Ecology, 83(4), 2002, pp. 1105–1117

MULTIVARIATE REGRESSION TREES: A NEW TECHNIQUE FOR MODELING SPECIES–ENVIRONMENT RELATIONSHIPS

GLENN DE'ATH[1]

Cooperative Research Center for the Great Barrier Reef World Heritage Area, James Cook University, Townsville, Queensland 4811, Australia

Abstract. Multivariate regression trees (MRT) are a new statistical technique that can be used to explore, describe, and predict relationships between multispecies data and environmental characteristics. MRT forms clusters of sites by repeated splitting of the data, with each split defined by a simple rule based on environmental values. The splits are chosen to minimize the dissimilarity of sites within clusters. The measure of species dissimilarity can be selected by the user, and hence MRT can be used to relate any aspect of species composition to environmental data. The clusters and their dependence on the environmental data are represented graphically by a tree. Each cluster also represents a species assemblage, and its environmental values define its associated habitat. MRT can be used to analyze complex ecological data that may include imbalance, missing values, nonlinear relationships between variables, and high-order interactions. They can also predict species composition at sites for which only environmental data are available. MRT is compared with redundancy analysis and canonical correspondence analysis using simulated data and a field data set.

Key words: canonical correspondence analysis; CART; classification tree; cluster analysis; cross-validation; ecological distance; gradient analysis; multivariate regression tree; ordination; prediction; redundancy analysis; regression tree.

INTRODUCTION

The importance of being able to effectively model relationships between species and environment has long been recognized, and many methods have been applied to this task (see Franklin [1995] and Guisan and Zimmermann [2000] for reviews). Recently, the focus of such modeling has shifted towards prediction, with less emphasis on description and explanation. This shift towards a predictive ecology is profound, in both philosophy and method, and is not simply a case of using well-known methods for a new purpose. Driven by issues such as global climate change, accurate prediction is now often seen as the principal objective of species–environment analyses (Franklin 1998, Iverson and Prasad 1998, Vayssières et al. 2000). Predictive accuracy is also routinely used as a method of statistical model selection (Stone 1974, Breiman et al. 1984, Ripley 1996, Burnham and Anderson 1998), and can replace the widely used practice of repeated hypothesis tests (e.g., stepwise selection); an approach that frequently leads to the inclusion of spurious explanatory variables (Draper 1995). The recognition of deficiencies in the use of statistical hypothesis tests as a method for addressing ecological hypotheses (Burnham and Anderson 1998, Johnson 1999) has also led to increased emphasis on prediction.

Manuscript received 12 January 2001; revised 14 May 2001; accepted 14 May 2001; final version received 11 July 2001.

[1] Present address: Australian Institute of Marine Science, PMB No. 3, Townsville Mail Centre, Queensland 4810, Australia. E-mail: g.death@aims.gov.au

Species–environment relationships can be modeled using either individual species or assemblages. For individual species analyses, species are independently related to the environmental variables, whereas for community analyses, species are jointly related to the environment using a single model. Thus community analyses are more constrained. Popular methods used for individual species analysis include generalized linear models (GLM; McCullagh and Nelder 1983), generalized additive models (GAM; Hastie and Tibshirani 1990), and classification and regression trees (CART; Breiman et al. 1984). Comparisons of GLM and GAM vs. CART (Franklin 1998, Vayssières et al. 2000) found that the latter gave better predictions. This is not surprising as trees are well suited for analysis of complex ecological data that may include lack of balance, missing values, nonlinear relationships between variables, and high-order interactions (Breiman et al. 1984, Ripley 1996, De'ath and Fabricius 2000). Ordination methods are widely used for community analysis (Jongman et al. 1995, ter Braak and Prentice 1988, Legendre and Legendre 1998), and typically assume that abundances of individual species vary in a linear or unimodal manner along ecological gradients. The gradients are explained by linear combinations of environmental variables, the latter being used either directly or indirectly in the analysis. When used directly, the species and environmental data jointly determine the gradients. When used indirectly, the structure of the species data is first determined, and is then related to the environmental data. Clustering methods are also used for com-

munity analysis, and with clusters are often used to define community or assemblage types (Anderberg 1973, Legendre and Legendre 1998). The clusters can subsequently be related to environmental values. However, methods that directly determine clusters in terms of environment are less well developed (Gordon 1996). Notable exceptions are various forms of contiguous clustering that impose temporal or spatial ordering on the clusters (Lefkovitch 1978, 1980, Legendre and Legendre 1998).

This paper presents a new method for modeling species–environment data; namely, the multivariate regression tree (MRT). MRT is a natural extension of univariate regression trees, with the univariate response of the latter being replaced by a multivariate response. MRT analyzes community data, but makes no assumptions about the form of relationships between species and their environment. MRT can be used for exploration, description, and prediction, and trees are usually selected on the basis of their predictive accuracy. MRT is a form of multivariate regression in that the response is explained, and can be predicted, by the explanatory variables. However, it is also a method of constrained clustering, because it determines clusters that are similar in a chosen measure of species dissimilarity, with each cluster defined by a set of environmental values. As with unconstrained clustering, each cluster can define an assemblage type, but additionally the environmental values define an associated habitat type.

Following this introduction, a review of the principal concepts of univariate regression trees is presented, because they also form the basis for MRT. The sums of squares MRT is then outlined, and a detailed example follows, which includes additional techniques that help interpret the tree analysis. More general forms of MRT are then discussed, and two types of MRT are defined; one generalizes the sums of squares MRT, and the other is distance based, analogous to distance-based redundancy analysis (db-RDA; Legendre and Anderson 1999, McArdle and Anderson 2000). MRT is then compared to redundancy analysis (RDA; Rao 1964) and canonical correspondence analysis (CCA; ter Braak 1986) using simulated data and a previously published set of field data.

Throughout this paper, response variables are species, data cases are sites, and explanatory variables are environmental characteristics. Hence, for ease of reading, I will often simply refer to species, sites, and environment. Note, however, that applications of MRT are not limited to modeling species in terms of their environment. The choice of response and explanatory variables is unrestricted, other than the response being numeric.

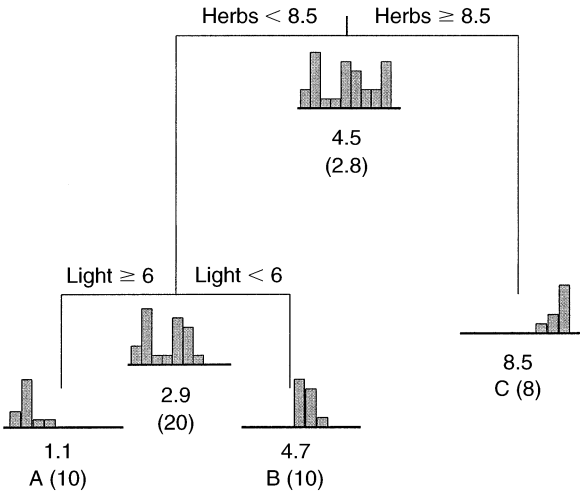

FIG. 1. An example of a univariate regression tree. The response variable is the abundance (0–9 scale) of a species of hunting spider, *Trochosa terricola*, and the explanatory variables are six environmental characteristics (water, sand, twigs, moss, herbs, and light; 0–9 scale). These data are from Van der Art and Smeenk-Enserink (1975). The mean values of *T. terricola* and the number of cases (in parentheses) are displayed at each node; e.g., the overall mean is 4.5 (n = 28). The histograms show the distributions of *T. terricola* values at the nodes. The three-leaf tree is formed by two splits; the first based on herbs (<8.5 and ≥8.5) and then, for low levels of herbs, a second split based on light (<6 and ≥6). The tree explains 93% of the variance of *T. terricola*, with virtually no overlap of values at the three leaves (A, B, and C). The depth of the tree following each split is proportional to the variance explained by the split.

Univariate Regression Trees

Univariate regression trees (URT) explain the variation of a single numeric response variable using explanatory variables that may be numeric and/or categorical (Breiman et al. 1984, Clark and Pregibon 1992, Ripley 1996, De'ath and Fabricius 2000). They do this by growing a tree structure that partitions the data set into mutually exclusive groups, each of which has similar values of the response variable. Starting with all the data represented by a single node at the top of the tree (Fig. 1), the tree is grown by repeated binary splitting of the data. Each split is defined by a simple rule, usually based on a single explanatory variable, and forms two nodes. Splits are (generally) chosen to maximize the homogeneity (minimize the impurity) of the resulting two nodes. The terminal nodes (i.e., unsplit nodes) represent the groups of data formed by the tree, and are also called the leaves of the tree. For a regression tree, impurity of a node is typically defined as the total sum of squares of the response variable values about the node mean, and each split minimizes the total sums of squares within the two nodes formed by the split. Equivalently, this maximizes the between-nodes sums of squares.

The splitting procedure is continued until an overlarge tree is grown, that is then pruned back to the desired size. The size may depend on the objective of the analysis; be it exploration, description, or prediction. For example, tree size can be selected choosing the most accurate predictor from a collection of trees

of varying size. Cross-validation is most often used to select tree size, with the chosen tree having the smallest (or close to) predicted mean square error (Breiman et al. 1984). This tree can be thought of as the ''best predictive tree'' in the sense that, on average, it should give the most accurate predictions.

Trees are summarized by their size (the number of leaves) and overall fit. For a sums of squares regression tree, variation of the response is expressed as variance and overall fit is simply the fraction of variance not explained by the tree. More generally, fit is defined by relative error (RE); the total impurity of the leaves divided by the impurity of the root node (the undivided data). RE gives an over-optimistic estimate of how accurately a tree will predict for new data, and predictive accuracy is better estimated from the cross-validated relative error (CVRE). CVRE varies from zero for a perfect predictor to close to one for a poor predictor.

Extending to Multivariate Regression Trees

A simple way to extend URT to MRT is to replace the univariate response by a multivariate response, and to redefine the impurity of a node by summing the univariate impurity measure over the multivariate response. Thus, to extend the sum of squares URT, impurity can be defined as the sum of squares about the multivariate mean (see Appendix). Geometrically, this is simply the sum of squared Euclidean distances of sites about the node centroid. Each split minimizes the sums of squared distances (SSD) of sites from the centroids of the nodes to which they belong. Equivalently, this maximizes the SSD between the node centroids. It can also be shown that this minimizes the SSD between all pairs of sites within the nodes, and maximizes the SSD between all pairs of sites in different nodes (Digby and Gower 1981, Gower and Hand 1996). This equivalence only holds for sums of squares. Pruning and selecting tree size by cross-validation carry over from the univariate tree provided the prediction error for a new observation is defined analogously (see Appendix). Finally, each leaf of the tree can be characterized by the multivariate mean of its sites, the number of sites at the leaf, and the environmental values that define it. I will refer to this tree as the sums of squares MRT (SS-MRT).

Although the growing and pruning of univariate trees carry over to MRT, new challenges arise. First, because the multivariate response introduces extra complexity, additional ways to interpret the results of an MRT analysis are needed. In *An introductory example of multivariate regression trees* below, three methods to assist that interpretation are outlined, namely (1) identification of species that most strongly determine the splits of the tree, (2) tree biplots to represent group means and species and site information, and (3) identification of species that best characterize the groups.

Second, because MRT is a form of constrained clustering, it is useful to compare the MRT solution to unconstrained clustering using a metric equivalent to the MRT impurity and the same numbers of clusters. If the unconstrained cluster analysis accounts for substantially more of the species variation than an MRT analysis, this could indicate that unobserved factors, additional to the explanatory variables of the tree analysis, are responsible for the difference in explained species variation. However, if the two forms of analysis explain similar amounts of species variation, it is likely that important environmental variables (or their surrogates) have been identified. In such cases, the groups from the two forms of analysis will often coincide substantially. In addition, clustering may be weak, but if there are strong relationships between species and environment, MRT analyses can detect distinct groups not detectable by unconstrained clustering.

Third, MRT can be extended to include impurity measures equivalent to the various measures of species dissimilarity currently used in ordination and clustering techniques. This issue is dealt with in *Beyond sums of squares multivariate regression trees.*

An Introductory Example of a Multivariate Regression Tree

In this example, a SS-MRT is used to relate abundances of 12 species of hunting spiders to six environmental characteristics (water, sand, twigs, moss, herbs, and light). These data, first presented by Van der Art and Smeenk-Enserink (1975), were later analyzed by ter Braak (1986), using various forms of correspondence analysis. Canonical correspondence analysis (CCA) suggested a curvilinear distribution of sites in two dimensions. The first axis was interpreted as a moisture gradient and the second axis as a contrast between sites high in twigs, and sites high in herbs and moss. These data have also been analyzed using principal curves (De'ath 1999*a*), which revealed a strong one-dimensional solution, with largely unimodal distributions of species along the estimated gradient.

In this analysis, species and then sites were standardized to the same mean. Site standardization converts abundances to relative abundances, and is widely used in gradient analyses methods because it increases the strength of the relationship between species dissimilarity and ecological distance for moderate or long gradients (Faith et al. 1987). The tree analysis was also run on data standardized to give chi-squared distances between sites. The tree groups from this analysis were identical in number and composition to those of the tree analysis of the mean standardized data, and are not reported. The species and site standardizations were chosen to permit subsequent comparisons with both redundancy analysis (RDA) and CCA.

The MRT analysis gave a four-leaf tree with the splits based only on twigs and water (Fig. 2). The tree explained 78.8% of the standardized species variance. Species composition varied strongly across the four groups, with sites in groups A (twigs <8; water <2.5)

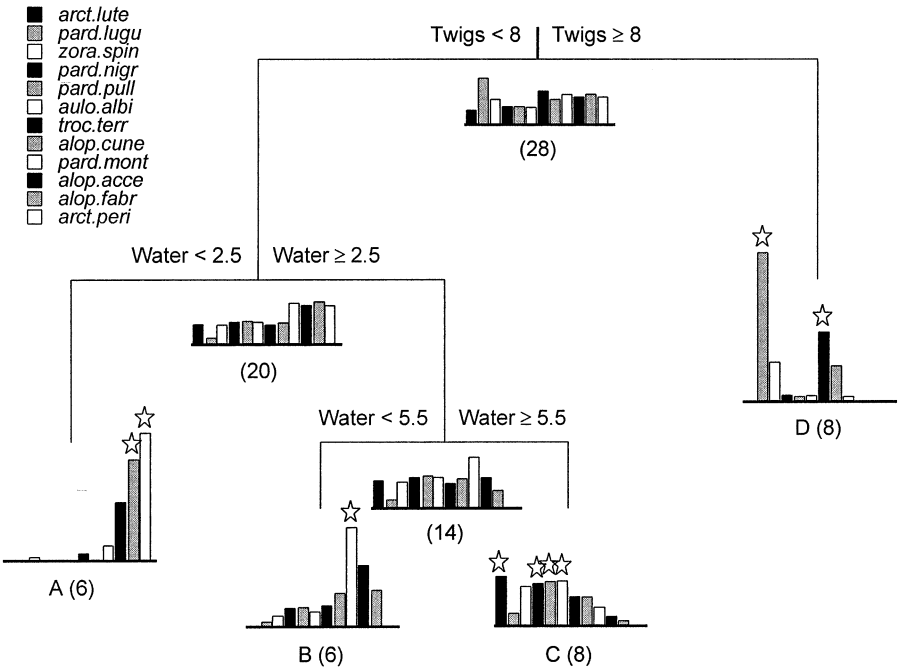

FIG. 2. Multivariate regression tree for the hunting spider data. The data were species standardized to the same mean and site standardized to the same mean; Euclidean distance was used for splitting. Barplots show the multivariate species mean at each node, and the numbers of sites are shown in parentheses. Indicator species for groups A–D are denoted by stars. The cyclical shadings (black, gray, and white) indicate the various species and run from left to right across the barplots; for example, *alop.fabr* and *arct.peri* are indicator species for group A. The species abbreviations are as follows: *alop.acce = Alopecosa accentuata*; *alop.cune = Alopecosa cuneata*; *alop.fabr = Alopecosa fabrilis*; *arct.lute = Arctosa lutetiana*; *arct.peri = Arctosa perita*; *aulo.albi = Aulonia albimana*; *pard.lugu = Pardosa lugubris*; *pard.mont = Pardosa monticola*; *pard.nigr = Pardosa nigriceps*; *pard.pull = Pardosa pullata*; *troc.terr = Trochosa terricola*; *zora.spin = Zora spinimana*.

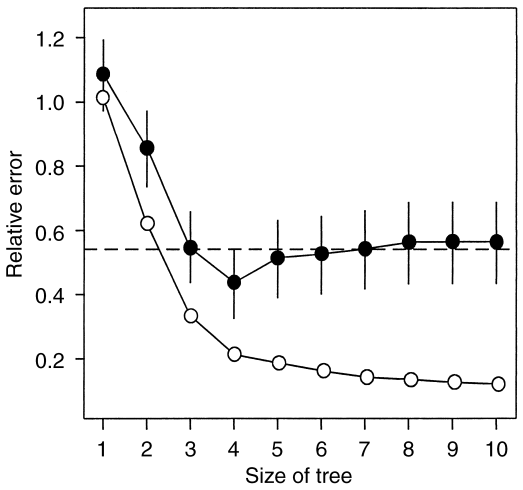

FIG. 3. Selection of the multivariate regression tree for the hunting spider data. The relative error (open circles) decreases with tree size, whereas the cross-validated relative error (filled circles) decreases to a minimum for a tree size of four, and then increases before flattening to a typical plateau. The vertical bars indicate one standard error for the cross-validated relative error, and the dashed line indicates one standard error above the minimum cross-validated relative error and suggests a tree size of four leaves.

and D (twigs ≥8) having only low levels of two species in common. The four associated habitats (A–D) are simply defined by high levels of twigs (D) or by moderate to low levels of twigs with three increasing levels of water (A, B, and C, respectively). The size of the tree was selected by cross-validation, with the four-leaf tree clearly identified as having the smallest estimated predictive error (Fig. 3). More typically, the cross-validations generate a series of trees that predict only marginally worse than the best predictive tree. From this series, the smallest tree within one standard error of the best is often selected (the 1-SE rule; Breiman et al. 1984).

Examining the splits

Inspection of the barplots at each node, and at the four leaves, shows how individual species contribute to each split, and to the composition of the four final groups. For example, *Pardosa lugubris* strongly determines the first split and is the dominant species of group D (Fig. 2). It also has a minimal presence in the left hand branch of the tree, and thus does not contribute to subsequent splits that form groups A, B, and C. In this example, the barplots effectively illustrate the species distributions throughout the tree, however, numerical summaries of tree characteristics are also

TABLE 1. Tabulation of species variance for the tree analysis of the hunting spider data.

Species	Species variance (%) explained by tree splits and whole tree				
	Twigs < 8	Water < 2.5	Water < 5.5	Tree total	Species total
Arctosa perita	1.90	15.54	0.00	17.45	20.85
Alopecosa fabrilis	2.33	6.72	0.79	9.83	13.44
Pardosa monticola	1.75	1.34	5.06	8.16	10.82
Alopecosa accentuata	1.93	0.70	2.15	4.78	5.77
Arctosa lutetiana	0.47	0.71	1.78	2.96	4.37
Aulonia albimana	0.35	0.90	0.70	1.96	3.28
Pardosa pullata	0.47	0.99	0.48	1.94	2.53
Pardosa nigriceps	0.20	0.88	0.41	1.49	2.23
Zora spinimana	0.80	0.64	0.64	2.08	4.04
Alopecosa cuneata	0.25	0.84	0.02	1.11	2.89
Pardosa lugubris	23.22	0.02	0.04	23.28	24.41
Trochosa terricola	3.29	0.40	0.05	3.74	5.38
Total species variance	36.97	29.69	12.12	78.78	100.00

Notes: The total species variance is partitioned by species, the whole tree, and the three splits of the tree. The first split is dominated by *Pardosa lugubris* (23 percentage points of 37% of the total species variance explained by that split), the second by *Arctosa perita* (15 percentage points of 30%), and the third by *Pardosa monticola* (5 percentage points of 12%).

useful, particularly when the number of species is large. For example, the contributions of individual species at each split and how well each species is explained by the tree can be quantified, by tabulating the explained variance at each split for each species (Table 1). The variance of *Pardosa lugubris* comprises 24.4% of the total species variance, of which 23.3% is explained by the tree, with 23.2% explained by the first split (Table 1). *Arctosa perita* and *Alopecosa fabrilis* largely determine the second split (water <2.5), and *Pardosa monticola* dominates the third split (water <5.5). Despite these four species accounting for 69.5% of total species variance, the remaining eight species are also well separated by the four tree groups, with the least well explained, *Alopecosa cuneata*, having 38.4% (1.11 of 2.89%) of its variance explained.

Tree biplots

For URT, each group is characterized by the mean response, and comparisons of groups is straightforward. However, for MRT, the mean response is multivariate, and this makes comparisons of groups more difficult. One way to examine the structure of multivariate data is to plot them in a low-dimensional space using a principal components biplot (Gabriel 1971, ter Braak 1994, Gower and Hand 1996: Fig. 4). The distance biplot (ter Braak 1994) gives the best least squares representation of the group means and is consistent with the MRT maximizing the sums of squares between groups; i.e., the biplot and MRT are using the same metric. For these data, the species have unimodal distributions, and thus each species is located by its weighted mean of the group means (also weighting on group size). Thus, species are located close to groups with relatively high abundance of that species. Species values of individual sites can be projected onto the plot and identified by group (Fig. 4). This shows the homogeneity of sites within groups and can also reveal outlier sites. These points represent the observed species values and are referred to as supplementary points. Observed values were originally used in constrained (canonical) biplots (ter Braak 1986), however fitted values from the underlying model are now more typical (Oksanen 1987, ter Braak 1994, Palmer 1993). For tree biplots, the fitted values are the group means and the observed values are species values of individual sites; hence, both are shown in the tree biplot.

Finally, the interset correlations can be calculated for each axis (ter Braak 1986) using the scores of the group means and the scores of the observed values. These correlations estimate the strength of the species–environment relationship. Axes with high interset correlations (typically >0.8), and that also account for a substantial proportion of the species variance explained by the environmental variables (i.e., the between-group variance), should be retained in the biplot representation.

Identifying indicator species

Species that characterize each group can be identified using an indicator species index based on relative abundance and relative frequency of occurrence (Dufrêne and Legendre 1997). The index is defined as the product of relative abundance and relative frequency of occurrence of the species within a group. If there are no occurrences of the species within a group, the index takes the value of zero, increasing to 100 if the species occurs at all sites within the group and does not occur at any other site. The index can be calculated for each species–group combination, and species with high index values for a group are indicator species for that group (Fig. 2). The index distinguishes between ubiquitous species that dominate many groups in absolute abundance, and species that occur consistently within single groups but have low abundance. For the four MRT groups, the index values for the 12 species vary from 100 for *Arctosa perita* (which occurs in all sites of group A and no other sites), down to 29 for *Alo-*

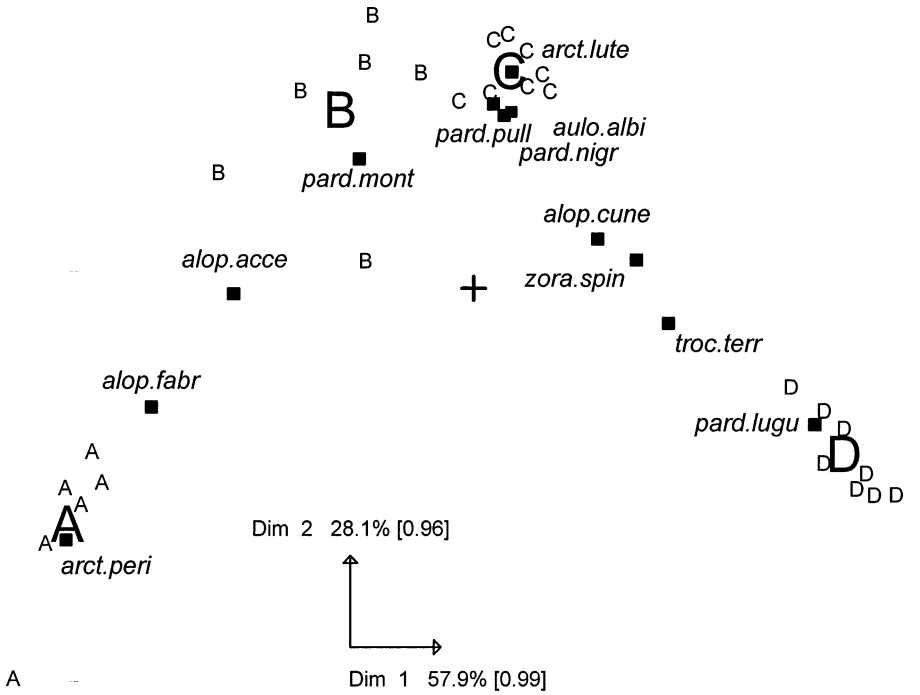

FIG. 4. Principal-components biplot of the four group means from the tree analysis of Fig. 2. The large letters A–D represent the multivariate group means and correspond to groups A–D in Fig. 2. The individual sites denoted by small letters A–D suggest a good fit at all sites. Each species label is located at its weighted mean from the four group means. The first two dimensions (Dim 1 and Dim 2) account for 57.9% and 29.1% of the between-groups sums of squares, respectively, with interset correlations of 0.99 and 0.96. Species names are as in Fig. 2.

pecosa cuneata (which has moderate abundance and presence across three groups). Nine of the 12 species had high indicator values for single group (range: 57–100). The remaining three species had moderate values for two or three groups (27–48). For each of the three species, the values were very similar across groups, thereby suggesting they were indicators for composite groups.

Comparing tree groups with unconstrained clusters

Unconstrained cluster solutions, comprising two to four clusters, were derived using complete linkage hierarchical clustering (Euclidean metric). For each solution, the clusters were refined using *k*-means clustering (Hartigan and Wong 1979) to minimize the within-cluster sums of squares. This procedure gives compact clusters, and thus provides a stringent comparison with the constrained clusters defined by the MRT. The tree and the unconstrained clusters had almost identical membership for two to four clusters and explained the species variance equally well. Together with the large proportion of species variance explained by the tree, this supports the conclusion that the two of the six environmental variables that formed the tree (twigs and water) adequately account for the species variance.

BEYOND SUMS OF SQUARES MULTIVARIATE REGRESSION TREES

Regression trees are modular in the sense that the impurity measure, splitting criteria and prediction error are all independent of the growing and pruning processes (Breiman et al. 1984). This presents opportunities to develop MRT using impurity measures, additional to SS, that are useful to ecologists, e.g., robust measures and measures that correspond to popular measures of species dissimilarity. To do this, two strategies are adopted, and these lead to additive MRT (A-MRT) and distance-based MRT (db-MRT).

Additive multivariate regression trees

Following the method of extending SS-URT to SS-MRT, i.e., adding the univariate measure of impurity over the multivariate response, measures other than sums of squares about the mean can be used, e.g., sums of absolute deviations about the median. These impurity measures are examples of additive distances (Gower and Hand 1996), with each variable of the multivariate response contributing to the impurity independently of the others. One useful aspect of A-MRT is the potential to make them robust to outliers. For URT, the two most widely used impurity measures are sums of squares about the mean, and sums of absolute deviations about the median (LAD; Breiman et al. 1984), with LAD being more robust than SS to outliers. Though the median is an extremely robust measure of location, the sum of absolute deviations about medians is sensitive to outliers, and this reduces the robustness of LAD trees. Sums of squares about trimmed means,

where extreme values are simply dropped from each group, are even more robust, but are less widely used (SPSS Inc. 1997).

Distance-based multivariate regression trees

Multivariate trees can be formulated to work directly from a dissimilarity matrix. Treating the dissimilarities as distances, the clusters can be formed by splitting the data on environmental values that minimize the intersite SSD (sums of squared distances) within clusters. SS-MRT is based on Euclidean distance, a measure of dissimilarity often used for the analysis of species–environment relationships. However, other measures are often preferred. For example, most forms of gradient analysis depend, either explicitly or implicitly, on a strong linear relationship between some measure of species dissimilarity and ecological distance. Analyses based on Euclidean distance often fail for moderate to long gradients, because compared to alternatives such as site standardized Bray-Curtis and extended dissimilarity, it is only weakly correlated with ecological distance (Faith et al. 1987, De'ath 1999*b*). The relative merits of various measures have been widely discussed (e.g., Faith et al. 1987, Legendre and Legendre 1998) and I will not add to that debate. Rather, I am assuming that a particular measure has been chosen, and that if the resulting matrix of site dissimilarities is treated as distances, then these distances adequately represent the structure in the data. Distance-based MRT (db-MRT) can be defined in terms of within-group intersite SSD, independent of the particular measure of dissimilarity chosen for the analysis. The impurity of the node is defined as the sum of within-group intersite SSD, the splitting criterion maximizes the reduction of impurity at a split, and the prediction error is defined as the SSD of the prediction site from all other sites, minus the within-group SSD of other sites (see the Appendix). If the dissimilarities used in a db-MRT are Euclidean distances, then SS-MRT and db-MRT are exactly equivalent, because as noted earlier, minimizing squared Euclidean distances of sites about node centroids is identical to minimizing within-node intersite squared Euclidean distances. Distance-based MRT extends SS-MRT, just as distance-based redundancy analysis (db-RDA: Legendre and Anderson 1999, McArdle and Anderson 2000) extends RDA. In both cases, the distance-based methods allow any measure of dissimilarity to be used, not just Euclidean distance.

Comparison of additive and distance-based trees

Despite the coincidence of SS-MRT and Euclidean db-MRT, A-MRT and db-MRT are quite different. For example, the impurity of A-MRT can be defined in many ways and focuses on the typical characteristic of the node, e.g., the multivariate mean. Conversely, the impurity of db-MRT is always defined as the within-group intersite SSD (see Appendix) and the focus is on individual sites with all sites contributing equally to the impurity of a node. The following two examples illustrate these differences. First, if the impurity of an A-MRT is defined as sums of squares about trimmed means, sites with extreme values may have no influence on any splits of the tree. This cannot occur for db-MRT because every site will always influence at least one split (the first). Second, because db-MRT depends solely on the dissimilarities and the environmental data, the species observations themselves are not needed. Thus db-MRT, but not A-MRT, can be used in pairwise comparison studies, for which only the dissimilarities (or similarities) between subjects are available (Thurstone 1927).

There are, however, links between A-MRT and db-MRT. First, there are several measures of species dissimilarity that are equivalent to Euclidean distances of suitably scaled species data; e.g., chi-squared and chord dissimilarity. For such measures, db-MRT can be determined by simply scaling the data appropriately and then using SS-MRT. This is not possible for species dissimilarities based on sums of absolute deviations; e.g., Bray-Curtis dissimilarity. Secondly, for many species dissimilarities, a principal coordinates analysis will generate site coordinates such that the Euclidean distances between the sites are exactly proportional to the dissimilarities (i.e., they are Euclidean embeddable [Gower and Hand 1996]). A SS-MRT analysis using these coordinates as the response data will give the identical tree as the db-MRT of the original dissimilarities. For these two situations, SS-MRT analysis can be used to determine db-MRT more efficiently; particularly for large dissimilarity matrices.

Comparison of Multivariate Regression Trees with Redundancy Analysis and Canonical Correspondence Analysis

Redundancy analysis (RDA; Rao 1964) and canonical correspondence analysis (CCA; ter Braak 1986) are widely used methods of direct gradient analysis (DGA). Similar to RDA and CCA, MRT seeks to maximize the species variation explained by the environmental variables. Unlike RDA and CCA, MRT is not a method of DGA, in the sense of locating sites along ecological gradients. However, MRT does locate sites (albeit grouped) in an ecological space defined by the environmental variables. For RDA and CCA, the solution lies in the linear space defined by the environmental values, whereas for MRT the solution space is nonlinear and includes interactions between the environmental variables. For RDA and CCA, interactions have to be individually added to the environmental variables, making such effects more difficult to detect, particularly so when the number of environmental variables is large. However, MRT automatically detects interactions because each split partitions the data into independent subsets, each of which is analyzed independently.

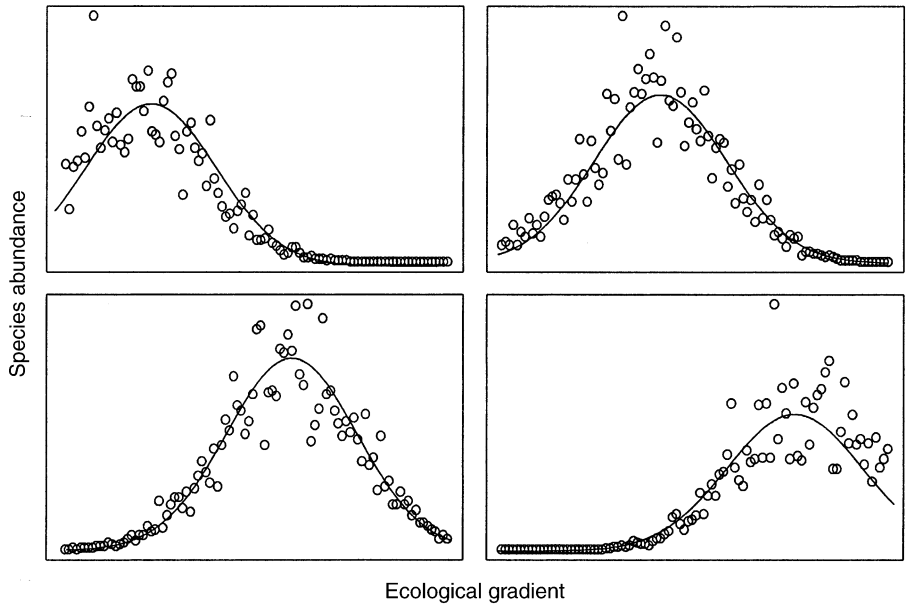

FIG. 5. Simulated species abundances along a one-dimensional ecological gradient. Sites and species modes are distributed evenly along the gradient.

Comparisons using simulated data

Abundance data were simulated for 100 sites and four species each with a Gaussian response curve driven by a single ecological gradient (Fig. 5). The sites and species optima were uniformly distributed along the gradient. The span (species tolerance) and maxima of the response curves were fixed for all species, and Poisson noise was added to the abundances. The gradient length was eight standard deviations of the response curves. These data closely satisfy the conditions for which correspondence analysis and CCA are effective estimators of ecological gradients (ter Braak 1986).

The data were analyzed using SS-MRT, RDA, and CCA. For each method, the data were: (1) untransformed, (2) square-root transformed, and (3) square-root transformed and site standardized to the same mean. CCA implicitly site standardizes the species and thus (3) was not used in the CCA analyses. The transformation and standardization increasingly remove the arch effect present in principal components biplots of the species data by linearizing the relationship between species dissimilarity (in this case Euclidean) and ecological distance. Linear and quadratic terms in site locations on the ecological gradient were used as explanatory variables for RDA and CCA. For MRT, only the linear term was used because the (noncentered) quadratic has the same rank order of values as the linear term and hence it does not contribute additional splits. For RDA and CCA, models were fitted with both linear and quadratic terms, and with only the linear term.

For MRT, the variance explained by the constrained and unconstrained models (unconstrained clustering) differed little and was >91%, irrespective of transformations and standardization (Table 2). Individual species analyses of the simulated data using univariate regression trees gave marginally higher mean CVRE values compared to MRT for all three analyses. RDA differed greatly between constrained and unconstrained solutions for other than the transformed and standardized data, whereas CCA showed good agreement between the two due to the implicit use of species and site standardization. The variance explained by the first two axes of ordination plots, the species–environment correlations and permutation tests ($P < 0.05$; ter Braak 1992) all indicated a two-dimensional solution was required for all RDA and CCA analyses. Strong arch effects were present in all RDA and CCA biplots (not shown). Differences between the transformed and standardized RDA analysis and the transformed CCA analysis were minimal (Table 2).

The cross-validated relative error (CVRE) of all methods decreased (i.e., predictions were more accurate) as the species–environment was linearized by the transformation and site standardization (Table 2). The differences were small for MRT, which predicted comparatively well irrespective of transformation and standardization. The MRT, using only linear terms in site locations, predicted far better than the linear RDA and CCA models. The differences in prediction for MRT (linear), RDA (quadratic), and CCA (quadratic) were small when the data were transformed and standardized.

These results show the necessity of satisfying model assumptions for RDA and CCA. They also demonstrate two clear advantages of MRT; namely (1) the absence

TABLE 2. Comparison of analyses of simulated data using three methods: sums of squares multivariate regression tree (MRT), redundancy analysis (RDA), and canonical correspondence analysis (CCA). The data are illustrated in Fig. 5.

Analysis	Parameters	Untransformed	Square-root transformation	Square-root + site standardization
MRT	Variance explained (%)	91.1 (94.2)	93.8 (94.7)	93.9 (94.2)
	Size of tree	10	12	10
	CVRE (L)	0.144	0.099	0.097
	CVRE (URT)	0.169	0.104	0.107
RDA	Variance explained (%)	55.5 (86.5)	74.4 (95.0)	90.3 (96.2)
	CVRE (Q)	0.475	0.276	0.098
	CVRE (L)	0.732	0.609	0.242
CCA	Variance explained (%)	88.3 (93.7)	91.3 (96.8)	Not applicable
	CVRE (Q)	0.124	0.107	
	CVRE (L)	0.308	0.237	

Notes: The response variables are the four species abundances, and the explanatory variable(s) are linear and quadratic terms in site locations on the ecological gradient. For MRT only the linear term was used, and for RDA and CCA linear and quadratic terms, and then only the linear term, were used. For each method, the data were (1) untransformed, (2) square-root transformed, and (3) square-root transformed and site standardized to the same mean (not applicable to CCA). For each method, the percentage of variance explained by the constrained and equivalent unconstrained analyses (in parentheses; clustering, principal components, and correspondence analysis, respectively) is shown. The cross-validated relative error (CVRE: L = linear, Q = quadratic, and URT = individual species analysis using univariate regression trees) of each method decreases (predictions are more accurate) with transformation and site standardization of the data.

of model assumptions that results in a greater robustness and (2) the invariance of trees to monotonic transformations of explanatory variables. These analyses also raise the issue of inclusion of polynomial effects or detrending of CCA analyses for moderate to long gradients (ter Braak 1986, Palmer 1993). If only linear environmental effects are included, gradients may be accurately estimated, but species composition may be both poorly explained and predicted (Table 2). It also follows that the proportion of explained species variance is not a good indicator of how well gradients have been estimated. For example, in the analyses of our untransformed simulated data, the RDA and CCA models with only linear explanatory variables exactly recover the site locations, despite having relatively lower explained species variance and species–environment correlations. This will always be the case when the latent variable(s) representing the gradient(s) are an exact linear combination of the environmental data.

Comparisons using the hunting spider data

As outlined in *An introductory example of multivariate regression trees*, the SS-MRT analysis of the standardized spider data explained 78.8% of scaled species variance with a four-leaf tree based on two environmental variables (twigs and water). For the same data, RDA explained 73.6% using all six environmental variables, with 61.7%, 28.9%, and 6.1% explained by the first three axes, respectively (Fig. 6). The species–environment correlations were 0.99, 0.93, and 0.61, suggesting a two-dimensional solution. Omitting the nonsignificant variables (herbs and moss; permutation tests; $P > 0.05$), the percentage variance explained drops to 71.2%. Using only twigs and water, the percentage variance explained drops to 61.5%, 12.1% less than the full RDA model and 17.3% less than the four-leaf tree. In addition to MRT explaining more of the species variance than RDA (RE = 0.21 for MRT vs. 0.26 for RDA), it was also a marginally more accurate predictor (CVRE = 0.44 [SE = 0.07] for MRT vs. 0.47 [0.04] for the two-dimensional RDA model based on all six environmental variables). Individual species analyses using univariate regression trees were similarly accurate (CVRE = 0.45 [SE = 0.09]) to MRT.

It could be argued that RDA explains less species variance due the unimodal distributions of the species of spider, and that CCA is a more appropriate form of analysis. However, the species and site standardizations of the RDA analysis largely removes that nonlinearity, in an analogous manner to the species and site standardizations implicit in CCA. Also, comparison of the configurations of site locations (gradients) from the RDA analysis and from the CCA analysis using untransformed species data showed the two configurations to be almost identical (two-dimensional Procrustes analysis [Gower 1975]: Sibson's stress = 0.018). The variances explained and species–environment correlations from the two analyses were also very similar, although CCA performs a weighted analysis of the species data, and hence statistics from the CCA and RDA analyses are only approximately equivalent.

DISCUSSION

Multivariate regression trees (MRT) are a powerful method for the community analysis of species–environment data; one that can be used for exploration, description, and prediction. MRT can be viewed and used in many ways. First, they are a regression technique that both explains and predicts species abundances from environmental variables. The regression is constrained in that all species are explained (and predicted) by the same environmental values. This is similar to redundancy analysis, which is a form of constrained multivariate linear regression. MRT can also be seen as a form of

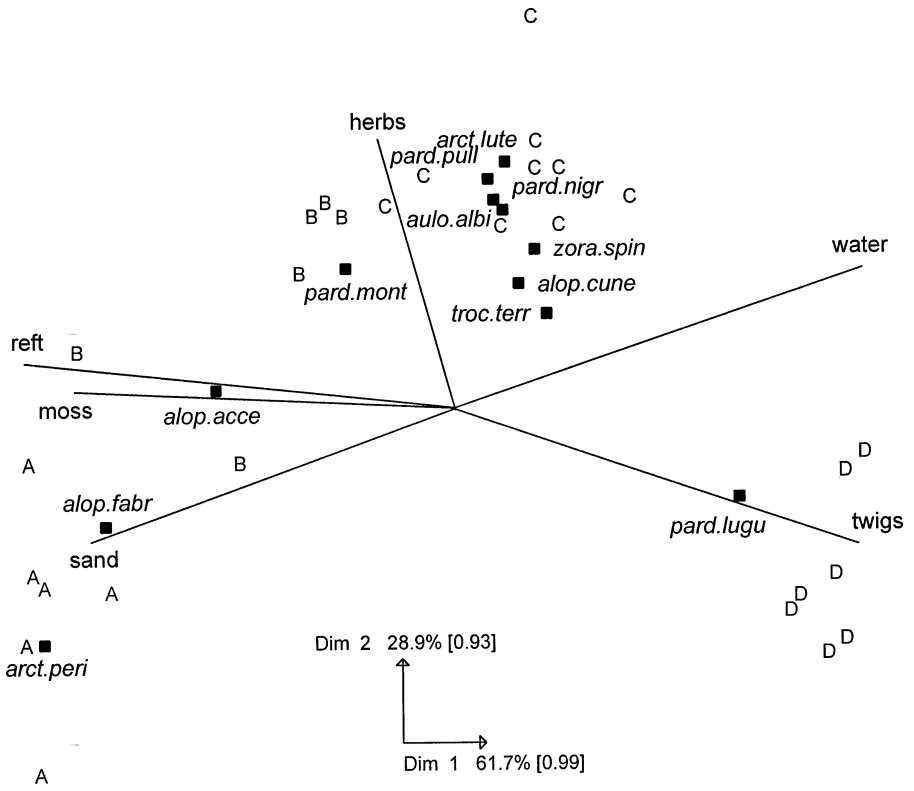

FIG. 6. Redundancy analysis of hunting spider data. The data were species standardized to the same mean and site standardized to the same mean. The six explanatory variables (twigs, water, herbs, reflected light, moss, and sand) explained 73.6% of the species variance compared to 78.8% of the variance explained by the multivariate tree analysis (Figs. 2 and 4). The first two dimensions (Dim 1 and Dim 2) account for 61.7% and 28.9% of the explained species variance, respectively, and the interset correlations are 0.99 and 0.93. The individual sites are denoted by small letters A–D corresponding to the tree groups, and each species label is located at its weighted mean from all sites. As will generally be the case, the plot is very similar to the tree biplot.

constrained cluster analysis that, dependent on transformations, standardizations, and choice of splitting criterion, can relate different aspects of species composition to environmental data. In this form, the clusters defined by MRT define species assemblages and associated environment types in a simple manner not available in other techniques. MRT can present a comprehensive view of species–environment relationships by (1) displaying the annotated tree; (2) tabulating variation at the splits of the tree; (3) identifying indicator species to characterize groups; (4) displaying the group means, species, and sites in a low-dimensional space; and (5) comparing the hierarchy of tree groupings with the equivalent unconstrained clusters.

Two forms of MRT have been outlined, namely, additive MRT (A-MRT) and distance-based MRT (db-MRT). A-MRT works directly from the species–environment data matrix, with the tree clusters defined by a chosen measure suitable to the problem at hand, e.g., a robust measure if outliers are problematic. Although this work has focused on sums of squares MRT, other measures can be used as a basis. For example, sums of absolute deviations about medians give a robust MRT. Also, if the species–environment data matrix is not available, or if a particular measure of dissimilarity is preferred, then db-MRT can be constructed directly from the dissimilarity matrix. These two forms of MRT, additive and distance based, are not exhaustive. For example, nonadditive impurity measures that take into account relationships between species (e.g., Mahalanobis distance) could be used. Established unconstrained clustering criteria based on the trace and determinant of the within-groups covariance matrix (Krzanowski and Marriot 1995) could also be used as measures of impurity. This flexibility in the choice of method for forming clusters is a significant strength of MRT.

Though inherently a clustering technique suited to identifying assemblage types, MRT is an alternative to, but also complements, constrained ordination techniques such as redundancy analysis (RDA) and canonical correspondence analysis (CCA). MRT differs from these techniques in that (1) MRT is a divisive technique, RDA and CCA model continuous structure; (2) MRT emphasizes local structure and interactions between environmental effects, RDA and CCA determine global structure; and (3) MRT does not assume particular relation-

ships between species abundances and environmental characteristics, RDA and CCA respectively assume linear and unimodal distributions of species along gradients. These differences make MRT and RDA/CCA ideal complementary techniques. For example, MRT can be used to determine interactions between environmental variables, a task that is difficult for RDA and CCA, and the interactions could then be included in an RDA or CCA analysis. MRT has been compared with RDA and CCA analysis in examples that should favor the latter techniques; namely strong gradients with a continuous distribution of sites. Despite this, MRT outperformed or matched these techniques in both explaining and predicting species composition. The advantage of MRT increases with the strength of interactions and the nonlinearity of relationships between species composition and the environmental variables. I have found this to be the case over a broad range of ecological data sets and simulations. MRT and RDA/CCA present different descriptions of the relationships between species abundances and their environment; the former based on discrete environmental types, and the latter on gradient(s) linearly related to the environmental values. Which of these descriptions is most useful is open to debate, but clearly MRT is an interesting alternative to RDA and CCA.

As a form of community analysis, MRT compares favorably with the equivalent individual species analysis based on univariate regression trees (URT). For both the data simulation and field data examples, the accuracy of MRT was at least as good as that of URT. Given MRT is a single tree, and is thus easy to interpret compared to the multiple trees of the URT analysis, this suggests the potential to both accurately predict, and also to generate a simple descriptive model, using a single analysis.

Most univariate tree concepts and practices carry over to MRT, e.g., tree selection based on cross-validation, and methods for determining splits of the tree. MRT also inherits characteristics of univariate regression trees in that (1) mixtures of unlimited numbers of numeric and categorical variables can be used as explanatory variables, (2) they are invariant to monotonic transformations of explanatory variables, (3) interactions between explanatory variables are automatically detected, (4) they are easy to construct, and the resulting groups are often simple to interpret, (5) they are robust to the addition of pure noise response and/or explanatory variables, and to the collinearity of explanatory variables, and (6) they handle missing values in explanatory variables with minimal loss of information.

Statistical Analyses

All data analyses used the S-Plus statistical software (Statistical Sciences, a division of Mathsoft, Seattle, Washington, USA). The library of tree routines (RPART; Recursive PARTitioning) developed by T. Therneau (*unpublished manuscript*) was extended by inclusion of additional C routines to fit multivariate regression trees. The index for determining indicator species was calculated as in Dufrêne and Legendre (1997). Code to simulate data sets was written in S-Plus. The hierarchical and *k*-means clustering used S-Plus routines.

A library of S-Plus functions for tree analyses (including multivariate regression trees) is available in ESA's Electronic Data Archive; see Supplementary Materials.

Acknowledgments

Thanks are due to Dave Roberts and two anonymous referees whose constructive comments substantially improved this paper. Comments from Pierre Legendre, Steve Delean, and Katharina Fabricius are also appreciated. I would also like to thank John Birks, Cajo ter Braak, Allan Stewart-Oaten, Lee Belbin, and Mike Palmer for their creative works and encouragement over recent years. Finally but not least, thanks are due to Terry Therneau for the development of the RPART library. This work was funded by the Cooperative Research Centre for Great Barrier Reef World Heritage Area.

References

Anderberg, M. R. 1973. Cluster analysis for applications. Academic Press, London, UK.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. G. Stone. 1984. Classification and regression trees. Wadsworth International Group, Belmont, California, USA.

Burnham, K. P., and D. R. Anderson. 1998. Model selection and inference: a practical information theoretic approach. Springer Verlag, New York, New York, USA.

Clark, L. A., and D. Pregibon. 1992. Tree-based models. Pages 377–420 *in* J. M. Chambers and T. J. Hastie, editors. Statistical models in S. Wadsworth and Brooks. Pacific Grove, California, USA.

De'ath, G. 1999*a*. Principal curves: a new technique for indirect and direct gradient analysis. Ecology **80**:2237–2253.

De'ath, G. 1999*b*. Extended dissimilarity: a method of robust estimation of ecological distances from high beta diversity data. Plant Ecology **144**(2):191–199.

De'ath, G., and K. E. Fabricius. 2000. Classification and regression trees: a powerful yet simple technique for the analysis of complex ecological data. Ecology **81**:3178–3192.

Digby, P. G. N., and J. C. Gower. 1981. Ordination between and within groups applied to soil classification. Pages 53–75 *in* D. F. Merriam, editor. Down to earth statistics: solutions looking for geological problems. Syracuse University Geology Contributions. Syracuse, New York, USA.

Draper, D. 1995. Assessment and propagation of model uncertainty (with discussion). Journal of the Royal Society Series B **57**:45–97.

Dufrêne, M., and P. Legendre. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecological Monographs **67**:345–366.

Faith, D. P., P. R. Minchin, and L. Belbin. 1987. Compositional dissimilarity as a robust measure of ecological distance. Vegetatio **69**:57–68.

Franklin, J. 1995. Predictive vegetation mapping: geographic modeling of biospatial patterns in relation to environmental gradients. Progress in Physical Geography **19**:474–499.

Franklin, J. 1998. Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. Journal of Vegetation Science **9**:733–748.

Gabriel, K. R. 1971. The biplot graphical display of matrices with application to principal component analysis. Biometrika **58**:453–467.

Gordon, A. D. 1996. A survey of constrained classification. Computational Statistical Data Analysis **21**:17–29.

Gower, J. C. 1975. Generalized procrustes analysis. Psychometrika **40**:33–51.
Gower, J. C., and D. J. Hand. 1996. Biplots. Chapman and Hall, London, UK.
Guisan, A., and N. E. Zimmerman. 2000. Predictive habitat distribution models in ecology. Ecological Modeling **135**: 147–186.
Hartigan, J. A., and M. A. Wong. 1979. Algorithm AS136. A *K*-means clustering algorithm. Applied Statistics **28**: 100–108.
Hastie, T. J., and R. J. Tibshirani. 1990. Generalized additive models. Chapman and Hall, London, UK.
Iverson, L. R., and A. M. Prasad. 1998. Predicting abundance of 80 tree species following climate change in the eastern United States. Ecological Monographs **68**:465–485.
Johnson, D. H. 1999. The insignificance of hypothesis testing. Journal of Wildlife Management, **63**(3):763–772.
Jongman, R. H. G., C. F. J. ter Braak, and O. F. R. Tongeren. 1995. Data analysis in community and landscape ecology. Second edition. Cambridge University Press, Cambridge, UK.
Krzanowski, W. J., and F. H. C. Marriot. 1995. Multivariate analysis: part 2. Classification, covariance structures and repeated measurements. Arnold, London, UK.
Lefkovitch, L. P. 1978. Cluster generation and grouping using mathematical programming. Mathematical BioScience **41**: 91–110.
Lefkovitch, L. P. 1980. Conditional clustering. Biometrics **36**:43–58.
Legendre, P., and M. J. Anderson. 1999. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. Ecological Monographs **69**:1–24.
Legendre, P., and L. Legendre. 1998. Numerical ecology. Second English edition. Elsevier, Amsterdam, The Netherlands.
McArdle, B. H., and M. J. Anderson. 2000. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology **82**:290–297.
McCullagh, P., and J. A. Nelder. 1983. Generalized linear models. Chapman and Hall, London, UK.
Oksanen, J. 1987. Problems of joint display of species and site scores in correspondence analysis. Vegetatio **72**:51–57.
Palmer, M. W. 1993. Putting things in even better order: the advantages of canonical correspondence analysis. Ecology **74**:2215–2230.
Rao, C. D. 1964. The use and interpretation of principal components analysis in applied research. Sankhyã A **26**: 329–358.
Ripley, B. D. 1996. Pattern recognition and neural networks. Cambridge University Press, Cambridge, UK.
SPSS Inc. 1997. SYSTAT Version 7.0 for Windows. SPSS, Prentice Hall, New Jersey, USA.
Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions (with discussions). Biometrika **64**: 29–35.
ter Braak, C. J. F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology **67**:1167–1179.
ter Braak, C. J. F. 1992. Permutation versus bootstrap significance tests in multiple regression and ANOVA. Pages 79–86 *in* K. H. Jöckel, G. Rothe, and W. Sendler, editors. Bootstrapping and related techniques. Springer Verlag, Berlin, Germany.
ter Braak, C. J. F. 1994. Canonical community ordination. Part 1: basic theory and linear methods. Ecoscience **1**(2): 127–140.
ter Braak, C. J. F., and I. C. Prentice. 1988. A theory of gradient analysis. Advances in Ecological Research **18**: 271–317.
Thurstone, L. L. 1927. The method of paired comparisons for social values. Journal of Abnormal Psychology **31**:384–400.
Van der Aart, P. J., and N. Smeenk-Enserink. 1975. Correlations between distributions of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area. Netherlands Journal of Zoology **25**:1–45.
Vayssières, M. P., R. E. Plant, and B. H. Allen-Diaz. 2000. Classification trees: an alternative non-parametric approach for predicting species distributions. Journal of Vegetation Science **11**:679–694.

APPENDIX

Growing multivariate regression trees

In order to grow and prune any tree, and to select the tree size by cross-validation, it is necessary to define (1) the impurity of a node, (2) the rule for splitting nodes, and (3) the prediction error for a new observation (Breiman et al. 1984). The overall objective is to minimize the total impurity of the groups, and hence the splitting rule (generally) maximizes the reduction of impurity at each split. The prediction error for new observations is needed to determine the best predictive tree by cross-validation. The impurity and prediction error for A-MRT using sums of squares and least absolute deviations, and for db-MRT using any dissimilarity measures, are shown in Table 3.

TABLE A1. Impurity measures and prediction error for multivariate regression trees.

Tree description	Impurity	Prediction error
Multivariate sums of squared deviations about the mean (SS-MRT)	$\sum_{i,j} (x_{ij} - \bar{x}_j)^2$	$\sum_{j} (x^* - \bar{x}_j)^2$
Multivariate sums of absolute deviations about the median (LAD-MRT)	$\sum_{i,j} \lvert x_{ij} - \tilde{x}_j \rvert$	$\sum_{j} \lvert x^* - \tilde{x}_j \rvert$
Distance-based (db-MRT)	$\sum_{i>k,k} d_{ik}^2$	$\sum_{i} \frac{d^{*2}}{n} - \sum_{i>k,k} \frac{d_{ik}^2}{n^2}$

Notes: Notation: x_{ij} denotes the species data for site i and species j, x^* denotes a new observation, $\bar{x}$ and $\tilde{x}$ denote mean and median, respectively, d_{ik}^2 and d_i^{*2} denote the squared dissimilarities between sites i and k, and between a new observation and site i, respectively, and n denotes the number of cases in the prediction error group.

Tree biplots

The tree biplots are simply a distance biplot of the fitted and observed species values from the MRT analysis. The information needed to generate the plot can be calculated as follows. Suppose $\mathbf{Y}$ is the $n \times p$ matrix of species observations (columns are species and rows are sites), and $\hat{\mathbf{Y}}$ is the $n \times p$ matrix of fitted values (i.e., the group means). Let $\hat{\mathbf{Y}} = \mathbf{U}\,\mathbf{d}\,\mathbf{V}^{\mathrm{T}}$ be the singular value decomposition of $\hat{\mathbf{Y}}$. The display points for the group means (the fitted values) are $\hat{\mathbf{Y}}\mathbf{V} = \mathbf{U}\,\mathbf{d}$, where we need to plot only one point from each group. The display points for sites (observed species values) are $\mathbf{Y}\,\mathbf{V}$. The species can be represented by either vectors from the plot origin given by $\mathbf{U}$, or by locating them at their weighted averages from the group means given by $\mathbf{S}^{-1}\,\hat{\mathbf{Y}}^{\mathrm{T}}\,\hat{\mathbf{Y}}\,\mathbf{V}$ where $\mathbf{S}$ is the column sums of $\hat{\mathbf{Y}}$.

The tree biplot is equivalent to the RDA biplot in the fol-

lowing way. First, run the MRT analysis. Then run an RDA using the same species data as the multivariate response, and for explanatory variables use indicator variables, one for each cluster of the tree, where the variable takes the value 1 if a site is in a cluster and 0 otherwise. The fitted values of the RDA are the same as those of the tree, and the RDA biplot will be identical to the tree biplot, provided both fitted and observed values are plotted.

SUPPLEMENTARY MATERIAL

A library of S-Plus functions for tree analyses (including multivariate regression trees) is available in ESA's Electronic Data Archive: *Ecological Archives* E083-017-S1.