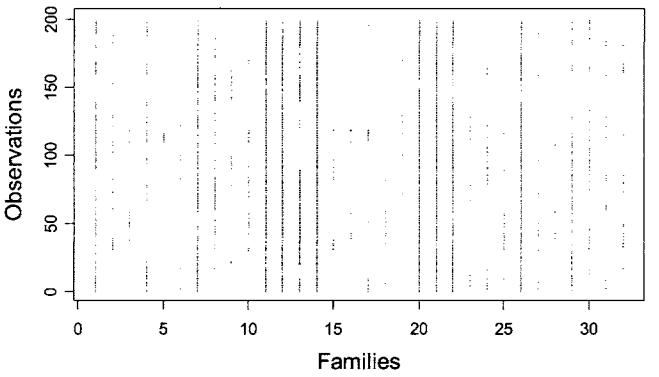

Figure 32.3. Visualisation of the number of observations equal to zero. The horizontal axis represents the different families and the vertical axis the 200 observations (sampling plots). The order of the observations corresponds to the order in the spreadsheet. A '-' indicates an observation equal to 0.

Instead of applying NMDS or RDA (combined with the Chord transformation), there is an alternative. The data contain 10 replicate observations per pasture (Figure 32.2), and we decided to use the average of these 10 replicates. The reasons for this are that (i) the main underlying question for this study is whether there is a block effect, and (ii) most explanatory variables have the same value for all replicates within a pasture.

The average index of family cover per pasture was calculated, and as a result, we have a data set with 32 families and 20 observations (average index per pasture). Most of the explanatory variables had the same value for all replicates in a pasture. For those that differed within pastures, we took the average. Hence, we have a new data set of dimension 20-by-32 with average indices, and nine explanatory variables measured at the same pastures. This new data set has considerably fewer zeros for the family data.

Outliers

Cleveland dotplots and boxplots (not shown here) showed that none of the families had extreme observations. Considering the explanatory variables, bared soil had one observation that was twice as large as the second largest observation and therefore bared soil was log transformed. Weeding frequency had 16 pastures with the same value, and only three unique values overall, so with respect to weeding frequency the data are highly unbalanced and we decided to omit this variable from the analysis. Five families (Boraginaceae, Adiantaceae, Schizaeaceae, Selaginellaceae, Sapindaceae) were measured at less than five pastures and therefore were omitted from the analysis.

Collinearity

In the next step of the data exploration, we investigate the relationship between the explanatory variables. Figure 32.4 shows a pairplot of the continuous explanatory variables, and there are some problems. Minimum and maximum vegetation height have a cross-correlation of 0.82 indicating a strong linear relationship. The scatter of points for these two variables (panel opposite "0.82") confirms the strong linear relationship. This means that we have to omit one of the vegetation height variables (it does not matter which one because both variables are basically representing the same ecological signal). We decide to drop minimum vegetation height.

The fact that the pairplot indicates that there are no other strong two-way interactions does not mean that there is no further collinearity. We have not included the two nominal variables 'block' and 'plague', and there might be three- or multiway interactions between the explanatory variables. A dotplot or boxplot conditional on 'block' (not shown here) gives the impression that there is a block effect in some of the explanatory variables. We will discuss this further when applying the multivariate analysis.

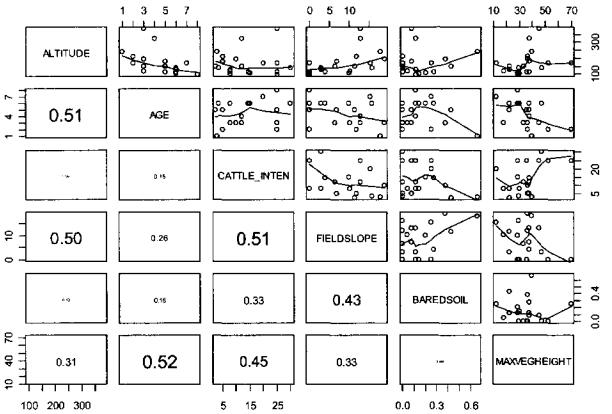

Figure 32.4. Pairplot of the continuous explanatory variables. The lower diagonal panels contain the (absolute) correlation coefficients, and the upper diagonal panels contain scatterplots of each combination. The font size of the correlation coefficient is proportional to its value.

32.5 Canonical correspondence analysis results

We know *a priori* that the gradients are relatively long. For example, altitude ranges from 100 to 400 m and it is unlikely that species-environmental relationships are linear along such a long gradient. For this reason (Chapter 13), we apply canonical correspondence analysis on the pooled family data (20 pastures).

The first question we have to address is which triplot type we want to use: a species-conditional or a site-conditional triplot (Chapters 12 and 13). The underlying question in this study is whether there is a block effect. This means that we want to compare observations with each other and that we are less interested in comparing families with each other. Hence, we should use the site-conditional biplot scaling (also called: distance scaling). In this scaling, distances between observations represent two-dimensional approximations of Chi-square distances, but angles between families cannot directly be interpreted as correlations. All we can say is that lines (families) pointing in the same direction mean that those families appear at the same pastures.

The nominal explanatory variable 'plague' is coded as 0–1 and can be used in the CCA. For 'block' this is slightly more complicated. It has four levels, and therefore we create four new dummy variables B_1, B_2, B_3 and B_4, with B_j equal to 1 if the observation was taken in block j, and 0 otherwise. To avoid 100% collinearity, one of these levels has to be omitted. We select B_4. The CCA triplot is presented in Figure 32.5. The block effect seems to dominate the triplot.

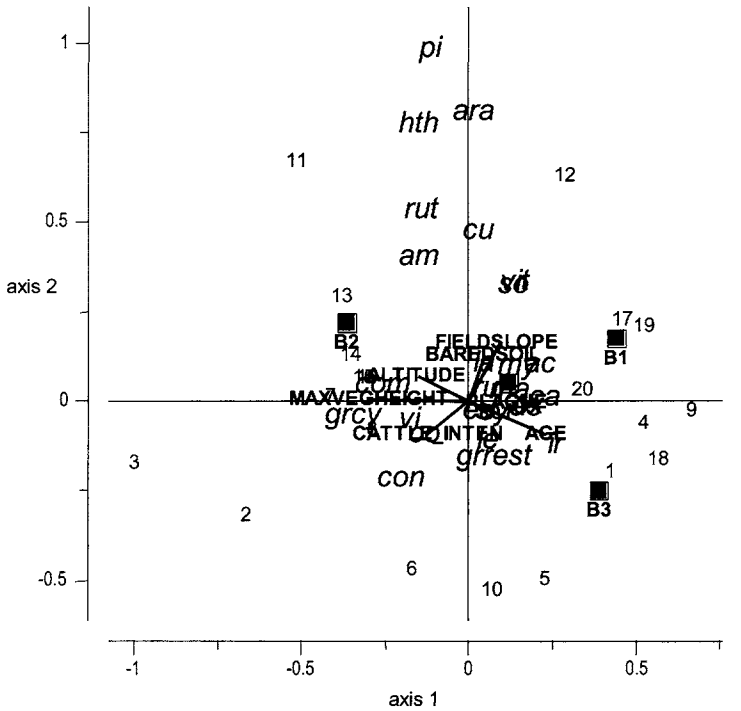

Figure 32.5. CCA triplot for pooled family data. The continuous explanatory variables are represented as lines and the nominal explanatory variables as square blocks. The numbers 1–20 refer to the pastures, and plant families are plotted as labels.

The total inertia (variation) in the family data is 0.45, and the inertia explained by all explanatory variables is 0.29. Hence, all explanatory variables explain 100 × 0.29/0.45 = 64% of the variation in the data. The first two axes explain 59% of this 64%, which means that they explain 38% of the variation in the family data. This is quite good compared with other ecological studies.

The next question is which explanatory variables are important, and this is typically investigated with a backward selection combined with a permutation test (Chapters 12 and 13). The results in Table 32.2 indicate that maximum vegetation height is highly significant, and that field slope, B_2 and B_1 are significant as well at the 5% level. The p-values for B_1 and B_2 are close to 0.05, which suggests that their role is less important as was inferred from the triplot. However, the situation is slightly more complicated. Recall that in the data exploration, we mentioned that a conditional dotplot showed a block effect in some of the explanatory variables. A good way to assess collinearity between all variables are Variance

Inflation Factors (VIFs). These were discussed in detail in Chapter 26. The VIF values for the explanatory variables are given in Table 32.3 and indicate that a considerable part of the variation in B_1 is explained by the other explanatory variables, and the same holds for B_3 and field slope.

Table 32.2 Results of the forward selection process. The number of permutations was 9999.

Explanatory Variable	F-statistic	p-value
MAXVEGHEIGHT	4.497	<0.001
FIELDSLOPE	2.214	0.013
B_2	1.841	0.047
B_1	1.951	0.037
ALTITUDE	1.637	0.091
BAREDSOIL	1.472	0.149
AGE	1.115	0.326
PLAGUE	0.862	0.531
CATTLE_INTEN	0.563	0.799
B_3	0.362	0.934

Table 32.3.VIF values for all explanatory variables. The higher a VIF value, the more variation in the explanatory variable is explained by the other explanatory variables (collinearity). VIF values larger than 5 can be considered as a problem.

Explanatory variable	VIF
B_1	6.96
B_2	3.22
B_3	7.12
ALTITUDE	4
AGE	3.59
CATTLE_INTEN	2.77
FIELDSLOPE	4.17
PLAGUE	1.78
BAREDSOIL	1.71
MAXVEGHEIGHT	3.39

There are now two ways to proceed. The first option is to remove some of the explanatory variables and identify which are collinear with the block dummy variables. The problem is that this will be a difficult and arbitrary process, and therefore, a more objective tool is needed. The second option is to apply a partial CCA and calculate the pure bock effect.

Partial CCA and variance partitioning

In a partial CCA, five different CCAs are applied on the family data. In each analysis, a different set of explanatory variables is used and the total sum of all

canonical eigenvalues of each CCA is used to calculate the pure block effect, the shared information, the (pure) effect of the other explanatory variables and the residual information. The results of these five CCA steps are given in Table 32.4, and Table 32.5 summarises the results. All explanatory variables explain 64% of the inertia (variation) in the family data, and 36% of the variation cannot be explained by these explanatory variables. Decomposing the 64% shows that the pure block effect is 13%, and the variation explained purely by the other variables is 44%. The shared explained variation is 7%, and this is due to collinearity. The percentages 13, 44 and 7 add up to 64. Summarising, the block effect explains between 13% and 20% of the variation in the family data.

Table 32.4. Results of the partial CCA. Total variation is 0.45. Percentages are obtained by dividing the explained variance by total variance. The block variables are B_1, B_2 and B_3 and 'Others' represent the remaining seven explanatory variables.

Step	Explanatory Variables	Explained Inertia	%
1	Block and others	0.29	64%
2	Block	0.15	33%
3	Others	0.23	51%
4	Block with others as covariable	0.06	13%
5	Others with Block as covariable	0.20	44%

Table 32.5. Variance decomposition table showing the effects of block and the other variables. Components A and B are equal to the explained variances in steps 5 and 4, respectively. C is equal to the variance in step 3 minus the variance in step 5, and D is calculated as Total inertia minus the explained inertia in step 1.

Component	Source	Calculation	Inertia	%
A	Pure others		0.20	44%
B	Pure Block		0.06	13%
C	Shared (3–5)	0.23–0.20	0.03	7%
D	Residual	0.45–0.29	0.16	36%
Total				100

32.6 African star grass

In the previous section we analysed whether there was any difference in the species community, and we were particularly interested in the variation in different blocks as this might represent the effect of grcyn (i.e., the cover of African star grass; *C. plectostachyus*). However, the complete family data contained this species (grcyn) as a separate family. A Cleveland dotplot (Figure 32.6) shows that this species is mainly observed in blocks 2 and 4 (see also Figure 32.2). In fact, the CCA applied in the previous section also produced diagnostics for each family (not shown here), and these show that the African star grass is fitted rather well.

Although it is difficult to verify, it might be the case that the CCA was mostly depicting only the grcyn-block relationship. One way to avoid this would be to use grcyn as an explanatory variable, but this would cause trouble with the other explanatory variables as grcyn might be influenced by altitude, field slope, vegetation height, etc. As an alternative, we apply the same analysis as in the previous section but without grcyn and it then becomes interesting to know what the pure block effect is as it may be hypothesised that it represents the grcyn effect. The triplot (not shown here) looks similar as in Figure 32.5, and the results of the variance partitioning (Table 32.6 and Table 32.7) show that the pure block effect is now 15%.

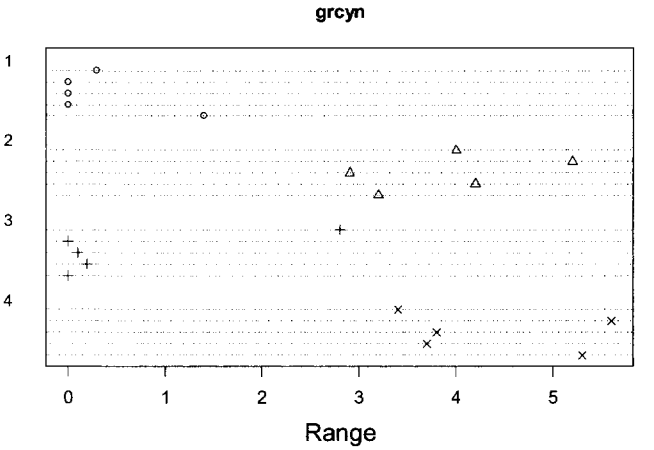

Figure 32.6. Cleveland dotplot of the African star grass *Cynodon plectostachyus* (grcyn) conditional on Block ('o' = grama pasture in Balzapote, 'Δ' = star pasture in Balzapote, '+' = Grama pasture in La Palma, "x" = star pasture in La Palma). The horizontal axis shows the value of the observations and the vertical axis the observations (within groups).

Table 32.6. Results of the partial CCA. The African grass species grcyn was not used in the analysis. Total variation is 0.39. Percentages are obtained by dividing the explained variance by total variance. The Block variables are B_1, B_2 and B_3 and the 'other' variables are the remaining seven explanatory variables.

Step	Explanatory Variables	Explained Inertia	%
1	Block and others	0.23	59%
2	Block	0.08	21%
3	Others	0.17	44%
4	Block with others as covariable	0.06	15%
5	Others with Block as covariable	0.15	38%

Table 32.7. Variance decomposition table showing the effects of Block and the other variables. The African grass species grcyn was not used in the analysis. Components A and B are equal to the explained variances in steps 5 and 4, respectively. C is equal to the variance in step 3 minus the variance in step 5, and D is calculated as Total inertia minus the explained inertia in step 1.

Component	Source	Calculation	Inertia	%
A	Pure others		0.15	38%
B	Pure Block		0.06	15%
C	Shared (3–5)	0.17–0.15	0.02	5%
D	Residual	0.39–0.23	0.16	41%
Total				100

32.7 Discussion and conclusion

In this chapter, CCA was applied to pooled data. The motivation for pooling the data was the large number of zeros in the original data. Alternative analyses would have been non-metric multidimensional scaling using the Jaccard index, or RDA (combined with the Chord transformation).

The results of variance partitioning show that the block effect explains 13–20% of the variation in the family data. The problem with this approach is that the block effect may be caused by the presence of the African star grass *C. plectostachyus* (grcyn) in the data matrix. Hence, the CCA may pick up only the grcyn-block relationships. If the African star grass grcyn is omitted from the analysis, the block effect represents 15–20% of the variation. In this case, the block effect may represent the effect of grcyn. If there are no other factors that differ between the blocks, then grcyn may be the reason for the 15–20%. However, simple dotplots and boxplots give some indication that not only grcyn differs per block as do a few of the other explanatory variables.

The main underlying question was whether there is a block effect, and the answer to this question is positive. However, understanding why there is a block effect is more difficult as the block effect not only represents differences in the presence of the introduced African star grass, but also differences between explanatory variables, as shown by the VIF values in Table 32.3. A large proportion of the variation in the block variables is explained by other variables so it is difficult to hypothesise what exactly the block effect means in terms of ecology.

As suggested by the CCA triplot (Figure 32.5), the method of pasture establishment (induced cover of native grasses or sowing the introduced African grass) has a strong effect on pasture family composition (the block effect), and this not only represents differences in the presence of the introduced African star grass, but also differences between explanatory variables, as shown by the VIF values (Table 32.3). Of all the explanatory variables considered, vegetation height and the slope of the terrain had the strongest effect on family composition. Vegetation height is highly variable and depends directly on the duration and timing of resting vs. grazing in each pasture. Slope is directly linked to grazing intensity; in flat

pastures, more cows are left to graze for longer periods than on steep terrain. Unexpectedly, other variables such as the age of the pasture and the number of cows that are kept throughout the year in each pasture, did not have an important effect on family composition. Overall, the analysis shows that a pasture's management intensity and history have a notable effect on its vegetation composition in Los Tuxtlas. Of the dozens of variables originally selected for study, the results of the CCA indicate that in order to understand the spatial variation in family composition and cover in a pasture, cattle density and grazing regime in particular require further study.

Acknowledgement

We thank Bianca Delfosse for translating parts of the manuscript into English and are grateful to all those who participated in the vegetation sampling, and to the specialists who identified the plants collected in this study. This research was funded by the Departamento de Ecología Funcional (902-17) of the Instituto de Ecología, A.C. and by the Consejo Nacional de Ciencia y Tecnología (project CONACYT 0239-N9107). We would like to thank Toby Matthews for valuable comments on an earlier draft.