31 Redundancy analysis and additive modelling applied on savanna tree data

Lykke, A.M., Sambou, B., Mbow, C., Zuur, A.F., Ieno, E.N. and Smith, G.M.

31.1 Introduction

Between 1930 and 1970, the colonial administration and the Senegalese state established 213 protected areas aimed at preserving the natural heritage and ensuring a future supply of natural resources for the local population. Today, the woody resources from the protected savannas still provide an important source of firewood, construction materials, food, animal fodder and medicine for the local people. The management of these protected areas has until recently been centralised and directed by the authorities without reference to the views of the local societies. This has often led to a lack of concern and understanding by the local people, and the protected savannas have continued to decline through uncontrolled fires, grazing animals, agriculture and logging. The decline in tree density within the savannas has been drastic during the last decades, and the remaining areas of savanna are under increasing pressure as the demands on their resources continue to grow.

Today a general agreement is emerging that locally based management of protected areas is the way forward, particularly in the light of an increased concern about the state of the natural resources among local people. For integrated sustainable use and conservation of habitats and biodiversity, it is necessary to combine local needs with a scientifically based comprehensive understanding of the biodiversity and ecology of savannas.

The comprehension of ecological processes is complicated because the changes are gradual, dynamic and related to the land use patterns, including fire and grazing by domestic animals. Furthermore, major vegetation changes took place 20–70 years ago, before most scientific investigations started. The ecology and dynamics therefore need to be investigated on the basis of current vegetation data. During the last two decades, satellite images have become available, and several studies have used remote sensing to assess vegetation status; however, the relations between vegetation characteristics and satellite-based indices are poorly understood and no consensus on methods exists. The current study aims to identify whether species composition and vegetation characteristics can be recognized from satellite images or by other simple vegetation parameters.

31.2 Study area

Field data were collected from a typical savanna in western Senegal (Figure 31.1). The area received protection status as classified forest in 1933 under the name Patako Classified Forest; since then, the vegetation has changed from dry forest dominated to savanna dominated: a change that is perceived as negative by local people who rely on its woody resources for subsistence and revenue. Today, Patako Forest is surrounded by agricultural land and functions as an important reserve of natural resources for the local population as well as being commercially logged for firewood. The climate is seasonal tropical with 900 mm annual precipitation falling within a four-month period from July to October. Savanna fires burn the vegetation every year during the dry season.

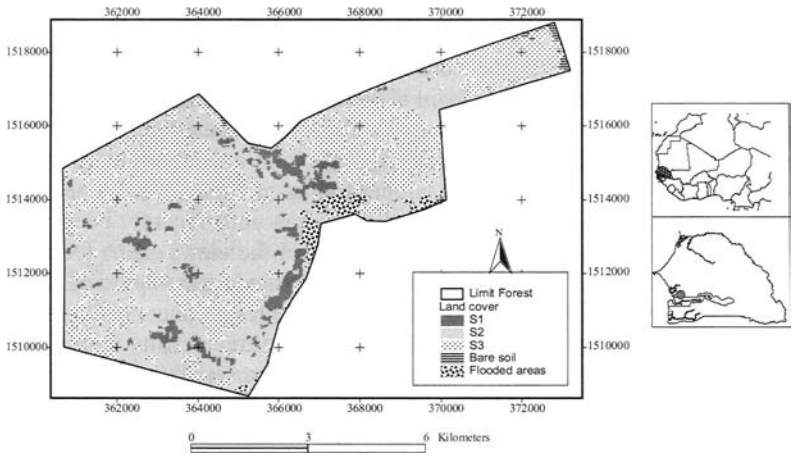

Figure 31.1. Left: Map of study area. Right: West Africa and Senegal.

31.3 Methods

The sampling was designed (i) to cover all habitat types in a vast and heterogeneous area, (ii) to be efficient in the field, and (iii) to allow for statistical analysis. The study area was divided into 250 m × 250 m quadrats and stratified into homogenous zones based on satellite images (Figure 31.2). Across the strata, 22 quadrats were selected randomly for sampling. One sample was taken in each of the selected quadrats. A sample consisted of eight, 20 m × 20 m sub-plots placed at random as follows: A pole was placed at a random point within the selected quadrat and from that point, a 115 m long line was located in the following directions N, NE, E, SE, S, SW, W and NW (Figure 31.3). The sub-plots were placed

on each line at a random distance from the pole. In total, 7.04 ha were investigated. All woody plants over 5 cm dbh (diameter at 1.3 m above the ground) within sub-plots were identified to species. Smaller individuals were counted in two groups (less than 1 m and over 1 m tall).

For the statistical analysis, the eight sub-plots in each strata were pooled in order to eliminate fine-scale heterogeneity. To eliminate rare species that were only measured at a few sites, a cut off level of five species was chosen. This reduced the number of species from 50 to 16 (Table 31.1). Thus, the final dataset contained the abundance of 16 woody plant species measured at 22 sites. Several diversity indices were calculated on the basis of the 16 selected species.

There are two types of explanatory variables, namely those derived from the satellite images (Table 31.2) and those derived on the basis of other vegetation parameters (Table 31.3). Because of extreme large cross-correlations (>0.98), some of the satellite variables were not used in the analyses (band 3, 4, 5, 7 and ndvi).

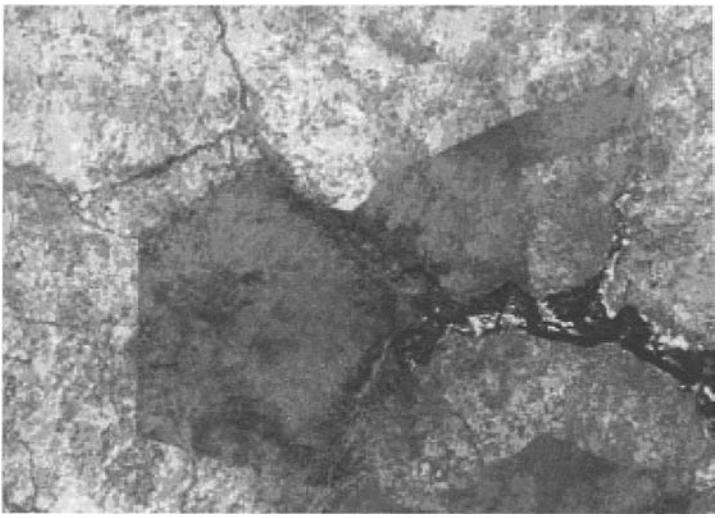

Figure 31.2. Satellite image of study area (LANDSAT — ETM data from 9 December 1999).

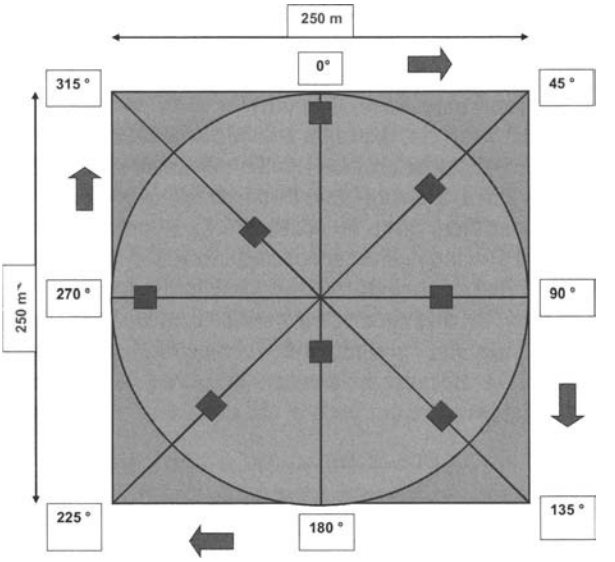

Figure 31.3. Sampling method. One sample consists of eight sub-plots.

Table 31.1. List of 16 species used as response variables in the statistical analysis.

Species	Code	Family	Total Abundance
Acacia macrostachya	Acamac	Mimosaceae	93
Bombax costatum	Bomcos	Bombacaceae	17
Combretum glutinosum	Comglu	Combretaceae	721
Combretum nigricans	Comnig	Combretaceae	208
Cordyla pinnata	Corpin	Caesalpiniaceae	64
Daniellia oliveri	Danoli	Caesalpiniaceae	45
Detarium micrantum	Detmic	Caesalpiniaceae	33
Lannea acida	Lanaci	Anacardiaceae	51
Prosopis africana	Proafr	Mimosaceae	16
Pterocarpus erinaceus	Pteeri	Fabaceae	36
Sclerocarya birrea	Sclbir	Anacardiaceae	8
Securidaca longepedunculata	Seclon	Polygalaceae	39
Sterculia setigera	Steset	Sterculiaceae	11
Struchnos spinosa	Strspi	Loganiaceae	9
Terminalia macroptera	Termac	Combretaceae	142
Xeroderris stuhlmanii	Xerstu	Fabaceae	18

Table 31.2. Explanatory variables based on satellite images (Landsat ETM).

Explanatory Variable	Definition
Strata	
Band 1	band1 (0.45 – 0.52 μm / blue)
Band 2	band2 (0.53 – 0.61 μm / green)
Band 3	band3 (0.63 – 0.69 μm / red)
Band 4	band4 (0.78 – 0.90 μm / near infrared)
Band 5	band5 (1.55 – 1.75 μm / short wave infrared)
Band 7	band7 (2.09 – 2.35 μm / medium infrared)
Brightness	0.2909 × band1 + 0.2493 × band2 + 0.4806 × band3 + 0.5568 × band4 + 0.4438 × band5 + 0.1706 × band7
Greenness	0.2728 × band1 + 0.2174 × band2 + 0.5508 × band3 + 0.7221 × band4 + 0.0733 × band5 + 0.1648 × band7
Wetness	0.1446 × band1 + 0.1761 × band2 + 0.3322 × band3 + 0.3396 × band4 + 0.6210 × band5 + 0.4189 × band7
Ratio72	band7 / band2
Savi	((band4 / band3) / (band4 + band3 + 0.5)) × (1 + 0.5)
Ndvi	(band4 - band3) / (band4 + band3)

Table 31.3. Explanatory variables based on vegetation parameters.

Explanatory variable	Definition
cl_1	regeneration, woody plants ≤ 5 cm dbh and > 1 m tall
cl_2	regeneration, woody plants ≤ 5 cm dbh and ≤ 1 m tall
Dead trunks	No. of standing dead trunks

31.4 Results

Data exploration

A data exploration was carried out to identify extreme observations and the type of relationship between species, between explanatory variables and between species and explanatory variables. Cleveland dotplots (Chapter 4) of various species showed that a data transformation was needed, as there are various extreme observations. Figure 31.4 shows two examples of this where the largest value (on the right-hand side) is considerably larger than the majority of the other

observations. Cleveland dotplots of the satellite variables indicated that there are no variables with extreme observations, but there are a few variables that show a strata effect, the two most obvious are greenness and savi (Figure 31.5). It might be an option to subtract the mean in each strata for each variable. This would remove a high correlation between variables due to strata differences. On the other hand, differences between the strata might be important; so there are arguments for and against removing it. We decided not to remove it.

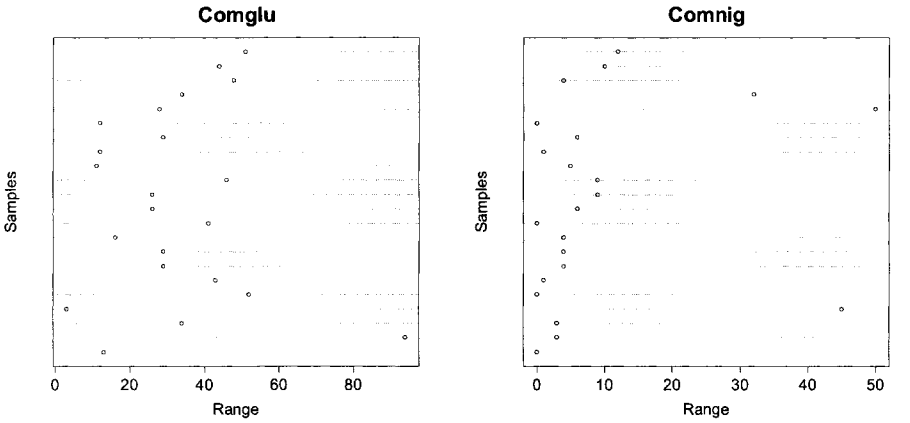

Figure 31.4. Cleveland dotplots of two species showing the need for a data transformation. The horizontal axis shows the value at a site and the vertical axis the sample number, as imported from the spreadsheet. The value at the top is the last value in the spreadsheet and the sample at the bottom the first.

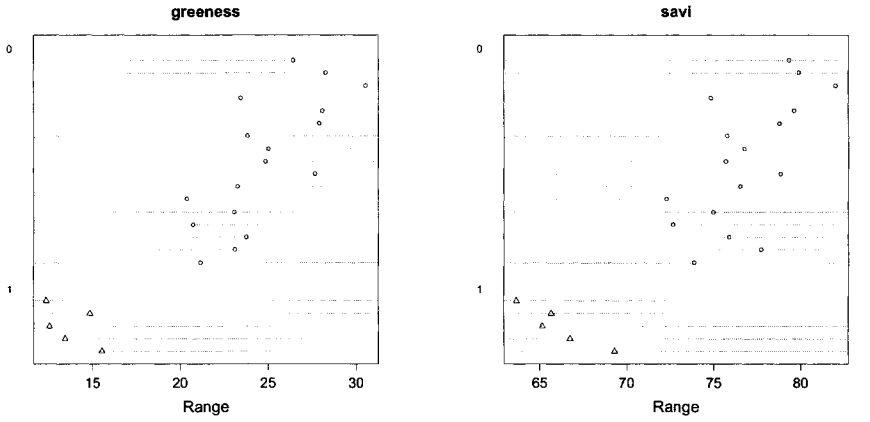

Figure 31.5. Dotplots of two explanatory variables illustrating a strata effect. The circles are observations from strata 0 and the triangles from strata 1. The horizontal axis shows the value at a site and the vertical axis the sample number, grouped by strata.

A data exploration applied on the regeneration variables (cl_1 and cl_2) and the dead trunks indicated that a transformation on these variables is required because they have a few samples with considerably larger values than the rest. We decided to apply a square root transformation on the species and a $\log(X + 1)$ transformation on cl_1, cl_2 and dead trunks. The square root transformation also improved the linear relationship between species and (untransformed) satellite variables, as indicted by scatterplots and correlation coefficients. Some cross-correlations between species and satellite variables were around 0.5. The reason we used these two transformations is based on the range of the original data; the log transformation is considerably stronger than a square root transformation.

A scatterplot between the satellite variables (Figure 31.6) indicated serious collinearity, and it was decided to omit the variables band 2, ratio72 and savi. The choice for these variables is based on the correlations in Figure 31.6, histograms of the explanatory variables (showing the coverage of the gradient) and a principal component analysis on the satellite variables (Figure 31.7). Ratio72 is negatively correlated to wetness. Greenness and savi are related. Band 2 and brightness are related. An alternative is to use VIF values (Chapter 26).

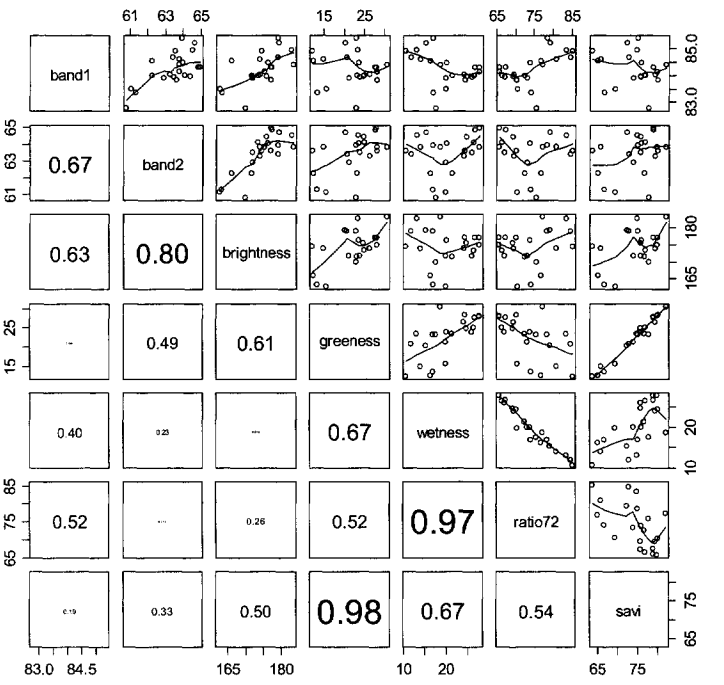

Figure 31.6. Pairplot between the satellite variables showing serious collinearity. The graphs above the diagonal are scatterplots, and numbers below the diagonal represent (absolute) correlations between the variables. Font size is proportional to the value of the correlation.

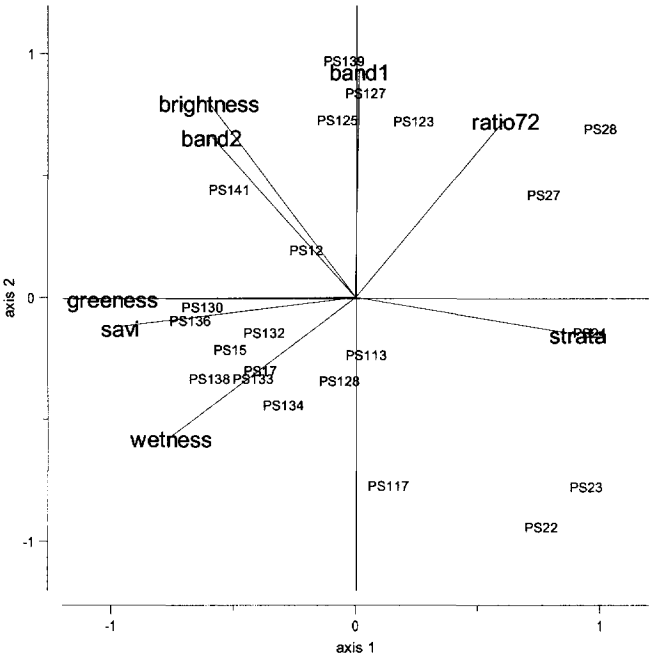

Figure 31.7. PCA applied on the satellite variables, some showing collinearity. The first two axes explain 87% of the variation in the data (53% on axis 1 and 34% on axis 2).

Univariate analysis: Species diversity versus satellite-based explanatory variables

Regression trees were used to analyse the relationship between various diversity indices (richness, Shannon–Weaver, total abundance, Simpson, Berger–Parker, Macintosh) and the selected satellite variables. However, no satisfactory model could be found. Other univariate methods like linear regression and additive modelling were also applied, but these methods did not give any convincing result neither.

Multivariate analysis: Species versus satellite-based explanatory variables

To relate the species data to the satellite variables, either redundancy analysis (RDA) or canonical correspondence analysis (CCA) could be applied. Data exploration, including coenoclines (Chapter 13), showed that most relationships between species and explanatory variables are approximately linear, which justifies the application of PCA and RDA instead of correspondence analysis (CA) and CCA, as the first methods are based on linear relationships and the second on

unimodal relationships. The additive modelling, carried out later on, also confirmed that most relationships between species and explanatory variables are approximately linear.

There are two main points to consider when applying RDA, namely (i) the covariance versus correlation and (ii) species conditional or site conditional scaling (Chapters 12 and 29). The choice between covariance and correlation requires some thought. The RDA based on the correlation matrix considers all species to be equally important. Making a small variation in a less abundant species is as important as a larger variation in a more abundant species (if the relative change is the same). The RDA, based on the covariance, focuses more on the abundant species. Here, it was decided to base the analysis on the covariance matrix in order to give the common species more weight in the analysis as this reflects the often more important ecological role of common species and better corresponds to way the vegetation is perceived by the local people. However, the effect of the more extreme values in the common species has been dampened by the square root transformation.

The scaling determines the interpretation of the triplot. The species conditional scaling gives a triplot in which angles between species and satellite variables can be interpreted in terms of correlation or covariance, but distances between samples are more difficult to interpret. In the site conditional scaling, sites can be compared with each other but angles between species do not have any formal interpretation. As we are primarily interested in the relationship between the species and the species and satellite variables, the species conditional scaling was used.

The resulting triplot is presented in Figure 31.8. Before discussing the graphical output, we discuss the numerical output. All five explanatory variables (the four satellite variables and strata) explain 35% of the variation in the species data. The two-dimensional approximation in Figure 31.8 explains 81.49% of this (53.58% on axis 1 and 17.91% on axis 2). Therefore, the first two axes explain 28.35% of the total variation in the species data.

The results of a forward selection and permutation tests, presented in Table 31.4, indicate that brightness is significantly related to the species data ($p < 0.001$). There is also a weak strata effect ($p = 0.026$). The triplot in Figure 31.8 indicates that strata is related to Comnig, and brightness is negatively related to the species Pteeri, Danoli and Comglu.

Because RDA explained only 35% of the variation, we decided to verify the results with another statistical method. A possible way of doing this is applying additive modelling (Chapter 7) in which each species is used in turn as response variable and brightness and strata as explanatory variables. Although RDA is based on linear relationships, we decided to use additive modelling as it allows for more flexibility than a parametric model. An alternative method is GAM using the Poisson distribution and log link function as the data are count data. However, this would complicate comparing the RDA (which is based on covariance) and GAM results. And as the species were square root transformed, which should stabilise the mean-variance relationship, a GAM was considered unnecessary, and the following additive model was applied on each of the 16 species:

$$\text{Species} = \text{constant} + f(\text{brightness}) + \text{strata} + \text{noise}$$

where $f(.)$ stands for a smoothing function and strata is modelled as a nominal variable. We used cross-validation (Chapter 7) to estimate the optimal amount of smoothing for brightness. Results indicated that brightness was significantly related to seven species, namely: Bomcos, Corpin, Danoli, Pteeri, Sclbir, Termac and Xerstu. Strata was significantly related to Proafr and Termac. The smoothing curves for these species are presented in Figure 31.9.

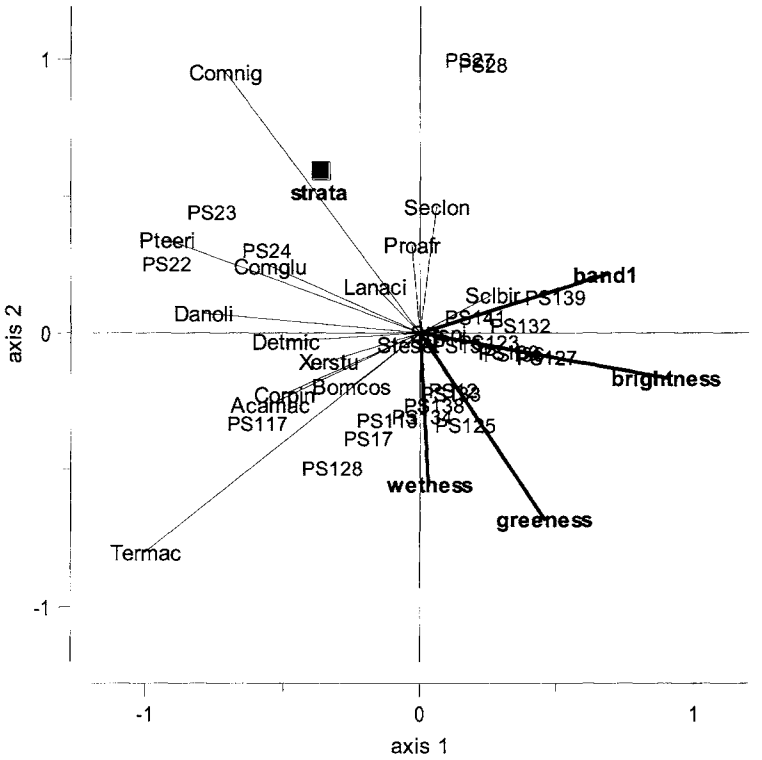

Figure 31.8. Triplot obtained by RDA.

Table 31.4. F-statistic and p-values of conditional effects obtained by a forward selection and permutation test in RDA. The number of permutations was 9999.

Variable	F-statistic	p-value
Brightness	3.685	0.000
Strata	2.248	0.026
Band1	0.828	0.570
Wetness	0.634	0.757
Greenness	1.073	0.365

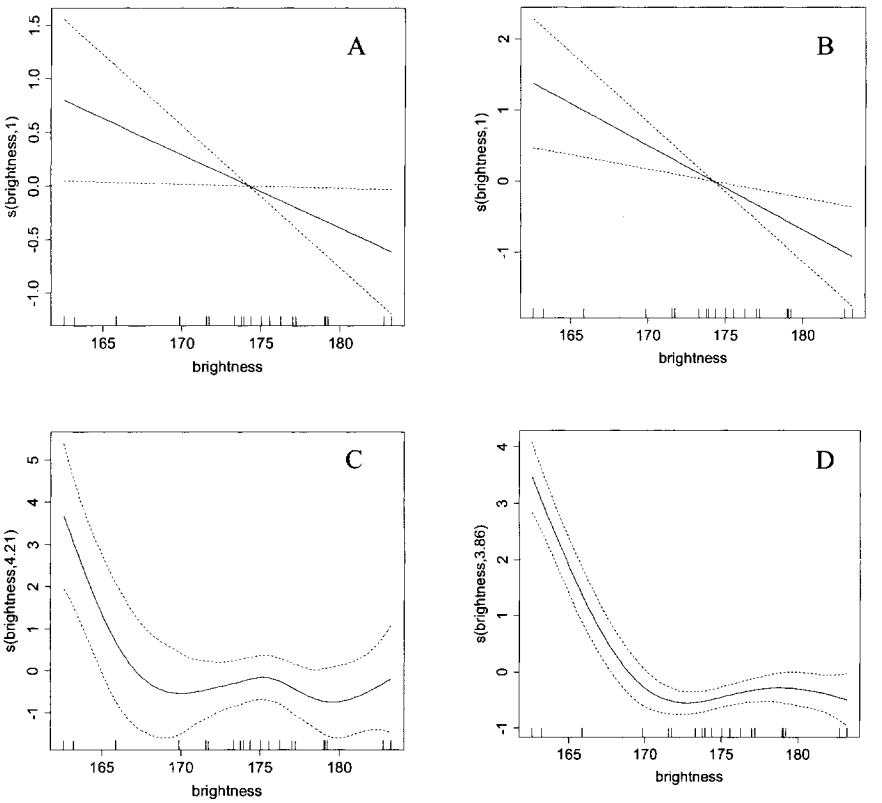

Figure 31.9. Smoothing curves for all additive models in which the smoothing curve for brightness was significant at the 5% level for the species Bomcor (A), Corpin (B), Danol (C) , Pteeri (D), Sclbir (E) and Termac (G).

Multivariate analysis: Species versus onsite explanatory variables

A similar analysis was applied on the species data using the vegetation parameters cl_1, cl_2 and dead trunks as explanatory variables. The triplot is presented in Figure 31.10. All three explanatory variables explain 24% of the variation, and the first two axes represent 89% of this (73.51% on axis 1 and 15.15% on axis 2). A forward selection and permutation test showed that cl_2 is significantly related to the species data. Other explanatory variables were not significant at the 5% level. Just as before, a more detailed analysis was applied using additive modelling in which cl_2 was used as the only explanatory variable. Results indicated that the cl_2 was significantly related to the following three species: Comnig (linear and positive). Danoli (approximately linear and positive) and Pteeri (step function going from low to high, indicating a positive relationship).

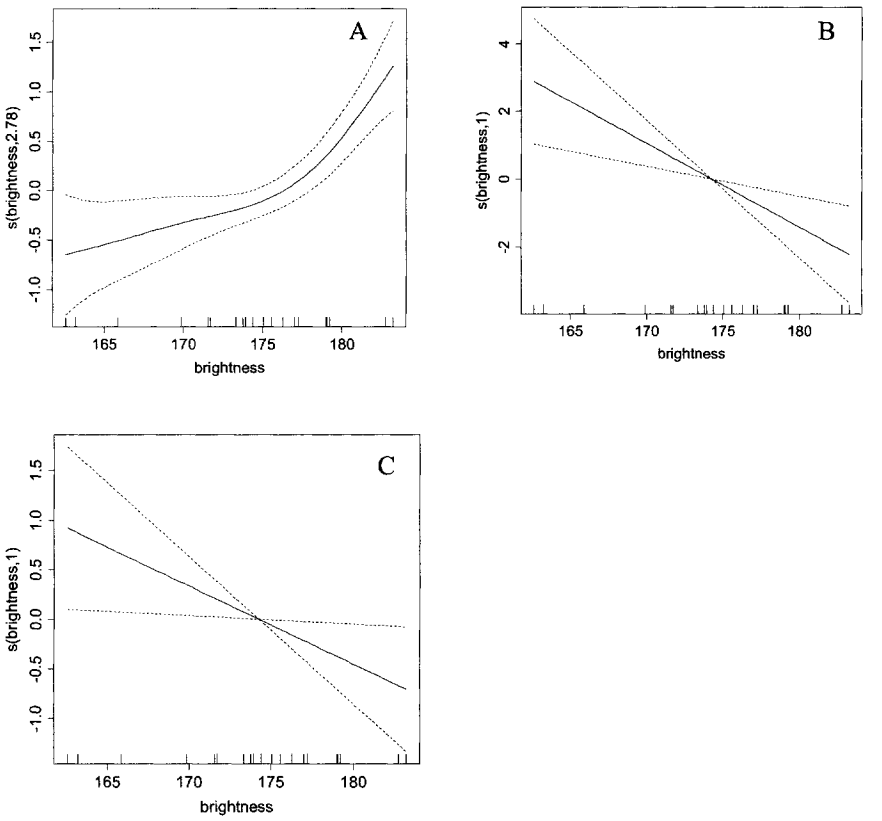

Figure 31.9 (continued). Smoothing curves for all additive models in which the smoothing curve for brightness was significant at the 5% level: Sclbir (A), Termac (B) and Xerstu (C).

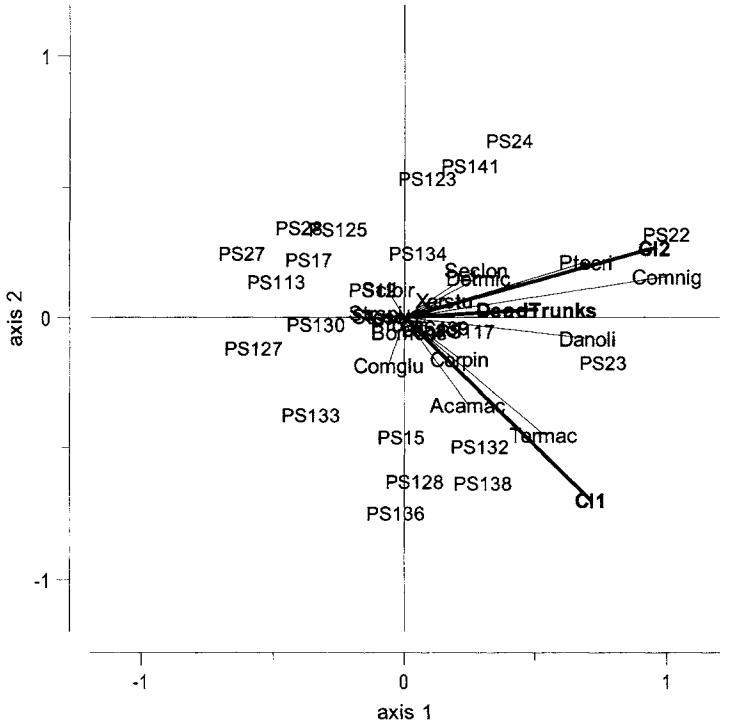

Figure 31.10. RDA triplot using *cl1*, *cl2* and dead trunks as explanatory variables.

31.5 Discussion

In this chapter, we looked at the relationship between woody plant species and satellite-derived variables. The (univariate) diversity indices are useful from an ecological point of view, as they are measures of biodiversity and abundance, but no relationship could be detected between these variables and satellite-based explanatory variables; i.e., satellite images were unable to detect patterns of diversity and density.

The RDA focussed on multiple species and satellite-based variables. We showed by aid of forward selection that two satellite-based variables, brightness and strata, were significantly related to the species data. Using these two variables in an additive model showed that seven species were significantly related to brightness and two species were significantly related to strata. Six species (Bomcos, Corpin, Danoli, Pteeri, Termac and Xerstu), all relatively large tree species, and quality species from local people's point of view, were negatively related to brightness. This indicates that brightness might be a measure of poor woody cover

and poor vegetation quality. One species (Sclbir), characteristic for drier and more open environments, was positively related to brightness. Strata was significantly related to Proafr and Termac after the effect of brightness was partialled out.

In a PCA applied on the species data using the covariance matrix, the first two axes explain 53% of the variation. Because RDA is a restricted form of PCA, it is interesting to compare the amount of explained variation in the species data obtained by both methods along the first two axes. If the explained variation by PCA and RDA along the first two axes is similar, then the selected explanatory variables explain the data rather well. On the other hand, a large difference indicates that a poor selection of explanatory variables was used. The first two axes in RDA explain 28% of the total variation in the species data. Hence, by putting restrictions on the PCA axes, we explain 25% less variation. This means that there are other explanatory variables, not used in the analysis, that are responsible for this 25%.

In a second analysis, RDA was applied on the species and regeneration information (cl_1, cl_2, dead trunks) to identify which species were abundant at sites with high (or low) regeneration. Just as before, a more detailed analysis using additive modelling was applied, indicating that cl_2 (regeneration > 1 m tall) was significantly and positively related to the species Comnig, Danoli and Pteeri. This type of regeneration indicates vegetation that has had time to restore without heavy fire impact. In areas that are heavily burned annually, the regeneration is arrested at an early stage (usually less that 1 m tall) or eliminated.

Conclusions

Scientific results play a critical role in the comprehension of vegetation changes and in identifying and evaluating the value of management solutions. The statistical investigations revealed no clear relation between satellite-based variables and typical measures of vegetation quality, such as diversity and density, which makes it risky to build general conclusions about vegetation quality in this environment on the basis of remote sensing.

More detailed statistical investigations based on RDA and additive modelling, however, indicate a relation between some of the larger tree species and some satellite-based variables such as strata and brightness. A more detailed understanding of such relations can improve the use of remote sensing in vegetation management.

The abundance of regeneration >1 m tall was also found to be an indicator of vegetation quality as it was significantly related to a number of large tree species of high quality. This could be explained by the high abundance of regeneration >1 m tall in less fire affected areas.

Acknowledgement

We thank ENRECA/Danida for financing the fieldwork.