Ecology, 81(11), 2000, pp. 3178–3192

CLASSIFICATION AND REGRESSION TREES: A POWERFUL YET SIMPLE TECHNIQUE FOR ECOLOGICAL DATA ANALYSIS

GLENN DE'ATH[1] AND KATHARINA E. FABRICIUS[2]

[1]*Tropical Environment Studies and Geography, James Cook University, Townsville, Qld 4811, Australia*
[2]*Australian Institute of Marine Science, P.M.B. 3, Townsville, Qld 4810, Australia*

Abstract. Classification and regression trees are ideally suited for the analysis of complex ecological data. For such data, we require flexible and robust analytical methods, which can deal with nonlinear relationships, high-order interactions, and missing values. Despite such difficulties, the methods should be simple to understand and give easily interpretable results. Trees explain variation of a single response variable by repeatedly splitting the data into more homogeneous groups, using combinations of explanatory variables that may be categorical and/or numeric. Each group is characterized by a typical value of the response variable, the number of observations in the group, and the values of the explanatory variables that define it. The tree is represented graphically, and this aids exploration and understanding.

Trees can be used for interactive exploration and for description and prediction of patterns and processes. Advantages of trees include: (1) the flexibility to handle a broad range of response types, including numeric, categorical, ratings, and survival data; (2) invariance to monotonic transformations of the explanatory variables; (3) ease and robustness of construction; (4) ease of interpretation; and (5) the ability to handle missing values in both response and explanatory variables. Thus, trees complement or represent an alternative to many traditional statistical techniques, including multiple regression, analysis of variance, logistic regression, log-linear models, linear discriminant analysis, and survival models.

We use classification and regression trees to analyze survey data from the Australian central Great Barrier Reef, comprising abundances of soft coral taxa (Cnidaria: Octocorallia) and physical and spatial environmental information. Regression tree analyses showed that dense aggregations, typically formed by three taxa, were restricted to distinct habitat types, each of which was defined by combinations of 3–4 environmental variables. The habitat definitions were consistent with known experimental findings on the nutrition of these taxa. When used separately, physical and spatial variables were similarly strong predictors of abundances and lost little in comparison with their joint use. The spatial variables are thus effective surrogates for the physical variables in this extensive reef complex, where information on the physical environment is often not available.

Finally, we compare the use of regression trees and linear models for the analysis of these data and show how linear models fail to find patterns uncovered by the trees.

Key words: analysis of variance; CART; classification tree; coral reef; Great Barrier Reef; habitat characteristic; Octocorallia; regression tree; soft coral; surrogate.

INTRODUCTION

Ecological data are often complex, unbalanced, and contain missing values. Relationships between variables may be strongly nonlinear and involve high-order interactions. The commonly used exploratory and statistical modeling techniques often fail to find meaningful ecological patterns from such data. Classification and regression trees (Breiman et al. 1984, Clark and Pregibon 1992, Ripley 1996) are modern statistical techniques ideally suited for both exploring and modeling such data, but have seldom been used in ecology (Staub et al. 1992, Baker 1993, Rejwan et al. 1999).

Trees explain variation of a single response variable by one or more explanatory variables. The response variable is usually either categorical (classification trees) or numeric (regression trees), and the explanatory variables can be categorical and/or numeric. The tree is constructed by repeatedly splitting the data, defined by a simple rule based on a single explanatory variable. At each split the data is partitioned into two mutually exclusive groups, each of which is as homogeneous as possible. The splitting procedure is then applied to each group separately. The objective is to partition the response into homogeneous groups, but also to keep the tree reasonably small. The size of a tree equals the number of final groups. Splitting is continued until an overlarge tree is grown, which is then pruned back to the desired size. Each group is typically characterized by either the distribution (categorical response) or mean value (numeric response) of the re-

Manuscript received 16 November 1998; revised 4 November 1999; accepted 6 November 1999.

sponse variable, group size, and the values of the explanatory variables that define it.

The way that explanatory variables are used to form splits depends on their type. For a categorical explanatory variable with two levels, only one split is possible, with each level defining a group. For categorical variables with >2 levels, any combinations of levels can be used to form a split, and for k levels, there are $2^{k-1} - 1$ possible splits. For numeric explanatory variables, a split is defined by values less than, and greater than, some chosen value. Thus, only the rank order of numeric variables determines a split, and for u unique values there are $u - 1$ possible splits. From all possible splits of all explanatory variables, we select the one that maximizes the homogeneity of the two resulting groups. Homogeneity can be defined in many ways, with the choice depending on the type of response variable.

Trees are represented graphically, with the root node, which represents the undivided data, at the top, and the branches and leaves (each leaf represents one of the final groups) beneath. Additional information can be displayed on the tree, e.g., summary statistics of nodes, or distributional plots.

We will show how trees can deal with complex ecological data sets using soft coral (Cnidaria: Octocorallia) survey data from the Australian central Great Barrier Reef. This, together with a detailed exposition of trees, follows this introduction. First, we describe the soft coral survey data, and ecological issues that we investigate with trees. We then illustrate the basics of classification and regression trees with two analyses of a soft coral species. A more detailed discussion of trees follows, and includes: (1) exploration, description, and prediction of data; (2) technical aspects of growing trees with different splitting criteria; (3) pruning trees to size by cross-validation; and (4) data transformations, missing values, and tree diagnostics. Tree analyses of the soft coral data then address the following ecological issues: (1) relationships between physical and spatial environmental variables, (2) habitat characteristics associated with aggregations of three soft coral taxa, and (3) comparison of physical and spatial variables as predictors of soft coral abundance. Finally, we compare the performance of trees to equivalent linear model analyses.

The Soft Coral Study

Soft corals (class Anthozoa, Octocorallia: Order Alcyonacea) occur in high abundances on many types of coral reefs. They can numerically dominate reefs in turbid nearshore regions, as well as in clear water reefs away from coastal influences (Dinesen 1983, Fabricius 1997). Abundances of soft corals are strongly related to their physical environment (Fabricius and De'ath 1997), but their role in reef communities is not well understood.

We analyze three groups of taxa: (1) *Efflatounaria* (family Xeniidae) comprises three species (Gohar 1939, Versefeldt 1977) which are not reliably distinguished; (2) *Sinularia* spp. (family Alcyoniidae; Versefeldt 1980) comprises five ill-defined species with very similar morphology and distribution, and includes *S. capitalis* and *S. polydactyla*; and (3) the distinct species *Sinularia flexibilis* (Versefeldt 1980).

Efflatounaria is locally dominant in clear offshore waters, whereas *Sinularia* spp. and *S. flexibilis* are highly abundant and conspicuous nearshore taxa. They can form dense aggregations, to the extent of monopolizing space and excluding the reef-building hard corals on the scale of thousands of square meters (Fabricius 1998).

Additionally, we use *Asterospicularia laurae* (family Asterospiculariidae; Utinomi 1951) as an example of an uncommon species. It is one of the few soft corals that are reliably identified to species level in the field.

Zonations along the gradients of depth and distance to land have been extensively used to explain patterns in abundances of individual taxa on the Great Barrier Reef (Done 1982, Dinesen 1983). However, spatial variables are only proxies for a range of physical environmental variables, with which they are highly correlated. The relationships between spatial and physical variables are often complex. Hence the question of which physical variables determine the distribution of a taxon often remains unresolved—a question we attempt to address in this study.

Data comprising abundances of 38 genera of soft coral, four physical and five spatial variables (Table 1), were collected during surveys of 374 sites at 92 locations on 32 reefs within the Australian central Great Barrier Reef (Table 1 and Fig. 1). Each site was visually surveyed by one experienced observer (K. Fabricius), by scuba diving over typically 300–500 m (≈ 900–2000 m^2), for 15–20 min, within each of five defined depth ranges. The distribution of sites was highly unbalanced with respect to their defining characteristics (Table 2).

Examples of Classification and Regression Trees

A regression tree example

As an illustrative example, we use the ratings of abundances (row 1 of Table 1) of *Asterospicularia laurae* as the numeric response variable (Fig. 2a). The species is uncommon, occurring on 15% of sites (mean rating = 0.241, n = 373) with <1% of sites having a rating >2. The explanatory variables used in the model are cross-shelf position, location, and depth; all of which appear in the tree. Splits minimize the sums of squares (ss) within groups. The first split is based on shelf position, with inner- and mid-shelf reefs in the left branch, and outer-shelf reefs in the right branch. The left node is strongly homogeneous, and is not subsequently divided, forming a leaf with mean rating of

TABLE 1. Description of the variables used in the soft coral study. The character of variables is denoted by B = biotic, P = physical, or S = spatial; and the type by N = numeric or C = categorical.

Variable	Character	Type	Values
Abundances of soft corals (38 taxa)	B	N	0 (absent), 1 (few), 2 (uncommon), 3 (common), 4 (abundant), 5 (dominant)
Sediment	P	C	0 (none), 1 (thin), 2 (moderate), 3 (thick)
Visibility	P	N	1–33 m
Wave action	P	C	0 (none), 1 (moderate), 2 (strong)
Slope angle	P	N	0–90° in 5° increments
Cross-shelf position	S	C	Inner, mid-, or outer shelf
Reef type	S	C	Fringing around islands or platform
Within-reef location	S	C	Front, back, channel, flank
Depth zone	S	C	0–1, 1–3, 3–8, 8–13, 13–18 m
Reef identity	S	C	32 levels

Note: Four of the 38 soft coral taxa were analyzed in terms of the physical and spatial variables.

0.038. For regression trees, the proportion of the total sum of squares explained by each split is important information, and could be displayed on the tree. However, we can also represent this graphically by the relative lengths of the vertical lines associated with each split (Fig. 2a); a practice we use for all trees in this paper. Continuing with the right branch comprising all outer reefs, it is now divided into back and flank reef locations to the left and front locations to the right; there are no channel sites on outer-shelf reefs. The splitting process is repeated, separating front-reef sites of depths above and below 3 m, which completes the tree with four leaves, and 49.2% of the total sum of squares is explained. The bar charts at each leaf show the distribution of observed ratings (0–3). They show *A. laurae* to be relatively common on the fronts of outer-shelf reefs, particularly at depths ≥3 m, but virtually absent from inner- and mid-shelf reefs.

A classification tree example

For the example in Fig. 2b, the response variable is the presence–absence of *A. laurae*. In this case, the tree is identical in structure to the regression tree, but the splits have relatively different strengths, as represented by their vertical lengths. Splits are based on the proportions of presences and absences in the groups. The leaves of the tree are characterized by their dominant category (present or absent), and the proportion of sites of that category; e.g., for the leftmost leaf, *A. laurae* is absent on 97% of inner and mid-shelf reefs (n = 263). When the response has more than two categories, leaves are characterized by their dominant category and the proportions in each category. A classification tree, treating the ratings of *A. laurae* as four distinct categories (not shown), gave identical leaves to the presence–absence tree, but had a stronger split for depth.

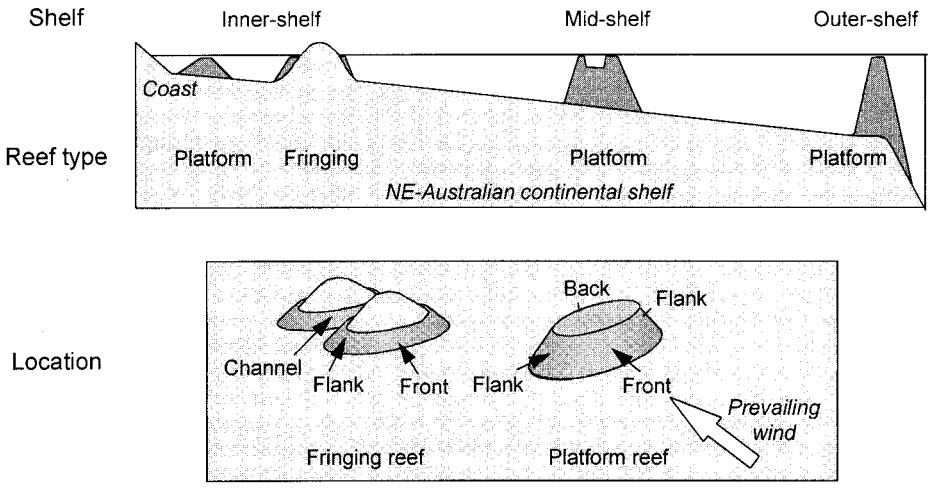

FIG. 1. Schematic representation of cross-shelf position (inner, mid, and outer), reef type (platform or fringing), and types of site location on the reef (front, back, flank, and channel). Fringing reefs form around islands, whereas platform reefs rise from the sea floor. In the survey area fringing reefs occurred only on the inner shelf. Fronts of reefs face the prevailing wind, and backs are on the leeward side, the two being joined by flanks. Channel sites occur on fringing reefs between closely located islands and typically have high currents.

TABLE 2. Spatial distribution of the survey sites in terms of cross-shelf position (inner, mid, and outer), reef type (fringing or platform), reef location (Ch = channel, Bck = back, Flnk = flank, and Frnt = front), and depth (m).

Depth (m)	Shelf position and reef type											
	Inner-fringing				Inner-platform		Mid-platform			Outer-platform		
	Ch	Bck	Flnk	Frnt	Bck	Frnt	Bck	Flnk	Frnt	Bck	Flnk	Frnt
0–1	5	7	1	4	3	3	2	1	3	6	0	2
1–3	19	17	3	8	3	4	9	2	8	13	1	7
3–8	18	20	2	12	3	4	9	2	10	16	1	10
8–13	3	12	0	8	4	5	6	2	9	14	1	12
13–18	4	7	0	6	1	3	3	1	7	11	1	15

Notes: Due to the varying structure of reefs and survey constraints, the distribution of sites with respect to these four spatial characteristics is highly unbalanced. In particular: (1) fringing reefs, and thus channel sites, occur only on the inner shelf, (2) flanks are underrepresented, (3) depths of 1–8 m are overrepresented on inner fringing reefs, and (4) depths of >8 m are overrepresented on outer reefs.

The proportion of total SS explained by the tree is useful for summarizing regression trees, but, for classification trees, misclassification rates are used. For the whole tree, 34 out of 373 cases were misclassified, giving an error rate of 9.1%. This compares with 50% for ''blind guessing,'' and 15% when we use the ''go with the majority rule,'' in this case predict *A. laurae* to be absent on all sites.

TREES: WHAT THEY CAN DO AND HOW THEY DO IT

Exploration, description, and prediction of data

It is useful to think of statistical analyses as exploration and modeling of data, with the former often preceding the latter. Exploration comprises both graphical and numerical techniques, and trees, which have both of these characteristics, are an effective way to explore

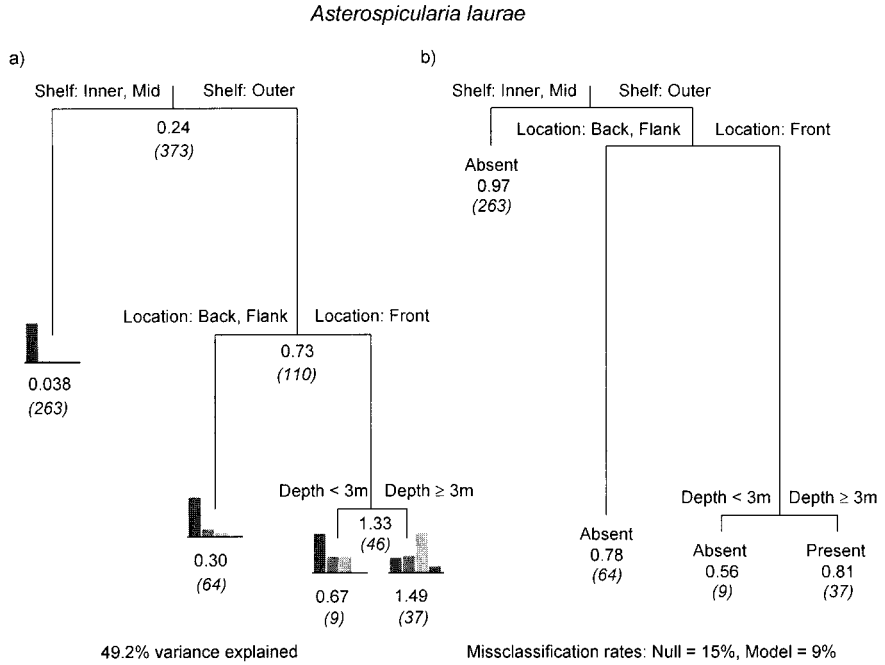

FIG. 2. (a) Regression tree analysis of the abundance of the soft coral species *Asterospicularia laurae* rated on a 0–5 scale; only values 0–3 were observed. The explanatory variables were shelf position (inner, mid, and outer), site location (back, flank, front, and channel), and depth (m). Each of the three splits (nonterminal nodes) is labeled with the variable and its values that determine the split. For each of the four leaves (terminal nodes), the distribution of the observed values of *A. laurae* is shown in a histogram. Each node is labeled with the mean rating and number of observations in the group (italic, in parentheses). *A. laurae* is least abundant on inner- and mid-reefs (mean rating = 0.038) and most abundant on front outer-reefs at depths ≥3 m (1.49). The tree explained 49.2% of the total SS, and the vertical depth of each split is proportional to the variation explained. (b) Classification tree on the presence–absence of *A. laurae*. Each leaf is labeled (classified) according to whether *A. laurae* is predominantly present or absent, the percentage of observations in that class, and the number of observations in the group (italic, in parentheses). The misclassification rate of the model was 9%, compared to 15% for the null model (guessing with the majority, in this case the 85% of absences).

complex relationships in data. Modeling often has one of two objectives: (1) description, in the sense that a model represents the systematic structure of the data as simply as possible, and (2) prediction, in that a model accurately predicts unobserved data. Trees can be used for both description and prediction.

Modeling requires the selection of one or more likely models from a range of possible alternatives, e.g., selecting which variables to include in a linear regression model. The methods available for model selection are numerous (Ripley 1996, Anderson and Burnham 1998), and depend on whether the objective is description or prediction.

Descriptive models can be selected by the two following methods:

1) Iterative comparison of nested models using hypothesis tests; e.g., forward, backward, and stepwise regression. Though a widely used method, the use of hypothesis tests for model selection has been increasingly questioned (Berger and Sellke 1987, Draper 1995, Stewart-Oaten 1996, Johnson 1999). For example, linear regression models selected by iterative hypothesis tests consistently include too many explanatory variables (Draper 1995).

2) Selection of the model with lowest true error (Breiman et al. 1984); also known as prediction error (Efron and Tibshirani 1993, Ripley 1996). For least squares regression, the prediction error of a model is defined as the expected squared error for new predictions, and for classification as the probability of misclassifying new observations. Obtaining realistic estimates of prediction error can be difficult. For example, the mean square error of least squares models (the resubstitution estimate of prediction error) gives overoptimistic estimates of error; an optimism that increases with model complexity. To counter this problem, we can either: (a) add a penalty for complexity; selection criteria such as Akaike's Information Criteria (AIC) (Sakamoto et al. 1986) and Bayes Information Criteria (BIC) (Schwarz 1978) adopt this approach; or (b) obtain more accurate estimates of the prediction error; cross-validation (Ripley 1996) and bootstrapping (Efron and Tibshirani 1993) are widely used for this approach.

Predictive models are selected on their accuracy of predictions for new data. Although we can select a single best model, predictive accuracy can be improved by averaging weighted predictions over several models, with the weights being the plausibility of the individual models (Draper 1995, Ripley 1996). Cross-validation (Breiman et al. 1984, Ripley 1996) is a widely used technique for selecting predictive models.

Growing trees and splitting criteria

The homogeneity of nodes is defined by impurity, a measure which takes the value zero for completely homogeneous nodes, and increases as homogeneity decreases. Thus maximizing the homogeneity of the groups is equivalent to minimizing their impurity. Many measures of impurity (splitting criteria) exist, and enable us to analyze many types of responses. There are five commonly used measures (Breiman et al. 1984); three for classification trees and two for regression trees.

For classification trees, impurity is defined in terms the proportions, c, of responses in each category. The three common criteria (indices) are:

1) The information (entropy) index takes the form $-\Sigma\ c\ \ln(c)$, where Σ indicates summation over categories. This index is identical to the Shannon-Weiner diversity index, and forms groups by minimizing the within-group diversity.

2) The Gini index takes the form $1 - \Sigma\ c^2$. At each split, the Gini index tends to split off the largest category into a separate group, whereas the information index tends to form groups comprising more than one category in the early splits.

3) The twoing index can be used for more than two categories. It defines two ''super categories'' at each split, for which the impurity is defined by the Gini index. It can also be used for ordered categories.

For regression trees, the two common forms of impurity are:

1) Sums of squares about the group means. This is equivalent to least squares linear models.

2) Sums of absolute deviations about the median. This gives a robust tree (Breiman et al. 1984). However, for ecological data dominated by zeros, this criterion can be ineffective, especially when the explanatory variables are categorical. In such cases all possible splits may result in groups with zero medians, and no splits will be formed.

Although the above measures are most commonly used, others include: (1) rank statistics, which result in invariance of the tree to any monotonic transformation of the response variable (Segal 1988), and (2) survival statistics for censored data (Segal 1988; T. Therneau, *unpublished manuscript*).

Trees can also be formulated as statistical models, akin to linear, generalized linear and generalized additive models (Clark and Pregibon 1992). In this approach, splits are based on an explicit statistical model, the deviance of which defines the dissimilarity measure. For classification trees, they use a multinomial model, equivalent to the information index, with the deviance defined by the multinomial log-likelihood. For regression trees, Clark and Pregibon (1992) use the Gaussian model, and the deviance for a node is simply the sums of squares about the mean. Summing over all leaves gives the overall deviance for the tree.

Pruning trees

A natural way of using a splitting criterion to grow a tree is to continue splitting until the improvement due to additional splits is less than a prespecified cutoff, and then take this as the best tree. The foundational

work of Breiman et al. (1984) points out two weaknesses of this approach. First, if the stopping rule is based on too small an improvement, then an overlarge tree will result. Second, if the criterion is too large, then splits based on interactions between explanatory variables will not be discovered unless at least one of the associated main effects is large enough to generate a split.

Breiman et al. (1984) introduce three basic ideas to solve the problem of finding the best tree. The first idea is tree pruning: rather than stop growth in progress, they grow an over-large tree and then seek ways to cut it back. This can be computationally infeasible, since the number of sub-trees is usually very large. To overcome this problem, their second idea is to find a sequence of nested trees of decreasing size, each of which is the best of all trees of its size. For this they use the resubstitution estimate of error, $R(T)$, which can be either the overall misclassification rate or the total residual SS, dependent on the type of tree. They show that, for any number α (≥ 0) there is a unique smallest tree that minimizes $R(T) + \alpha|T|$, where $|T|$ is tree size (number of leaves). By allowing α to increase from 0 to large, we obtain the desired sequence of nested trees of decreasing size, beginning with the initial overlarge tree and ending with the root tree with no splits at all. Since each tree in this sequence is the best of its size, choosing the best tree is reduced to the task of choosing the best size, a much simpler task than comparing all possible subtrees. $R(T)$ is not suitable for this choice because it will always be minimized by the largest tree (just as adding more explanatory variables reduces the residual SS of a regression). Thus, to complete the process, we require better estimates of error, and the third idea of Breiman et al. (1984) is to obtain ''honest'' estimates of error by cross-validation, as described in *Selecting tree size by cross-validation.* This can be computationally demanding, but is now feasible since we only have to consider one tree of each size, i.e., the trees of the nested sequence.

Selecting tree size by cross-validation

We have noted that descriptive models can be based on: (1) iterative comparisons between models, (2) model fit with a penalty for complexity, or (3) accurate estimates of prediction error. Selection of trees by either iterative comparisons or penalized complexity is ineffective (Clark and Pregibon 1992, Venables and Ripley 1999), and thus we require a method that accurately estimates prediction error.

Breiman et al. (1984) use cross-validation to obtain honest estimates of true (prediction) error for trees of a given size. For the sequence of trees, these estimates of error can be plotted against tree size, and the size with the minimum error selected. A single tree selected by cross-validation can be used for description and/or prediction. It should be interpreted as the tree which has the smallest estimated error and is the best estimated predictive single tree.

Cross-validation can be implemented in two ways. First, if enough data are available, we select a random subset of the data, typically comprising one-half to two-thirds of all data, and, using only this data, build the sequence of nested trees. For each tree, predict the response of the remaining data, and calculate the error from the predictions and the observed values. The tree with the smallest predicted error is then selected. One drawback of this technique is that there are often insufficient data to build good trees using only a subset of the data. The second way is to use V-fold cross-validation as follows: (1) divide the data into a number, V, of mutually exclusive subsets (typically $V = 10$) of approximately equal size; (2) drop out each subset in turn, build a tree using data from the remaining subsets, and use it to predict the responses for the omitted subset; (3) calculate the estimated error for each subset (e.g., for a sums of squares regression tree, the error is the sum of squared differences of the observations and predictions), and sum over all subsets; (4) repeat steps (2)–(3) for each size of tree; and (5) select the tree with the smallest estimated error rate. The subsets can be chosen randomly, but stratification into groups according to the value of the response variable gives smaller and more accurate estimates of the true error rate (Breiman et al. 1984).

When we have subsampling in the data, e.g., when we make repeated observations on the sampling units, then subsamples within units are usually correlated. V-fold cross-validation needs to be modified for such data, otherwise: (1) predictions of error rates are overoptimistic (Fig. 3), due to the fact that we would be using observations within a sampling unit to predict other values of the same unit; and (2) tree sizes tend to be overestimated (Fig. 3). We can overcome this problem by selecting only whole sampling units in our subsets, and thus units are predicted from other units. If the number of sampling units is large, each subset will contain several complete units, but if the number is $\leq V$, then each subset will comprise a single complete unit. Bootstrap procedures (Efron and Tibshirani 1993) have been modified in this way, but to our knowledge, cross-validation procedures have not.

For the nested sequence of trees, the difference between resubstitution and cross-validated estimates of error often increases with tree size (Fig. 3), with the latter often rapidly dropping to a minimum followed by a slow increase. If this plateau effect occurs, the tree size corresponding to the minimum error is imprecisely estimated.

Breiman et al. (1984) suggested the 1-SE rule whereby the best tree is taken as the smallest tree such that its estimated error rate is within one standard error of the minimum. The standard error of the estimate (Breiman et al. 1984) can be calculated for each tree size (Fig. 3). Use of the 1-SE rule can result in a much

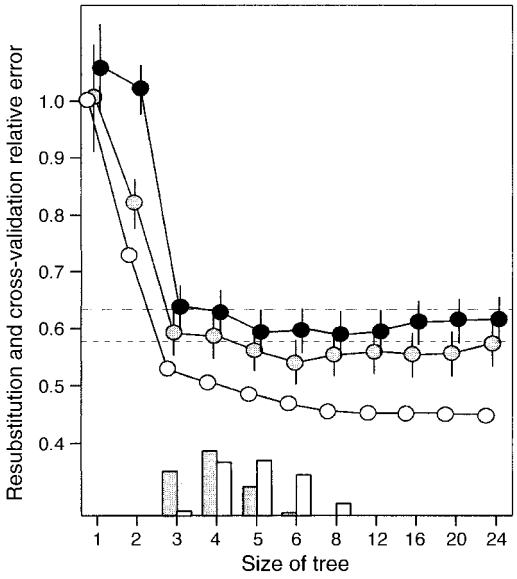

FIG. 3. Resubstitution (white circles) and cross-validation relative error (gray circles, unstratified; black circles, stratified by reef) for the regression tree of *Asterospicularia laurae* (Fig. 2). The cross-validation relative error plots are for a single 10-fold cross-validation and include 1-SE estimates for each tree size. The resubstitution error decreases monotonically with tree size, as will always be the case, reaching a minimum of 0.45 for the largest possible tree of size 24. Under unstratified cross-validation, a tree of size five is selected by the 1-SE rule, whereas for cross-validation stratified by reef, a tree of size four is selected. At the bottom of the plot, the bar chart shows the relative proportions of trees of each size selected under the 1-SE rule (gray) and minimum (white) rules from a series of 50 cross-validations. The modal tree (most likely) under the 1-SE rule has four leaves, and under the minimum rule a five-leaf tree is marginally more likely than the four-leaf tree.

smaller tree than suggested by the minimum cross-validated-error, but with minimal increase in the estimated error rate (at most <1 SE). Irrespective of whether the minimum or 1 SE rule is used, inspection of the cross-validated sequence is necessary to ensure that the sequence of trees has been grown large enough. For both the minimum and 1-SE rule, the size of the selected tree will vary under repeated cross-validation (Fig. 3), and it is advisable to run several cross-validations in order to assess the degree of variation in the sizes of the best tree, and ensure the chosen tree is not atypical.

If a single tree is required for description, and the data set is not overlarge, then either of the following strategies could be used. Each is based on a series of cross-validations, comprising perhaps 50 sets of 10-fold cross-validations. For the first strategy: (1) select the tree size from each cross-validation of the series according to either the minimum or 1-SE rule; and (2) from the distribution of selected tree sizes, select the most frequently occurring (modal) size. For the second, the estimated errors and their standard errors can be averaged over the cross-validations. This gives a smoother cross-validation curve from which the tree size can be selected according to either rule. We prefer the latter strategy, but it is difficult to implement for the available software.

If the objective of the analysis is prediction, then averaging over a collection of trees gives better performance (Ripley 1996). This can be done using these steps: (1) run a series of cross-validations, (2) select a tree from each cross-validation according to the minimum rule, (3) for each tree predict for the new data, and (4) average the predictions over the trees.

Transformations

For a numeric explanatory variable, x, only its rank order determines a split, i.e., the groups are defined by $x > x_s$ and $x < x_s$, where x_s maximizes their homogeneity. Trees are thus invariant to monotonic transformations of numeric explanatory variables. Hence, we do not have to deal with the difficult issue of the form of relationship between the response and numeric explanatory variables (e.g., linear–linear, log–log, linear–polynomial). However, for a strong linear (or smooth) relationship, regression trees will not perform as well as linear regression (or non- or semiparametric smoothers). For some splitting criteria, e.g., those based on ranks, trees are also invariant to monotonic transformations of the response variable (Breiman et al. 1984, Segal 1988). For regression trees based on sums of squares or sums of absolute deviations, residual plots are useful to check for outliers, and to show how variation of the response is related to mean values of the groups. Nonconstant variation gives greater weight to data with higher variation, and therefore it is often desirable to transform the response variable.

Missing values and surrogate variables

Missing data are a common occurrence in ecological studies, and for many types of models they can be problematic (Little and Rubin 1987). Cases with missing responses, or with missing explanatory variables are often deleted. This results in loss of information and possible bias. Alternatively, missing data can be replaced by some form of imputation, e.g., replacement by the mean value, but this too can result in bias. Trees handle missing data in a variety of ways. For regression trees, cases with missing responses are deleted, but for classification trees, the missing responses can be treated as a special category (Clark and Pregibon 1992), thereby providing information on response bias. If an explanatory variable used to form a split has missing values, we can either stop cases with missing values at the split, and give them the characteristic response value of all cases that pass through that node; or send them further down the tree, using an explanatory variable with nonmissing values for these cases. We choose the variable that best agrees with the original splitting variable, and quantify its performance by the percentage agreement in the allocation of cases to the

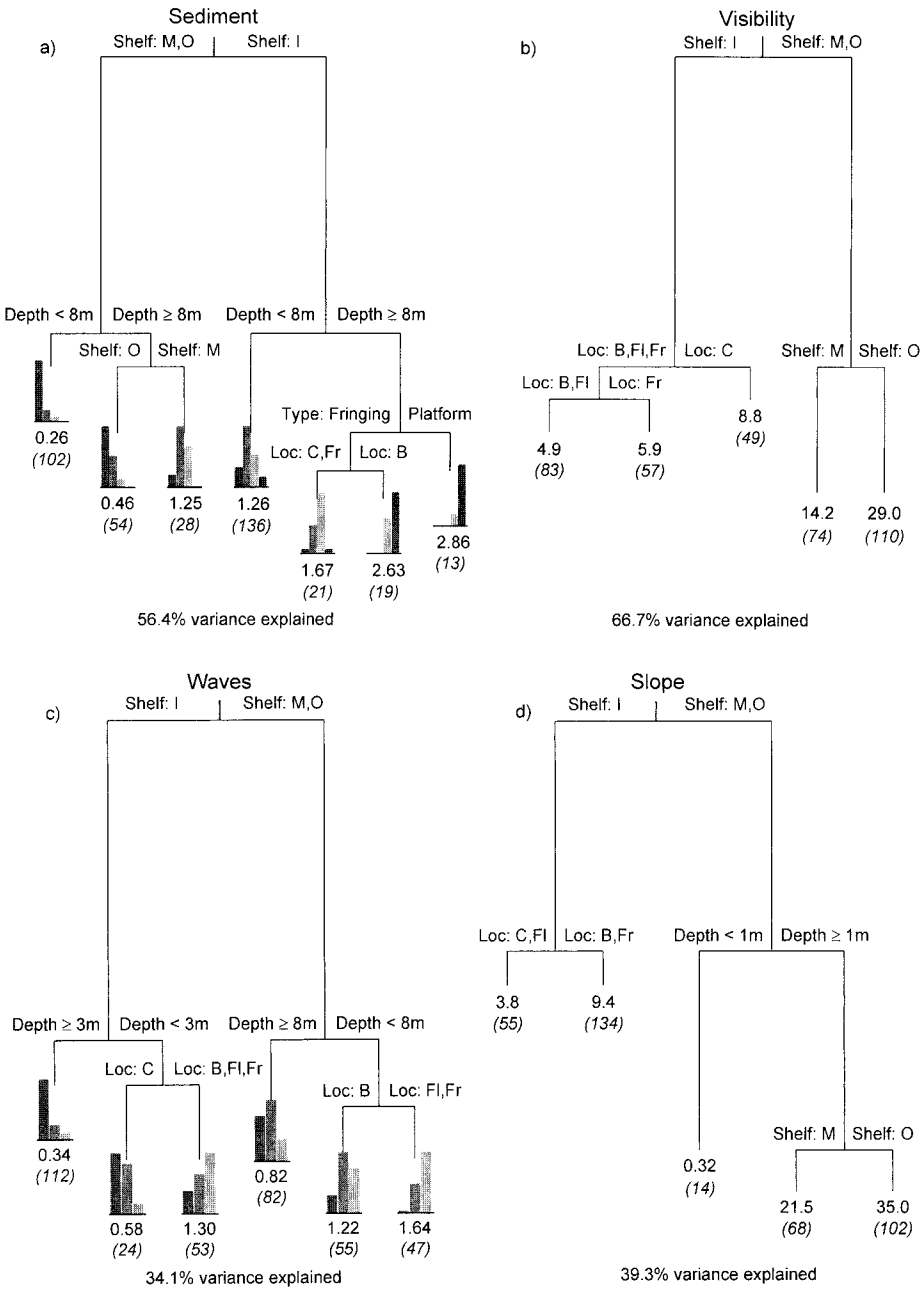

FIG. 4. Regression trees relating the distributions of the four physical variables (sediment, visibility, waves, and slope) to the four spatial variables (shelf position, location, reef type, and depth). The figures are labeled as described in Fig. 2. The regression trees for sediment and waves were similar, explaining 56.4% and 34.1% of the total SS, respectively. Visibility was the best-explained physical variable, followed by sediment, and finally, slope and waves. For all four physical variables, the strongest effects were due to cross-shelf position, followed by depth. The effects of location and reef type were relatively small.

two groups. The use of such variables, known as surrogate variables (Breiman et al. 1984), minimizes the information loss, which can render the case deletion approach unworkable when a large proportion of data are missing.

Alternative splits

The use of alternative splits can be used both as an exploratory technique, and also to generate alternative trees. For each split, we can compare the strength of the split due to the selected variable with the best splits of each of the remaining explanatory variables. A strongly competing alternative variable can be substituted for the original variable, and this can sometimes simplify a tree by reducing the number of explanatory variables, or, as we show later, lead to a better tree. Often, though not always, the best surrogate variable

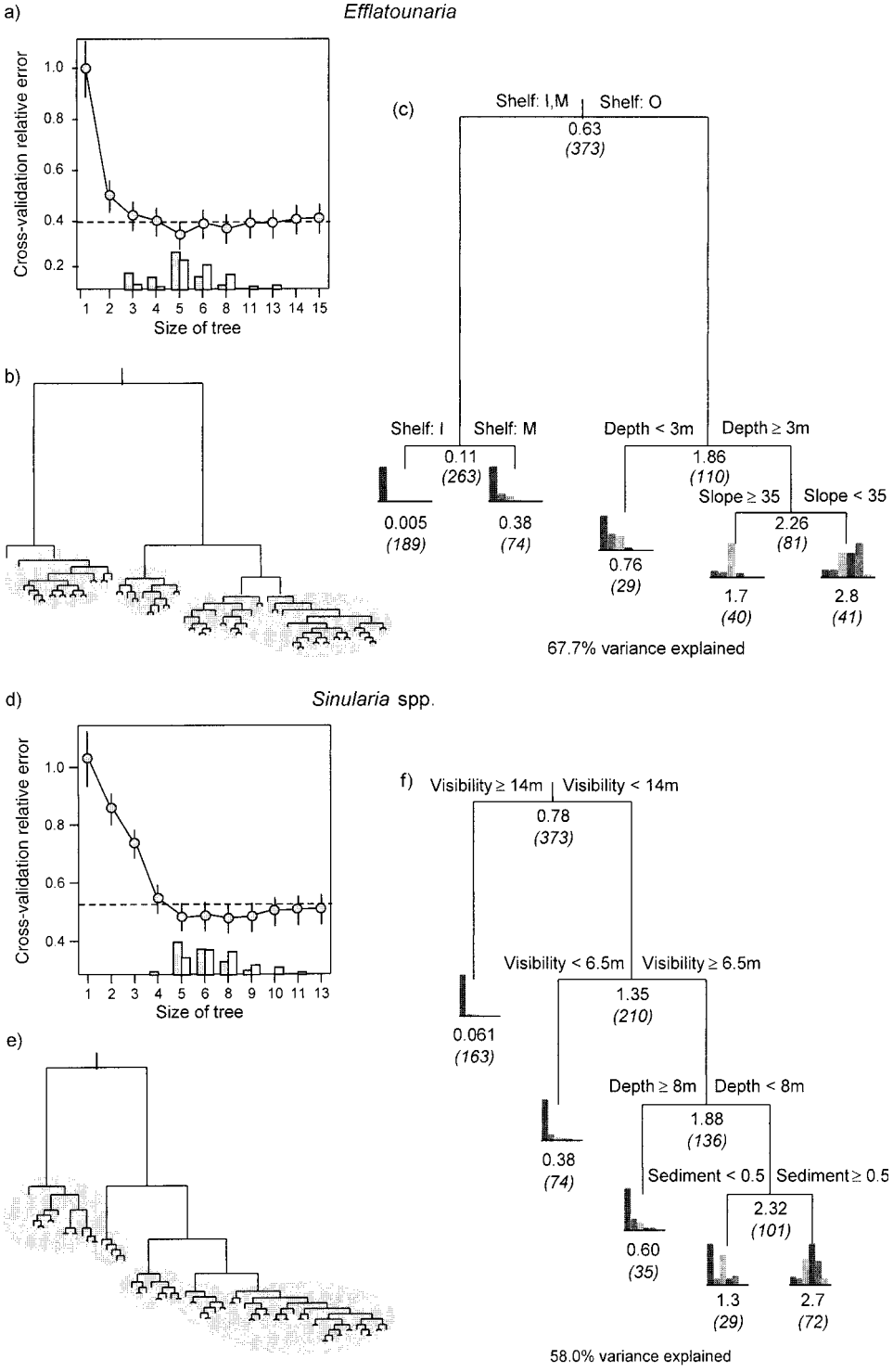

FIG. 5. Regression trees explaining the abundances of the soft coral taxa *Efflatounaria*, *Sinularia* spp., and *Sinularia flexibilis* in terms of the four spatial variables (shelf position, location, reef type, and depth) and four physical variables (sediment, visibility, waves, and slope). At the bottom of the cross-validation plots (a, d, g), the bar charts show the relative proportions of trees of each size selected under the 1-SE rule (gray) and minimum rules (white) from a series of 50 cross-validations. For *Efflatounaria* (a), a five-leaf tree is most likely by either the 1-SE or the minimum rule. For *Sinularia* spp. (d), five to eight-leaf trees have support, and for *S. flexibilis* (g), five- to nine-leaf trees have support. Cross-validation plots (a, d, g), representative of the modal choice for each taxa according to the 1-SE rule, are also shown. For all three taxa, a five-leaf tree was selected (c, f, i). The shaded ellipses enclose nodes pruned from the full trees (b, e, h), each of which accounted for >99% of the total SS.

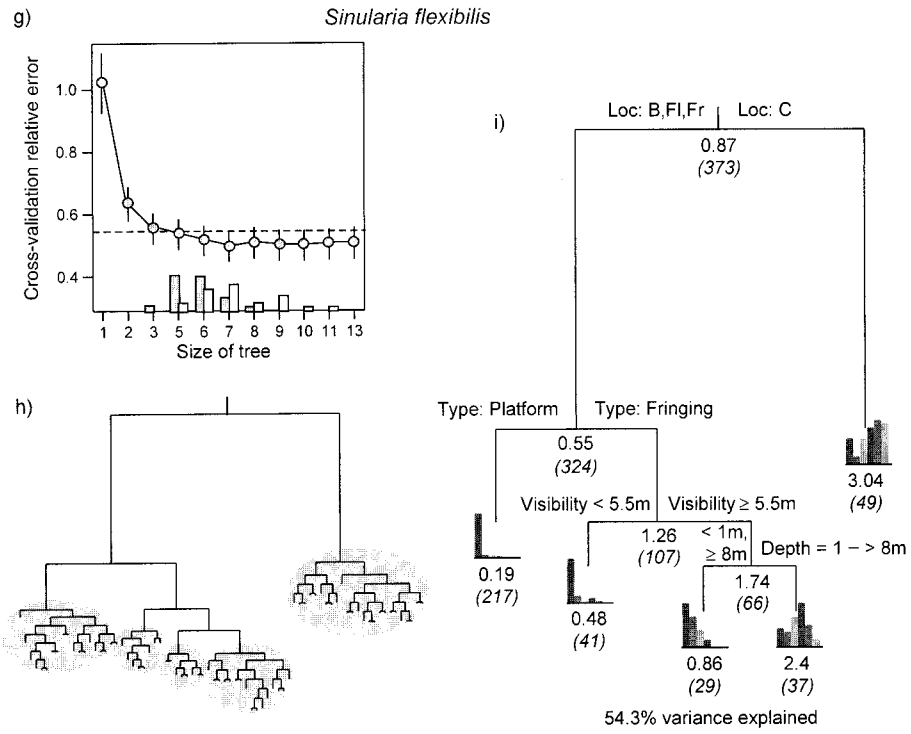

FIG. 5. Continued. For *Efflatounaria* the pruned nodes represent 33.3% of the total SS and show little structure, as can be seen from the vertical lengths that are proportional to the SS explained by each split. For *Sinularia* spp. and for *S. flexibilis*, the pruned nodes represent a greater loss of SS explained (42.0% and 45.7% respectively). Also some evidence of further useful splits is evident in the full trees, for the left and right branches, respectively. The habitats with highest levels of *Efflatounaria* were deep sites (≥3 m) of moderate slope (< 35°) on outer reefs (mean rating = 2.8). For *Sinularia* spp., moderate visibility (6.5–14 m) sites at moderate depth (<8 m) with at least some sediment were favored, and for *S. flexibilis*, channel sites had the highest levels of abundance.

will also give the best alternative split. Inspection of alternative splits and surrogate variables can lead to a more complete understanding of dependencies and relationships within the data.

STATISTICAL ANALYSES

All data analyses presented in this paper used S-Plus statistical software (Statistical Sciences, 1999). The library of tree routines (RPART: Recursive PARTitioning) developed by T. Therneau (*unpublished manuscript*) was used for all classification and regression tree analyses. For all tree analyses, models were selected by cross-validation. For each tree, we ran a series of 50 10-fold cross-validations and chose the most frequently occurring tree size using the 1-SE rule (see *Selecting tree size by cross-validation* above). Typical cross-validations are shown in the results, each selecting the modal-size tree.

For all tree analyses, reefs were the sampling units, and hence reef identity was not used in the same way as other spatial and environmental variables. Reefs were not included in the models as explanatory variables, but were used to form the subsets for the cross-validations (see *Selecting tree size by cross-validation* above).

ANALYSIS OF THE SOFT CORAL DATA USING TREES

Exploring relationships between physical and spatial variables

Classification and regression trees were used to analyze the relationships between each of the four physical variables and the four spatial variables. The responses of sediment and wave action were analyzed using both types of tree. For both responses, the classification and regression trees were highly similar, suggesting the categories were behaving as numerical variables. Only the regression trees are presented in the results (Fig. 4a and c). The data for visibility and slopes were fourth-root and square-root transformed, respectively, to stabilize the variance at the leaves. The means of these variables shown on the trees (Fig. 4b and d) were back transformed to assist interpretation.

Visibility and sediment levels were best explained (R^2 = 66.7 and 56.4% respectively) by the spatial variables, whereas wave action and slope were less well explained (R^2 = 34.1 and 39.3%) (Fig. 4). This is possibly due to high temporal variation of waves and high local variation in slopes. Cross-shelf position, followed by depth, were consistently the strongest explanatory variables of the physical variables (Fig. 4).

Predicting abundances and habitat characteristics of taxa

We used regression trees to relate abundances of *Efflatounaria, Sinularia* spp., and *S. flexibilis* to the spatial and physical environmental variables. For each of these taxa, cross-validation using the 1-SE rule selected a five-leaf tree (Fig. 5a, d, and g). The choice for *Efflatounaria* was clear from both the cross-validation plot and the full tree (Fig. 5a and b), with the selected tree explaining 67.7% of the total SS. For *Sinularia* spp. and *Sinularia flexibilis* the evidence was less clear-cut, with trees from four to ten leaves receiving support (Fig. 5d and e). Five-leaf trees were selected for both taxa, which explained 58.2% of the total SS for *Sinularia* spp., and 55.4% for *Sinularia flexibilis.* The fact that a particular size of tree was not favored over all others for *Sinularia* spp. is not a weakness of the cross-validation procedure, but simply reflects the fact that several models were almost equally plausible. This situation also occurs for other types of statistical models, such as linear models, and neatly illustrates that blind adherence to any single model selection procedure, be it 1-SE for trees, or $P < 0.05$ for linear models, is not the most informative strategy. Both the cross-validation and the full tree plots are useful in assessing the certainty with which a particular model is chosen.

High abundances of all three taxa were restricted to a narrow set of environmental conditions. *Efflatounaria* occurred on only one of 189 inner-reef sites (Fig. 5c) and was relatively rare on mid-shelf sites (mean rating: 0.38, $n = 74$), compared with outer-shelf sites (1.86, $n = 110$). Abundance was highest (2.8, $n = 41$) on outer-shelf sites of ≥3 m water depth on terraces with slopes < 35°. *Sinularia* spp. was most abundant (mean rating = 2.7, $n = 72$) on sites of intermediate visibility (6.5–14 m), with depths <8 m, and at least some sediment (Fig. 5f). These are largely inner-shelf sites (Fig. 4b). *Sinularia flexibilis* (Fig. 5i) was most abundant in channels (mean rating = 3.04, $n = 49$). These are inner-shelf sites with gentle slopes (Fig. 4d) and low wave action (Fig. 4c). It also had high abundances (mean rating 2.4, $n = 37$) at intermediate depths (1–8 m) on fringing reefs where visibility was ≥5.5 m.

Surrogates and alternative splits were examined to better understand the relative effects of physical and spatial variables. Visibility was a strong surrogate and alternative split for shelf position and vice versa, in agreement with Fig. 4b. For example, dropping shelf from the *Efflatounaria* model resulted in a five-leaf tree with the same splits (Fig. 5c), but with visibility replacing shelf, and explained only marginally less of the total SS (65.4 vs. 67.7%). The surrogate and alternative splits relating other physical and spatial variables were weaker and less consistent.

Inspection of alternative splits led to a stronger tree for *Sinularia* spp.. The initial tree used reef type for the first split, however, visibility was a strong alternative split, and when reef type was dropped from the model, the resulting tree better explained the data (58.0 vs. 53.5%) for a same-size tree.

TABLE 3. Comparison of regression tree analyses of *Efflatounaria, Sinularia* spp., and *S. flexibilis* using either physical and spatial predictors (P + S), only physical predictors (P), or only spatial predictors (S).

Taxa	Predictors	Size of tree	Error (%)	Predicted error (%)
Efflatounaria	P + S	5	32	38
	S	4	37	44
	P	6	40	46
Sinularia spp.	P + S	5	46	55
	S	5	54	61
	P	5	48	58
S. flexibilis	P + S	5	45	61
	S	5	49	63
	P	6	44	61

Notes: The trees were selected using the modal tree from 50 cross validations and the 1-SE rule. The error is the percentage of unexplained variance, and the predicted error is estimated error rate of predictions standardized by the total variance. The use of only physical or only spatial predictors compared to joint use shows little difference for *Sinularia* spp. and *S. flexibilis*, and a moderate effect for *Efflatounaria.*

The analyses above clearly identified habitats of high abundance. They also showed how examination of surrogates and alternative splits can lead to a more complete understanding of competing explanatory variables and better (or simpler) models when explanatory variables with strong surrogates are dropped from the models.

Physical or spatial determinants of abundance?

In order to explore the degree to which the groups of physical and spatial variables separately explained abundance, the three taxa were also analyzed using regression trees based on each group of variables separately. The comparisons (Table 3) show that trees of size five were selected for most analyses, differing at most by one from this size. Spatial variables both explained and predicted *Efflatounaria* marginally better than the physical variables, whereas for *Sinularia* spp. and *S. flexibilis* the reverse was true. The degree to which either of the physical or the spatial data explained the abundances of all three taxa was on average only 4% less than the combined physical-spatial data.

The explanatory strength of the physical variables could be improved by additional physical data (e.g., light levels and currents), and by long-term measurements (e.g., waves were poorly estimated due to their high temporal variability). None were available due to the large-scale nature of the study. However, the finding of similar explanatory strength by the two groups of spatial and physical variables does suggest that the observed physical variables are possibly major determinants of soft coral abundances and that, in the absence of observations on the physical data, the more easily available spatial variables could be used as surrogates.

TABLE 4. Analysis of variance for the abundances of the soft coral species *A. laurae*.

Source	df	SS	MS	*F*	*P*
Shelf	2	37.8	18.9	22.8	<0.001
Reef(Shelf)	28	20.8	0.74	4.25	<0.001
Location	3	9.15	3.05	17.5	<0.001
Depth	4	0.82	0.20	1.17	0.323
Shelf × Location	4	11.5	2.87	16.5	<0.001
Shelf × Depth	8	1.98	0.25	1.42	0.189
Location × Depth	12	2.37	0.20	1.13	0.333
Shelf × Location × Depth	13	5.86	0.45	2.58	0.002
Residuals	298	52.0	0.17		

Notes: The fixed effects are shelf position, location of site, and depth; and reefs were treated as random effects. The sums of squares are sequential. The total treatment degrees of freedom are 46 (i.e., the number of non-empty cells minus 1).

We have noted that only visibility and shelf position were strong surrogates for each other, and that for many splits of the three physical-spatial trees, there were no strong surrogates or alternative splits. Coupled with the strong explanations given by the physical and spatial models separately, this suggests that, for other than visibility and shelf position, the splits determined by spatial factors are equivalent to complex combinations of physical factors, and vice versa.

COMPARISON OF TREES AND LINEAR MODELS

Example one: mixed effects analyses of variance

In this example we use *A. laurae* to show how regression trees can be used to interpret complex analysis of variance tables. Analysis of variance is widely used by ecologists, and, in the context of a well-designed balanced study, is an effective method for determining which factors affect a numeric response. However, as the number of explanatory variables and the complexity of the data increase (e.g., high-order interactions, lack of balance, empty cells), then analysis of variance and linear models become less effective. The inclusion of random effects also greatly adds to the complexity of the analysis, especially for unbalanced designs.

In the introductory example, we used a regression tree to explain the abundance of *A. laurae* (Fig. 2a) in terms of shelf position (inner, mid, outer), location (front, back, channel, flank), depth (five levels) and reefs (32 levels). The data were highly unbalanced since channel locations only occur on inner reefs, flank locations were underrepresented and inner reefs were overrepresented (Table 2). Thus of the 60 possible cells generated by the three fixed factors, only 47 occurred in the data. In the mixed effects analysis of variance, shelf, location, and depth were fixed effects, and reefs nested in shelf were random effects and used as the error term for shelf position. Sequential sums of squares were used (Table 4). The analysis shows strong effects of shelf, location, and shelf by location; a moderate interaction between shelf, location, and depth; and strong variation between reefs (Table 4).

Interpretation of the analysis is difficult due to the significant high-order interaction with many levels and empty cells. Tabulations of means broken down by shelf position, location, and depth, showed the strong shelf and location effects, but failed to clearly identify the three-way interaction. In comparison, the regression tree analysis led to a simple interpretation (Fig. 2a). The four-leaf tree explained 49.2% of the total SS compared to 55.4% explained by the full tree, or equivalently, the analysis of variance with reef effects omitted. The three splits were (1) inner- and mid-shelf reefs vs. outer-shelf reefs; (2) within outer-shelf reefs, flanks and back vs. front (there are no channel sites on outer-shelf reefs); and (3) within front locations on outer-shelf reefs, shallow (<3 m) vs. moderate to deep (≥3 m). Only the sites of split (3) had moderate or high abundances, with the 37 deep sites having highest abundances. The full tree (omitted) showed two strong splits, one moderate split, and an additional 20 minor splits. Each split represents a one-degree-of-freedom contrast among the 47 cell means. The 24 leaves of the full tree indicate there were only 24 unique cell means.

We assessed the information missed by the regression tree in two ways. First, we analyzed the residuals from the tree using the same analysis of variance model as for the full data set. Other than reef effects, no terms were significant ($P > 0.05$). Second, we note the variation not explained by the four-leaf tree, but explained by the full tree (or analysis of variance without reef effects) was nonsignificant ($F_{20,326} = 1.07$, $P = 0.620$). Thus, we conclude the three one-degree-of-freedom contrasts of the tree adequately explained all fixed effects of the analysis of variance.

By comparison with the analysis of variance, the regression tree was far simpler. The splits of the tree represent an optimum set of one-degree-of-freedom contrasts, as defined by the splitting process. The advantages of the tree analysis, in its simplicity and ease of interpretation are clear; advantages that increase as the number of explanatory variables and complexity increase.

Example two: linear regression

In this example, we compare the effectiveness of regression trees and linear regression models in ex-

TABLE 5. Linear regression analyses of *Efflatounaria.*

Source	Model 1				Model 2			
	df	SS	*F*	*P*	df	SS	*F*	*P*
Shelf	2	245.9	331	<0.001	2	245.9	312.7	<0.001
Depth	4	22.9	15.4	<0.001	1	20.3	50.9	<0.001
Slope	1	1.8	4.9	0.028	1	4.0	10.2	0.001
Shelf × Depth	8	38.1	12.8	<0.001	2	17.6	44.7	<0.001
Shelf × Slope	2	4.3	5.7	0.004	2	1.9	4.8	0.009
Depth × Slope	4	10.8	7.3	<0.001	1	9.0	23.1	<0.001
Shelf × Depth × Slope	7	15.8	6.1	<0.001	2	3.6	9.3	<0.001
Residuals	343	127.2			361	141.9		

Notes: For Model 1, shelf and depth are three- and five-level factors, respectively, and slope is numeric. For Model 2, shelf, depth, and slope are three-, two-, and two-level factors, respectively. Models were built by forward inclusion based on Mallow's *Cp,* and taking into account interaction hierarchies. Sequential sums of squares are displayed to assist comparison with the corresponding regression trees.

plaining the abundances of *Efflatounaria, Sinularia* spp., and *S. flexibilis* in terms of the physical and spatial variables. Given the number and type of explanatory variables in these data, which are not atypical, building linear models is difficult. We needed to include high-order interactions in our models, and account for the hierarchy of interactions. Models were constructed by forward inclusion, using $P < 0.01$ to be conservative. An interaction was considered for inclusion only when all lower order terms contained by it had been included in the model. Forward inclusion based on Mallow's *Cp* (Venables and Ripley 1999) was also used (results not shown) and gave the same final model for *Efflatounaria,* but slightly more complex models for *Sinularia* spp. and *S. flexibilis.* The effects of reef, which should be treated as random, were not included due to the complexity of determining appropriate error terms.

For *Efflatounaria* the final model (Model 1) comprised two factors, cross-shelf position and depth, crossed with the numeric variable reef slope, which resulted in 15 intercept and 14 slope parameters (one slope parameter was not estimable because one cell had constant reef slopes). Examination of parameter estimates failed to reveal patterns. One solution to the problem of interpretation was to group reef slope into two categories, reduce the number of categories for depth, and then cross tabulate the mean abundance by these reduced factors (Model 2 of Table 5, Table 6). The regression tree (Fig. 5c) was used to choose the split points: ≥35° for slope and ≥3 m for depth. This reduction was effective, and showed a similar pattern to the regression tree with high abundance in medium to deep sites (≥3 m) in outer reefs, particularly on flat sites (<35° slope). Comparing the reduced model (Model 2) against the full model (Model 1), showed that the simplification resulted in the loss of significant information ($F_{17,344} = 2.23$, $P = 0.003$, Table 5). Model 2, with 11 degrees of freedom (df), explained 68.2% of the total SS, almost identical to 67.7% for the tree with only five leaves (equivalent to 4 df). Fitting an eight-leaf tree, for which there is support under cross-validation (Fig. 5a), accounted for the depth effect within mid-shelf reefs (Table 6), and slope effects within mid-shelf reefs in the deep and outer-shelf reefs in the shallow (not shown). This tree accounted for 71.1% of the total SS compared to 72.6% for the full model, the difference being nonsignificant ($F_{21,344} = 0.950$, $P = 0.473$).

For *Sinularia* spp. and *S. flexibilis* the final linear models were extremely complex, and due to that complexity, uninterpretable. The model for *Sinularia* spp. involved six of the eight environmental variables, and five first-order interactions, and for *S. flexibilis* five of the eight environmental variables, and five first-order and one second-order interaction.

These analyses highlight some difficulties in using iterative selection (forward, backward, and stepwise methods) of linear models for complex data. The resulting models will often include many interactions between quantitative and categorical variables, which make interpretation difficult. Regression trees can help in the selection of a simpler regression model as was case with *Efflatounaria.* Given the problem of liberal inclusion of variables, when using iterative selection based on hypothesis tests (Draper 1995), and the demonstrated performance of regression trees, we suggest

TABLE 6. Mean ratings (*N*; 1 SE) of *Efflatounaria* abundance from Model 2 (see Table 5).

Position	Slope <35		Slope ≥35	
	Depth <3 m	Depth ≥3 m	Depth <3 m	Depth ≥3 m
Inner-shelf	0 (73; 0)	0.01 (103; 0.01)	0 (4; 0)	0 (9; 0)
Mid-shelf	0.04 (23; 0.04)	0.59 (39; 0.13)	0 (2; 0)	0.40 (10; 0.27)
Outer-shelf	0.52 (21; 0.16)	2.78 (41; 0.20)	1.38 (8; 0.38)	1.73 (40; 0.12)

the latter will often identify descriptive models for complex ecological data more effectively than linear regression models.

Discussion and Conclusions

We have outlined the fundamentals of classification and regression trees, and have shown how they can model complex ecological data. Trees were used to determine the environmental characteristics associated with high abundances of three taxa. The analyses indicated that, for these taxa, the ability to establish dense aggregations and monopolize space was restricted to clearly defined types of reef habitat.

Efflatounaria occurred mainly in deep clear water sites on the outer-shelf. These sites are also low in sediment, and have the lowest wave action of the outer-shelf. *Sinularia* spp. occupies mainly shallow inner-shelf habitats of relatively high visibility, low wave action, and high sediment levels. *Sinularia flexibilis* occurs on channel sites that have the highest visibility of the inner-shelf, are shallow, and have low wave action and moderate sediment levels. These findings are consistent with previous studies that indicate that nutrition plays an important role in determining abundances of soft corals.

The photosynthetic efficiency in zooxanthellate soft corals is generally low (Fabricius and Klumpp 1995). This has two important consequences for the distribution of soft coral taxa. Firstly, light levels are only high enough to saturate photosynthesis either in clear water, or at shallow depth in turbid water. Secondly, additional heterotrophic food intake is essential for supplementing the phototrophic carbon supply, and for providing nutrients. The two main food sources for soft corals are suspended particulate matter and dissolved nutrients. Members of the families Alcyoniidae (which includes *Sinularia*) and Nephtheidae are efficient suspension feeders that feed on small suspended particulate matter (Fabricius et al. 1995*a*, *b*, Fabricius and Domisse 2000). The highest concentrations of suspended particulate matter are found on the inner-shelf, due to import by rivers and resuspension from the shallow sea floor (Revelante and Gilmartin 1982). In contrast, members of the family Xeniidae (which includes *Efflatounaria*) efficiently take in dissolved nutrients (Schlichter 1982), concentrations of which can be high on the outer-shelf due to periodic upwelling of nutrient-enriched deep water bodies (Wolanski 1994). The habitats of *Sinularia* are thus rich in suspended particulate matter for suspension feeding, but also moderately good for photosynthesis due to their shallow depth. Conversely, habitats of *Efflatounaria* are rich in dissolved nutrients, good for photosynthesis due to the clear water of the outer-shelf, and protected from extreme wave action of the outer-shelf by being deep sites.

Surveys can, of course, only indicate relationships, not causality, and follow-up experiments have been established to determine reasons for the observed distributions of these three taxa, and their ability to establish dense aggregations.

Classification and regression trees are powerful tools for analysis of complex ecological data. Their features include: (1) the ability to use different types of response variables; (2) the capacity for interactive exploration, description, and prediction; (3) invariance to transformations of explanatory variables; (4) easy graphical interpretation of complex results involving interactions; (5) model selection by cross-validation; and (6) procedures for handling missing values. In summary, classification and regression trees are a valuable addition to the statistical toolbox of every ecologist and environmental scientist.

Software

Classification and regression trees are now available in some statistical packages. Three sources, which we have used, are listed here.

1. CART (1998) is specialized classification and regression tree software, and includes many advanced features of tree construction. It is powerful and easy to use.

2. S-Plus (Statistical Sciences 1999) is a high-level statistical programming language, which includes many tree routines. Though more difficult to use than CART, it has greater flexibility and enables users to develop their own applications.

3. SYSTAT (SPSS Inc. 1997) includes classification and regression tree routines and, though they are quite limited, they can serve as a useful introduction to the topic.

Other tree software is available but has not been examined by the authors.

Acknowledgments

This manuscript greatly benefited from substantial input from Allan Stewart-Oaten. Critical comments of Steve Delean, and two anonymous referees are also appreciated. 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, and the Australian Institute of Marine Science.

Literature Cited

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

Baker, F. A. 1993. Classification and regression tree analysis for assessing hazard of pine mortality caused by *Heterobasidion annosum.* Plant Disease **77**:(2)136–139.

Berger, J. O., and T. Sellke. 1987. Testing a point hypothesis: the irreconcilability of *P* values and evidence (with discussion). Journal of the American Statistical Association **82**:112–139.

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

CART. 1998. Salford Systems, San Diego, California, 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.
Dinesen, Z. D. 1983. Patterns in the distribution of soft corals across the central Great Barrier Reef. Coral Reefs **1**:229–236.
Done, T. J. 1982. Patterns in the distribution of coral communities across the central Great Barrier Reef. Coral Reefs **1**:95–107.
Draper, D. 1995. Assessment and propagation of model uncertainty (with discussion). Journal of the Royal Society series B **57**:45–97.
Efron, B., and R. J. Tibshirani. 1993. An introduction to the bootstrap. Chapman and Hall, New York, New York, USA.
Fabricius, K. E. 1997. Soft coral abundance in the central Great Barrier Reef: effects of *Acanthaster planci* and the physical environment. Coral Reefs **16**:159–167.
Fabricius, K. E. 1998. Reef invasion by soft corals: Which taxa and which habitats? Pages 77–90 *in* J. G. Greenwood and N. J. Hall, editors. Proceedings, Australian Coral Reef Society 75th Anniversary Conference, Heron Island, October 1997. School of Marine Science, The University of Queensland, Brisbane, Australia.
Fabricius, K. E., Y. Benayahu, A. Genin. 1995*a*. Herbivory in asymbiotic soft corals. Science **268**:90–92.
Fabricius, K. E., and G. De'ath. 1997. The effects of flow, depth and slope on cover of soft coral taxa and growth forms on Davies Reef, Great Barrier Reef. Pages 1071–1076 *in* H. A. Lessios, editor. Proceedings of the Eighth International Coral Reef Symposium. Smithsonian Tropical Research Institute, Balboa, Republic of Panama.
Fabricius, K. E., M. Dommisse. 2000. Depletion of suspended particulate matter over coastal reef communities dominated by zooxanthellate soft corals. Marine Ecology Progress Series **196**:157–167.
Fabricius, K. E., A. Genin, Y. Benayahu. 1995*b*. Flow-dependent herbivory and growth in asymbiotic soft corals. Limnology and Oceanography **40**:1290–1301.
Fabricius, K. E., D. W. Klumpp. 1995. Wide-spread mixotrophy in reef-inhabiting soft corals: the influence of depth, and colony expansion and contraction on photosynthesis. Marine Ecology Progress Series **125**:195–204.
Gohar, H. A. F. 1939. On a new xeniid genus *Efflatounaria*. Annals and Magazine of Natural History **11**:32–36.
Johnson, D. H. 1999. The insignificance of statistical significance testing. Journal of Wildlife Management **63**: (3)763–772.
Little, R. J. A., and D. B. Rubin. 1987. Statistical analysis with missing data. Wiley, New York, New York, USA.
Rejwan, C., N. C. Collins, L. J. Brunner, B. J. Shuter, M. S. Ridgway. 1999. Tree regression analysis on the nesting habitat of smallmouth bass. Ecology **80**:341–348.
Revelante, N., and M. Gilmartin. 1982. Dynamics of phytoplankton in the Great Barrier Reef Lagoon. Journal of Plankton Research **4**:47–76.
Ripley, B. D. 1996. Pattern recognition and neural networks. Cambridge University Press, Cambridge, UK.
Sakamoto, Y., M. Ishiguro, and G. Kitagawa. 1986. Akaike information criterion statistics. D. Reidel Publishing Company, New York, New York, USA.
Schlichter, D. 1982. Nutritional strategies in cnidarians: the absorption, translocation, and utilization of dissolved nutrients by *Heteroxenia fuscescens*. American Zoologist **22**: 659–669.
Schwarz, G. 1978. Estimating the dimension of a model. Annals of Statistics **6**:461–464.
Segal, M. R. 1988. Regression trees for censored data. Biometrics **44**:35–47.
Statistical Sciences. 1999. S-PLUS, version 2000 for Windows. Mathsoft Inc., Seattle, Washington, USA.
Staub, J. E., L. D. Knerr, D. J. Holder, and B. May. 1992. Phylogenetic relationships among several African *Cucumis* species. Canadian Journal of Botany **70**:509–517.
Stewart-Oaten, A. 1996. Goals in environmental monitoring. Pages 17–28 *in* R. J. Schmitt and C. W. Osenberg, editors. Detecting ecological impacts. Academic Press, San Diego, California, USA.
SPSS Inc. 1997. SYSTAT Version 7.0 for Windows. Prentice Hall, New Jersey, USA.
Utinomi, H. 1951. *Asterospicularia laurae*, n. gen. et n. sp., the type of a new family of alcyonarians with stellate spicules. Pacific Science **5**:190–196.
Venables, W. N., and B. D. Ripley. 1999. Modern applied statistics with S-Plus. Third Version. Springer Verlag, New York, New York, USA.
Versefeldt, J. 1977. Australian Octocorallia (Coelenterata). Australian Journal of Marine and Freshwater Research **28**: 171–240.
Versefeldt, J. 1980. A revision of the genus *Sinularia* May (Octocorallia, Alcyonacea). Zoologische Verhandelingen, Rijksmuseum van Natuurlijke Histstorie te Leiden **179**:1–166.
Wolanski, E. 1994. Physical oceanographic processes of the Great Barrier Reef. CRC Press, London, UK.