Ecology, 93(1), 2012, pp. 156–168
© 2012 by the Ecological Society of America

Gradient forests: calculating importance gradients on physical predictors

Nick Ellis,[1,3] Stephen J. Smith,[2] and C. Roland Pitcher[1]

[1]*CSIRO Marine and Atmospheric Research, Ecosciences Precinct, GPO Box 2583, Brisbane, Queensland 4001 Australia*
[2]*Bedford Institute of Oceanography, Fisheries and Oceans Canada, 1 Challenger Drive, Dartmouth, Nova Scotia B2Y 4A2 Canada*

Abstract. In ecological analyses of species and community distributions there is interest in the nature of their responses to environmental gradients and in identifying the most important environmental variables, which may be used for predicting patterns of biodiversity. Methods such as random forests already exist to assess predictor importance for individual species and to indicate where along gradients abundance changes. However, there is a need to extend these methods to whole assemblages, to establish where along the range of these gradients the important compositional changes occur, and to identify any important thresholds or change points. We develop such a method, called "gradient forest," which is an extension of the random forest approach. By synthesizing the cross-validated R^2 and accuracy importance measures from univariate random forest analyses across multiple species, sampling devices, and surveys, gradient forest obtains a monotonic function of each predictor that represents the compositional turnover along the gradient of the predictor. When applied to a synthetic data set, the method correctly identified the important predictors and delineated where the compositional change points occurred along these gradients. Application of gradient forest to a real data set from part of the Great Barrier Reef identified mud fraction of the sediment as the most important predictor, with highest compositional turnover occurring at mud fraction values around 25%, and provided similar information for other predictors. Such refined information allows for more accurate capturing of biodiversity patterns for the purposes of bioregionalization, delineation of protected areas, or designing of biodiversity surveys.

Key words: biodiversity; community data; compositional turnover; Great Barrier Reef; random forests; variable importance.

Introduction

The development of models to predict the geographic distribution of individual species in terms of their current environment dates back to the early 1990s in ecological research (see reviews in Guisan and Zimmermann 2000, Austin 2002, and Guisan and Thuiller 2005). Recent concerns about direct human impacts (e.g., land use, fishing) and potential climate change impacts on species distribution and abundance have focused attention on methods to develop predictive habitat distribution models (Guisan and Thuiller 2005, Elith and Leathwick 2009, Lawler et al. 2011).

A number of statistical methods (e.g., linear models, generalized linear or additive models, discriminant analysis, regression, or classification trees) have been used to relate a species' geographic occurrence or abundance to predictor variables representing topographic or biophysical features or both, depending upon the degree of spatial resolution of the data (e.g., Elith and Leathwick 2009). Empirical evaluations of the performance of these methods including the ensemble classification/regression tree method, Random Forest (Breiman 2001), have generally concluded that this latter method was superior to the others investigated in terms of prediction on test data and various other criteria (e.g., Lawler et al. 2006, Prasad et al. 2006, Cutler et al. 2007, Peters et al. 2007, Knudby et al. 2010).

For community modeling, Ferrier and Guisan (2006) reviewed the overall approaches that have been employed, and summarized them into three broad strategies: *assemble first*, *predict later*, where community indices are derived on the site data then extrapolated to geographical space by modeling on environmental predictors; *predict first*, *assemble later*, where species distributions predicted over geographical space then undergo classification, ordination, or aggregation to provide community information; and *assemble and predict together*, where the processes are combined in a simultaneous analysis. The first approach does not concern us here. Ecological applications of random forests have been limited to the separate analysis of data on individual species (e.g., Prasad et al. 2006, Evans and Cushman 2009). Such analyses could comprise the "predict" step of the second approach. In this paper, we provide a method, called "gradient forest" which extends the random forest method to the community

Manuscript received 11 February 2011; revised 22 July 2011; accepted 26 July 2011. Corresponding Editor: M. Fortin.
[3] E-mail: Nick.Ellis@csiro.au

level and sits within the assemble-and-predict-together approach defined by Ferrier and Guisan (2006).

Our method has some parallels with generalized dissimilarity modeling (GDM; Ferrier et al. 2007), which is also an assemble-and-predict-together approach. GDM assumes that along each environmental gradient there is a monotonic function that expresses the compositional turnover along the gradient. The Bray-Curtis dissimilarity between any pair of sites is expressed as a function of a linear combination of absolute differences between the sites of these turnover functions. Using a framework akin to generalized linear models (GLMs), the turnover functions can be estimated from the site-by-site dissimilarity matrix.

Our method also uses the concept of turnover functions. However, instead of defining the functions indirectly via their relationship with dissimilarity, we use the outputs of random forest models to construct the functions directly. Separate random forests are grown for each species, as in the predict-first, assemble-later approach. Each random forest consists of an ensemble of trees that each recursively split the observations into partitions, where the splitting occurs at certain split values of an environmental predictor. The degree to which abundance changes across adjoining partitions represents a quantum of compositional turnover occurring at the split value. By aggregating these quanta over all species, using weighting that takes into account predictor importance and the goodness of fit of each species distribution model, the method produces functions that represent the compositional turnover along each environmental gradient. These turnover functions, which may be used to support spatial planning, management, and conservation (see *Discussion*), are the primary output of the gradient forest method.

Methods

Regression trees and random forests

The basic element of a random forest model for species abundance data (count or biomass) over sites is a regression tree (or classification tree for presence/absence data). In a regression tree, the sites are partitioned into a left group and a right group in such a way that the abundances within the two partitions are as homogeneous as possible. This partitioning is constrained to be on a split value s for some predictor p; that is, the left group comprises sites having predictor value $\leq s$ and the right group having predictor value $> s$. At each stage the combination of predictor and split is chosen that leads to the smallest total *impurity*, which is the sum of squared deviations about the group mean. The partitions are further partitioned in the same way, the process repeating recursively until a minimum number of sites in the partition is attained, when the partition becomes a terminal node. To predict on data at a new site the tree is traversed according to the decision rules until a terminal node is reached, where the predicted value is the mean of the response in that node. At each node in the tree the *importance* of a split is the reduction in impurity in the node induced by the split. It measures how much of the variation has been explained by the partitioning.

Consider a set of species that are sensitive to an environmental gradient having a threshold, such that some species are present below, but absent above, the threshold, whereas others are present above, but absent below, the threshold. A regression tree model of the abundance of any of those species would be likely to have its first split point s close to the value of that threshold. Moreover, this split is likely to be the most important split in the tree because it includes all the sites. Therefore, the split values and importances hold information about the compositional turnover.

A random forest (Breiman 2001) is an ensemble of a large number of regression (or classification) trees, in which each tree is fit to a bootstrap sample (i.e., with replacement) of the observations, and each partition within a tree is split on the best of a random subsample of the predictor variables. The prediction of a regression or classification random forest is, respectively, the average or majority class of the predictions of each tree. Single-tree models have a high degree of flexibility for fitting complex dependencies on the explanatory variables and so have low bias, but they suffer from high prediction variability on new data. By aggregating over many such low-bias models and using randomization to reduce the correlation between trees in the ensemble, random forests retain the low bias of single trees but with greatly reduced prediction variance. In this paper we focus mainly on regression trees, although the method extends to classification trees, restricted to presence–absence response, using a modified form of R^2 (see *Discussion*). See Prasad et al. (2006) for an explanation of the random forest method written for ecologists; chapter 15 in Hastie et al. (2009) explains the random forest algorithm and why it works.

Gradient forest measures

Random forests provide three measures that we use to investigate biodiversity response: the goodness-of-fit measure R^2_f for the forest for species f, the accuracy importance I_{fp} for a predictor p within the forest, and the raw importance I_{fpts} for that predictor at a split value s in a particular tree t. Ultimately, the turnover functions on the physical predictors are obtained by distributing the R^2 values from all species among the predictors in proportion to their accuracy importance and along the predictor gradient according to the density of the raw importances (Fig. 1).

The observations that were not selected in the bootstrap sample for a tree are called the out-of-bag (OOB) sample. For each observation, the OOB prediction is the average of the predictions on all trees for which the observation was OOB. Because these observations were not used in building the trees, they provide a cross-validated estimate of the generalization error,

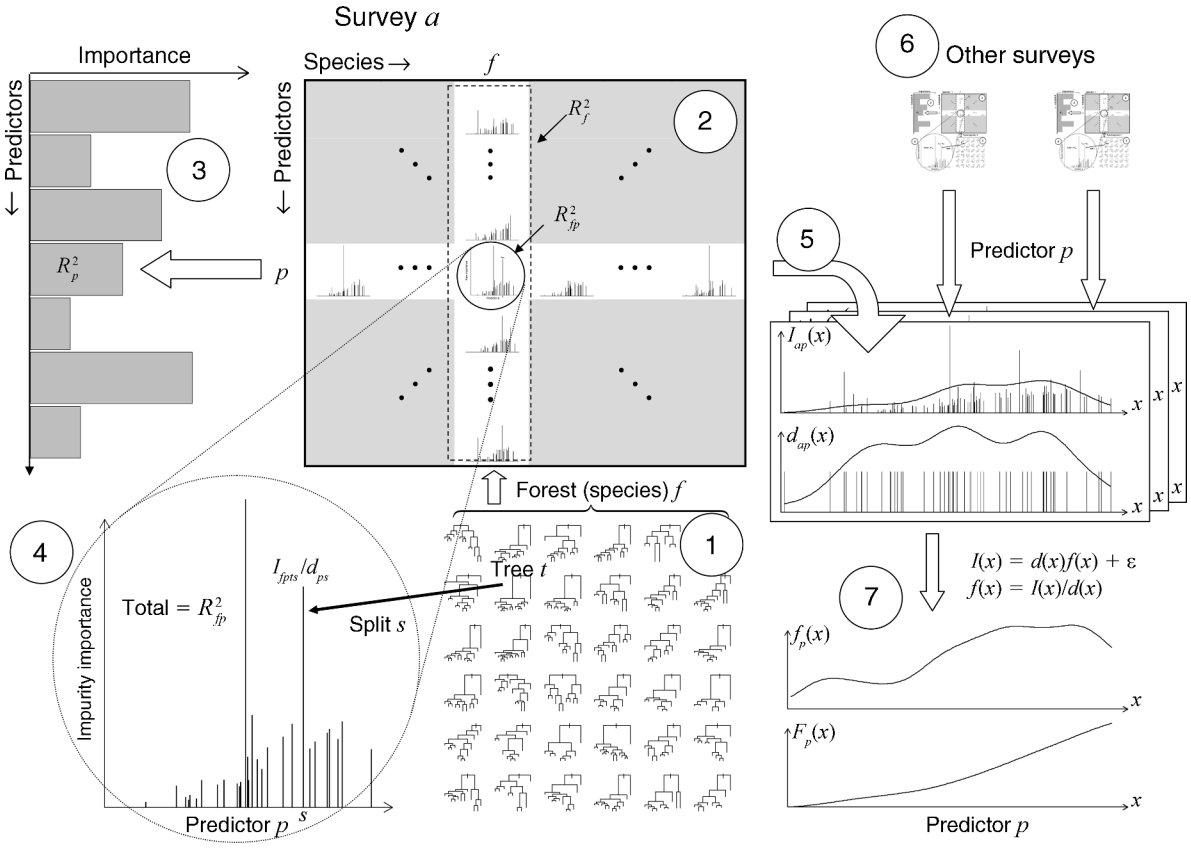

FIG. 1. Schematic of the processing involved in creating a gradient forest. (1) Within a survey a, a random forest is generated for each species f. (2) The goodness of fit R^2_f is partitioned among the predictors in proportion to their conditional importance, yielding R^2_{fp} for predictor p. (3) The overall importance R^2_p for a predictor is found by averaging across species. (4) For each predictor p in each forest f the splits s and impurity importances I_{fpts} are gathered from every tree t in the forest. The importances are standardized by the density and normalized to sum to R^2_{fp}. (5) For each predictor p the normalized importances are gathered across species, and a combined importance density I_{ap} is computed for each predictor value x. (6) The process is repeated for other surveys. (7) A combined estimate of the compositional turnover rate $f_p(x)$ is obtained by estimation from Eq. 4. In the case of a single survey, this is the ratio $I(x)/d(x)$. The compositional turnover function $F_p(x)$ is estimated as the integral of $f_p(x)$.

which is the expected variance of the residuals for new observations. By comparing this variance with the variance of the observations we obtain R^2_f; thus,

$$R^2_f = 1 - \sum_i (Y_{fi} - \hat{Y}_{fi})^2 \Big/ \sum_i (Y_{fi} - \bar{Y}_f)^2 \qquad (1)$$

where Y_{fi} is the ith abundance observation, $\hat{Y}_{fi}$ is the OOB prediction and $\bar{Y}_f$ is the overall mean for species f. The data Y_{fi} may have undergone transformation prior to modeling by random forest. R^2_f is the proportion of variance explained by the random forest; it can be used as a measure of the information on biodiversity provided by the predictors on that particular species. It is mathematically possible for the expected generalization error, as estimated by the OOB process, to exceed the observed variance, so for some species it may happen that $R^2_f < 0$. In such cases the predictors are assessed to have no predictive power within the random forest paradigm.

Each random forest also provides a measure of the importance of each predictor called the accuracy importance I_{fp}. This is the increase in the OOB mean square prediction error when the predictor p is randomly permuted, thus removing any signal there may have been due to the predictor. It is analogous to the error in dropping a term from a linear model. Large values of I_{fp} indicate the predictor has true predictive power which is eroded when the predictor is permuted, whereas small (or even negative) values indicate the predictor explains little or nothing. (For negative values we simply replace the value with 0.)

The quantity R^2_f can be partitioned into contributions R^2_{fp} from each predictor p in proportion to the accuracy importance; thus,

$$R^2_{fp} = \frac{R^2_f I_{fp}}{\sum_{p'} I_{fp'}} \qquad (2)$$

so that $\Sigma_p R^2_{fp} = R^2_f$ (Fig. 1, item 2). Because the R^2_{fp}

quantities are in units independent of the species, they can be averaged across all species within a survey to provide the overall importance of a predictor in a survey:

$$R_p^2 = \frac{1}{N^{\text{sp}}} \sum_f R_{fp}^2 \tag{3}$$

where N^{sp} the number of species in the survey (Fig. 1, item 3).

Within the forest for species f each tree t comprises many non-terminal nodes, each of which splits on some predictor p at split value s. Let I_{ftps} denote the raw importance of such a split (Fig. 1, items 1 and 4). This is the (in-bag) reduction in impurity due to the split, where, for regression trees, impurity is the total sum of square deviation from the node mean.

Correlated predictors

Many of the predictor variables used in ecological studies are correlated, either naturally (e.g., decreasing temperatures with water depth) or functionally (e.g., benthic irradiance calculated as a function of bottom depth and light attenuation). While some of these predictors may directly determine species distribution or abundance, other correlated predictors may have no influence.

At each node random forest chooses the predictor to supply the split from only a random subset of the available predictors. This feature of the process can result in a correlated but less influential predictor standing in for more highly influential predictors in the early splits of an individual tree, depending upon which predictor is selected in the subset. Breiman (2001) suggested this tendency could be lessened by increasing the subsample size of predictors for each node, but the trade-off would be an increase in correlation between trees in the forest, with concurrent increase in generalization error and a decrease in accuracy (Breiman 2001, see also Grömping 2009). Moreover, Nicodemus et al. (2010) have shown that even with increasing subsample size, the permutation method for estimating variable importance results in inflated importance measures for correlated predictors that are related to the response, relative to uncorrelated predictors that are equally related to the response. While increasing the subsample size of predictors at each node will reduce the tendency to choose correlated predictors when considered over all splits, the permutation method still results in higher selection frequencies for correlated predictors at the first node, which is associated with the largest (in-bag) reduction in impurity for all nodes in any one tree.

Strobl et al. (2008) have demonstrated that the underlying reason for this behavior has to do with the structure of the null hypothesis in the permutation method, i.e., independence between the response Y and the predictor X_j being permuted, implied by the importance measure. A small value for the importance measure would suggest that Y and X_j are independent but also assumes that X_j is independent of the other predictor variables Z in the model that were not permuted ($Z = X_i,\ i \neq j$). Correlation between X_j and Z will result in an apparent increase in importance reflecting the lack of independence between X_j and Z instead of only reflecting the lack of independence between X_j and Y.

To remedy this situation, Strobl et al. (2008) proposed a conditional permutation approach where the values for X_j in the OOB sample for each tree are permuted within partitions of the values of the predictors that are correlated with X_j. Permutation importance is calculated by passing the OOB samples reconfigured with this permutation grid through each respective tree in the standard way. The implementation of this method is explained in Appendix A and its use is demonstrated on a synthetic example in Supplement 1.

Compositional turnover

We define compositional turnover by aggregating over individual species turnover. In our random forest approach the basic building block of species turnover is the predictor split and associated importance within a tree. Where importance is high, the change in abundance, and therefore the turnover, is high. We therefore identify species turnover with split importance. We also identify the total turnover over the range of the predictor p with the amount of variation explained by the predictor R_{fp}^2. Combining these definitions, we define the species turnover $F_{fp}(x)$ as the monotonic function with minimum 0 and maximum R_{fp}^2, such that, within each interval (x_1,x_2), $F_{fp}(x_2) - F_{fp}(x_1)$ is proportional to the (suitably standardized) importance of all splits in the interval. The compositional turnover $F_p(x)$ becomes the average $F_{fp}(x)$ over all species.

In effect, we use an empirical definition for compositional turnover to be the R^2 attributable to the predictor over the range of x. Other definitions for turnover are possible, such as the change in mean absolute abundance between the child nodes or the raw importance. However, we prefer using R^2 as the unit of compositional turnover because it enables one to combine information across species within a survey, and also across multiple surveys. This is because R^2 is a dimensionless quantity that is independent of the units of abundance of the species. It also allows for the relative importance of predictors to be assessed.

The derivative $f_p(x)$ of $F_p(x)$ is the compositional turnover rate at predictor value x. Consider a single survey, denoted by subscript a. If p were sampled uniformly, $f_p(x)$ would be equal to the expected value of $I_{ap}(x)$, the observed importance density, suitably normalized. However, when the sampling is not uniform, the observed distribution of splits becomes biased toward values that are more densely sampled. Let $d_{ap}(x)$ be the *scaled* density of predictor values over the observed range Δ_a, normalized to satisfy $\int d_{ap}(x)dx =$

Δ_a. We propose the following standardizing relationship:

$$I_{ap}(x) = d_{ap}(x) f_p(x) + \varepsilon \quad (4)$$

where ε is a random variable. To ensure the empirical definition of $F_p(x)$, $I_{ap}(x)$ is normalized to satisfy $\int I_{ap}(x)/d_{ap}(x)dx = R^2_{ap}$ (Fig. 1, item 5).

Because $d_{ap}(x)$ occurs in the denominator of a ratio, it is important that it not have zero values where there is non-zero importance. This is a potential problem when a split point lies between two neighboring predictor values that are widely separated relative to the overall distribution of values. Therefore, first a (scaled) density $k_{ap}(x)$ is computed using a Gaussian kernel with bandwidth given by Silverman's rule-of-thumb (Silverman 1986: Eq. 3.31), as implemented in the R function *density*. This is then "whitened" by mixing in a uniform distribution $d_{ap}(x) = \lambda k_{ap}(x) + 1 - \lambda$. Setting $\lambda = 1$ uses the kernel density and $\lambda = 0$ a uniform density; in this paper $\lambda = 0.9$ was used as a compromise.

Assuming a distribution for ε and a parametric form for $f_p(x)$, one can estimate $f_p(x)$ by standard methods such as maximum likelihood. For example, if $\varepsilon \underset{\text{i.i.d}}{\sim} \mathcal{N}(0,\sigma)$ and $f_p(x)$ is parameterized using a B-spline basis in x, then the spline parameters can be estimated using ordinary least squares. It seems reasonable to assume that the variance of $I_{ap}(x)$ is inversely proportional to the number of predictor values in the neighborhood of x, so that $\text{var}[I_{ap}(x)] \propto 1/N_a^{\text{site}} d_{ap}(x)$, where N_a^{site} is the number of sample sites in survey a. In this case, $f_p(x)$ is estimated by weighted least squares.

An alternative to the additive error term in Eq. 4 is a multiplicative error $I_{ap}(x) = d_{ap}(x) f_p(x)\eta$, where η is a lognormal random variable with mean 1. Working on the log scale, one could estimate $\log f_p$ using $\log d_{ap}(x)$ as an offset. This alternative has the advantage of a more natural error model, with the disadvantage of having to back-transform the estimate from the log scale.

The density $I_{ap}(x)$ could be calculated from the observed splits $x = s$ by weighted kernel density estimation with weight equal to the raw importance at the split I_{aftps}. However random forests usually provide such a large number of splits that simply binning the raw importances is adequate. That is, if the predictor range is split into intervals $[X_i, X_{i+1}]$, $i = 1, \ldots, B_p$, each of width ΔX, then, for survey a, define

$$\frac{I_{ap}(x)\Delta X}{d_{ap}(x)} = \frac{1}{N_a^{\text{sp}}} \sum_f \left[R^2_{afp} \sum_{t,s \in \text{BIN}(x)} \frac{I_{afpts}}{d_{ap}(s)} \Big/ \sum_{t',s'} \frac{I_{afpt's'}}{d_{ap}(s')} \right] \quad (5)$$

where $\text{BIN}(x)$ is the interval in which x lies. The weighting of I_{aftps} inside the summation ensures that $\int [I_{ap}(x)/d_{ap}(x)]dx = R^2_{ap}$.

Given that $I_{ap}(x)$ has been binned, it is natural to consider a parameterization of $f_p(x)$ as a stepwise function on the same bins (so there is one parameter per bin). When there is only a single survey a, the estimate of $f_p(x)$ is simply $\hat{f}_{ap}(x) = I_{ap}(s)/d_{ap}(x)$, and the estimated compositional turnover function is then $\hat{F}_{ap}(x) = \int_{-\infty}^{x} \hat{f}_{ap}(y)dy$ (Fig. 1, item 7). At the resolution of a single species, we have the completely analogous results $\hat{f}_{afp}(x) = I_{afp}(x)/d_{ap}(x)$ and $\hat{F}_{afp}(x) = \int_{-\infty}^{x} \hat{f}_{afp}(y)dy$, where I_{afp} is obtained from Eq. 5 restricting the survey to a single species f.

When there are multiple surveys, the density at x varies among surveys and so the reliability of the information varies too. Moreover, since each species is potentially one unit of compositional turnover, different surveys carry different weight depending on how many species were observed. The combined estimate then becomes the following (see Appendix B):

$$\hat{f}_p(x) = \frac{\sum_a \left[\hat{f}_{ap}(x) N_a^{\text{site}} d_{ap}^3(x) N_a^{\text{sp}} W_a \right]}{\sum_a \left[N_a^{\text{site}} d_{ap}^3(x) N_a^{sp} W_a \right]} \quad (6)$$

and, as before, $\hat{F}_p(x) = \int_{-\infty}^{x} \hat{f}_p(y)dy$ (Fig. 1, items 6 and 7). The extra term W_a allows for alternative weighting of the surveys. For instance, setting $W_a = 1/N_a^{\text{sp}}$ would weight each survey according to its average R^2 over species instead of its total R^2 over species. The setting used in the examples in this paper is $W_a = 1$.

Software implementation

To implement our approach, we have written two packages in the R computing environment (R Development Core Team 2011). The first package, extendedForest, extends the existing package randomForest (Liaw and Wiener 2002, R Development Core Team 2011), which is an implementation of Breiman's random forest models. The new package has all the functionality of randomForest, such as calculating R^2_f and I_{fp}, but it also records the raw importances I_{ftps} required for the gradient forest technique, and allows for the computation of conditional instead of marginal permutation importance, with correlated predictors determined optionally by Pearson's ρ or Kendall's τ (Kendall 1938). The second package, gradientForest, uses the extensions provided by extendedForest to compute the gradient forests and turnover functions for both abundance and presence/absence data, and to provide functions for prediction and visualization. Both packages are available from R-Forge (*available online*).[4]

Results

Synthetic example

The method can be demonstrated and tested using a synthetic example. Consider 100 "sites" having 10 physical predictors *A–J*, each a uniform random sample on the interval [0,1]. We shall simulate a group of species whose abundance depends on the predictors *A* and *B*, but not on the remaining predictors *C–J*. Our method should identify the important predictors as well as where

[4] http://gradientforest.r-forge.r-project.org

along their range the compositional change points occur. We construct a compositional gradient on predictor A and populate it with three species, ($a_{1–3}$), centered on $\mu_{A,1–3}$, respectively with spread σ_A and abundance determined by a Poisson process with intensity proportional to λ_A. The species count at a site with predictor $A = x_A$ is generated thus:

$$a_i(x) \sim \text{Poisson}\left(\lambda_A \sigma_A^{-1} \phi(x_A - \mu_{A,i})\right) \quad i = 1, 2, 3 \quad (7)$$

where $\phi(\cdot)$ is the standard normal density. We also construct four species $b_{1–4}$ in a similar way with B as the environmental gradient having species locations $\mu_{B,1–4}$, spreads σ_B and intensities λ_B:

$$b_i(x) \sim \text{Poisson}\left(\lambda_B \sigma_B^{-1} \phi(x_B - \mu_{B,i})\right) \quad i = 1, \ldots, 4. \quad (8)$$

Finally we define five species $ab_{1–5}$ with bivariate compositional gradients on A and B having (bivariate) species means $\mu_{1–5}^{AB}$, spreads σ_{AB} and intensities λ_{AB}:

$$ab_i(x) \sim \text{Poisson}\left(\lambda_{AB} \sigma_{AB}^{-2} \phi(x_A - \mu_{A,i}^{AB}) \phi(x_B - \mu_{B,i}^{AB})\right)$$
$$i = 1, \ldots, 5. \quad (9)$$

The distributions of the species are shown in Fig. 2 for settings $\sigma_A = \sigma_B = \sigma_{AB} = 0.1$ and $\lambda_A = \lambda_B = \lambda_{AB} = 10$.

For each species a random forest of 500 trees was generated. The overall importance measures are shown in Fig. 3A (left). The measure based on increase in prediction error very clearly separates the important predictors A and B from the unimportant predictors C–J. Predictor B is found to be the most important, partly because four species depend directly on it, whereas only three depend directly on A, and partly because the fits of the b species are slightly better (Fig. 3A right).

Fig. 3B shows where along the predictor range the splits are and how important they are. The raw importances are first aggregated to narrow bins (vertical bars) from which a density $I(x)$ (black curve) is estimated by kernel density estimation. The estimated compositional turnover rate $\hat{f}(x)$ (blue curve) is obtained by dividing by the normalized density $d(x)$ (red curve) which has a mean of 1. $I(x)$, $\hat{f}(x)$ and the raw importances are together normalized to make the area under the blue curve equal to the overall importance R_p^2 as given in Fig. 3A. Thus, the compositional turnover rate can be compared directly across predictors as shown for predictors A, B, and E. For predictor A, splits lying on the shoulders of the species distributions either side of centers $\mu_{A,1–3}$ are important for explaining compositional change. Similarly for predictor B splits lying between the centers for species b_1 and b_2 and for species b_3 and b_4 are seen to have high importance. Portions of the gradient with relatively high density of splits delineate the boundaries between portions with relatively more homogeneous species composition.

In Fig. 3C (left and middle) the individual compositional turnover functions for each species are shown, on the same scale, for gradients A or B. The maximum value attained in each plot is R_{fA}^2 and R_{fB}^2, respectively. Naturally this value is highest for species $a_{1–3}$ for predictor A and for species $b_{1–4}$ for predictor B, and the values are close to the corresponding R_f^2 (Fig. 3A, right). For species $ab_{1–5}$ the importance is split roughly equally between A and B. Weighting each species equally, the average goodness of fit is higher for predictor B. Fig. 3C (right) shows the combined compositional turnover functions of the predictors, allowing them to be compared directly. The maximum value obtained by each line is the overall goodness of fit of all species, R_p^2, for each predictor A–J. The common y-axis allows for direct comparison among predictors: we can see which predictors are important and where the important compositional changes (steep slopes) occur along their range.

Real data example

We now turn to an example using real species data from a cross-shelf survey in the far northern (11° S–12° S) Great Barrier Reef (Poiner et al. 1998). There were 306 sites sampled and > 1000 species observed, but a subset consisting of the 93 most commonly occurring species were analyzed. There were 29 environmental predictors, either measured at each site or attributed to each site by interpolation in GIS (Pitcher et al. 2002).

The overall importance measure (Fig. 4A) shows that mud, oxygen standard deviation, and bottom stress are the most important predictors. In an earlier study in the same region, (Pitcher et al. 2002), using multiple methods (GLMs, cCCA, trees), found a similar ordering of predictor importance.

The density plots (Fig. 5A) show important splits around 5–10% and 20–25% for mud and around 0.15 N/m^2 for bottom stress. The incidence of data points for mud around 25% was relatively low, but the splits in that range accounted for a relatively larger reduction in impurity. For bottom stress there is also a small peak in the importance density around 0.4 N/m^2. This is not strongly evident in this figure because the window width used to calculate the density functions is rather wide, resulting in a smoothed-out curve. The threshold at 0.4 N/m^2 is more clearly seen from Fig. 5C. This is because the turnover curve is the integral of the binned importances, which have not undergone any smoothing.

The individual compositional turnover functions for each species for these two predictors are shown in Fig. 5B. The highest values attained (at the right of each curve) are the contribution to R_f^2 from each predictor (i.e., R_{fp}^2). Note that many species have strong changes in mud with preponderance around 5% and 25%, as expected. The strong change points for bottom stress are also evident.

When these species curves are averaged, we obtain the overall compositional turnover functions (Fig. 5C, line labeled "Full"). Steep parts of the curve indicate ranges of the predictors where species composition changes, and the flatter regions indicate more homogeneous

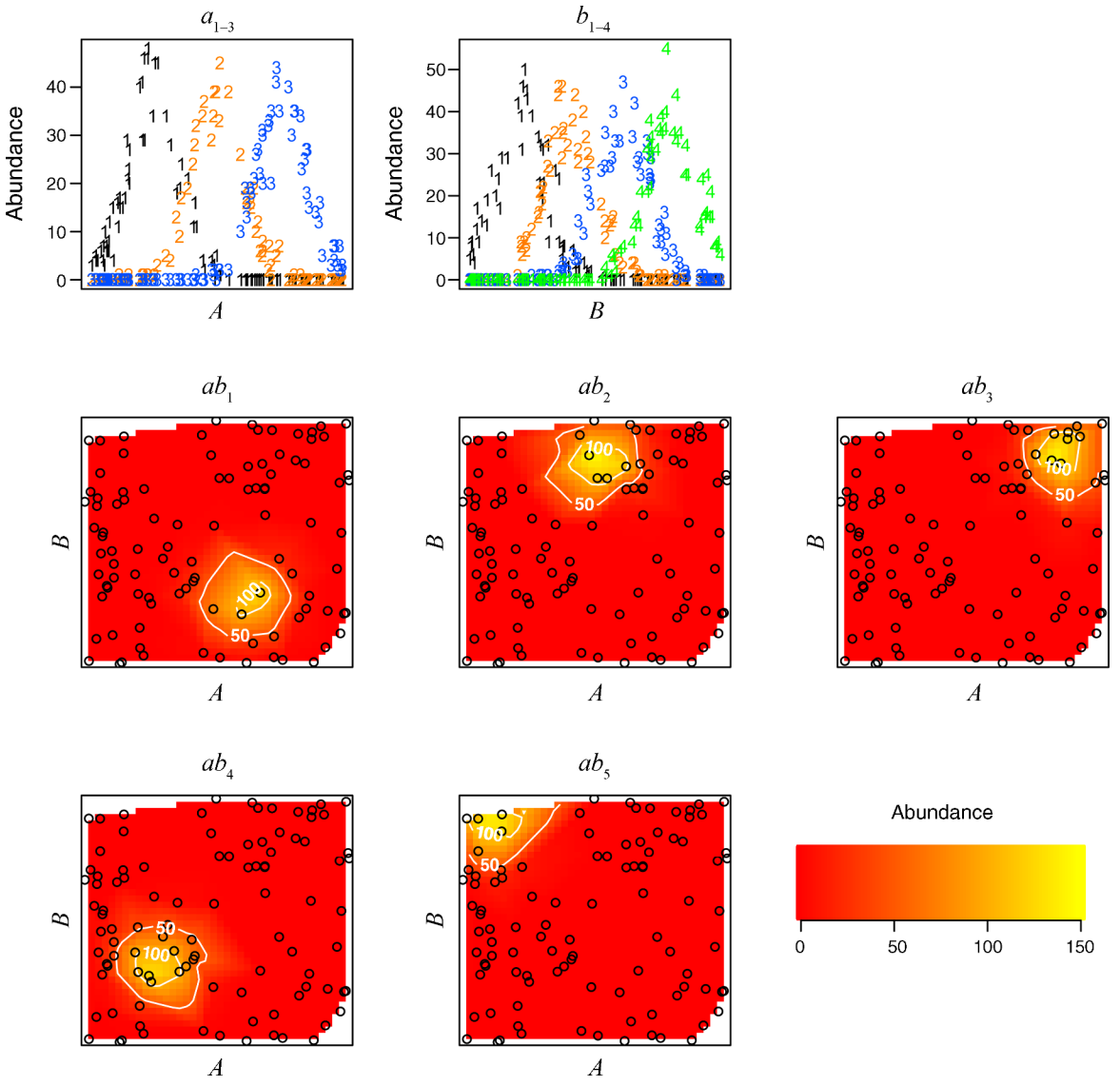

FIG. 2. Properties of various synthetic species. Species a_{1-3} respond to physical driver A, species b_{1-4} respond to physical driver B, and species ab_{1-5} respond jointly to A and B. Numbers (top row) and circles (bottom rows) denote sites. Abundance is shown on an interpolated color scale for species ab_{1-5}.

portions. Each curve can be interpreted as a transformation from environmental space to biological space, showing how environmental gradients translate to biological gradients. Moreover, the common vertical scale allows predictors to be assessed relative to one another in terms of their influence on patterns of species composition.

An important feature of the gradient forest method is that information can be combined across multiple surveys. To examine this we created two smaller data sets, one comprising the westernmost two-thirds and the other the easternmost two-thirds of sites, with an overlap of one third of the sites in the middle. We then estimated separate and combined compositional turnover functions (Fig. 5C). The combined curve approximates the shape and magnitude of the curve obtained from the original full data set ("Full"). The combined curve is obtained from accumulating weighted averages of the separate derivative curves for east and west. Where there is no overlap in the predictor range, the curve depends on only one survey; for example, for bottom stress < 0.12 the combined curve matches the compositional turnover function from the western survey.

DISCUSSION

Gradient forests provide a novel, highly flexible method for quantifying the pattern of change in

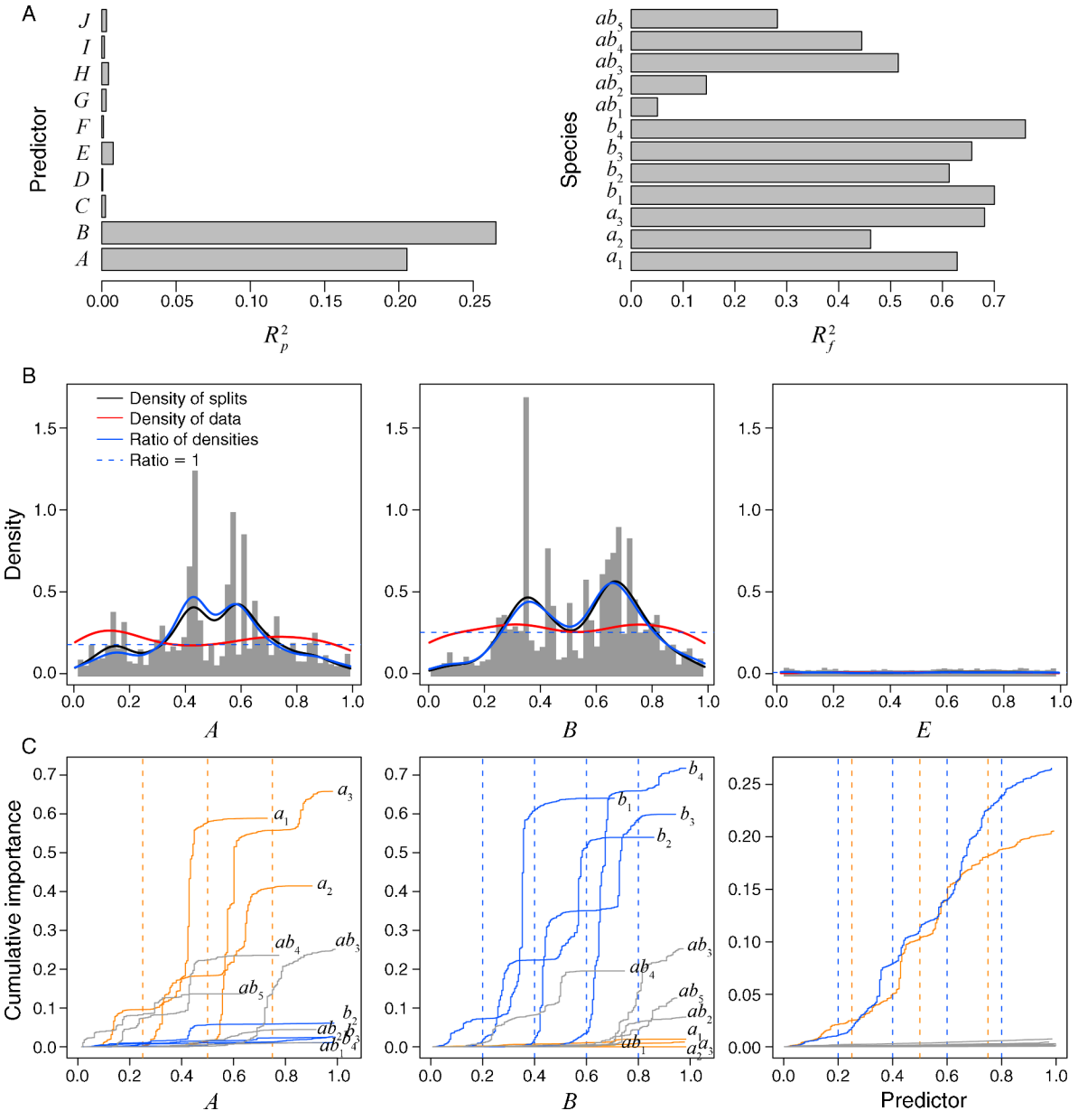

FIG. 3. Results for the synthetic data set. (A) Overall importance R^2_p of the physical drivers (left) and R^2 for each species (right). (B) Results for predictors *A*, *B*, and *E*, with *E* being the uninformative predictor with the highest R^2_p. The black line is the raw importance density $I(x)$ computed by kernel density estimation of split points weighted by importance; gray bars indicate binned raw importance density; the red line is density $d(x)$ of observed predictor values; the blue line is the estimated importance $\hat{I}(x)$ computed as the ratio of importance density to predictor value density, with the horizontal dashed line indicating where the ratio is 1. Each curve is normalized to integrate to R^2_p. (C) Compositional turnover functions $\hat{F}_{fp}(x)$ for predictor *A* (left) and predictor *B* (middle) for all 12 species. The *a* species are in orange, the *b* species are in blue, and the *ab* species are in gray. The rightmost panel shows compositional turnover functions $\hat{F}_p(x)$ for predictors *A* (orange), *B* (blue), and *C*–*J* (gray). For predictors *A* and *B*, the centers of the species distributions are given by dashed lines of the corresponding color.

biodiversity composition along gradients of environmental variables. The method also assesses predictor importance, and supplements existing techniques, such as variance partitioning among predictors (i.e., ANOVA), correlation measures or Akaike weighting (Murray and Conner 2009), permutation methods (Knudby et al. 2010) or predictor loadings in CCA (Ter Braak 1994), by also showing where along the gradients the important changes are occurring.

Like other species or community modeling methods, gradient forest is an exploratory method to seek out relationships between the community and the environ-

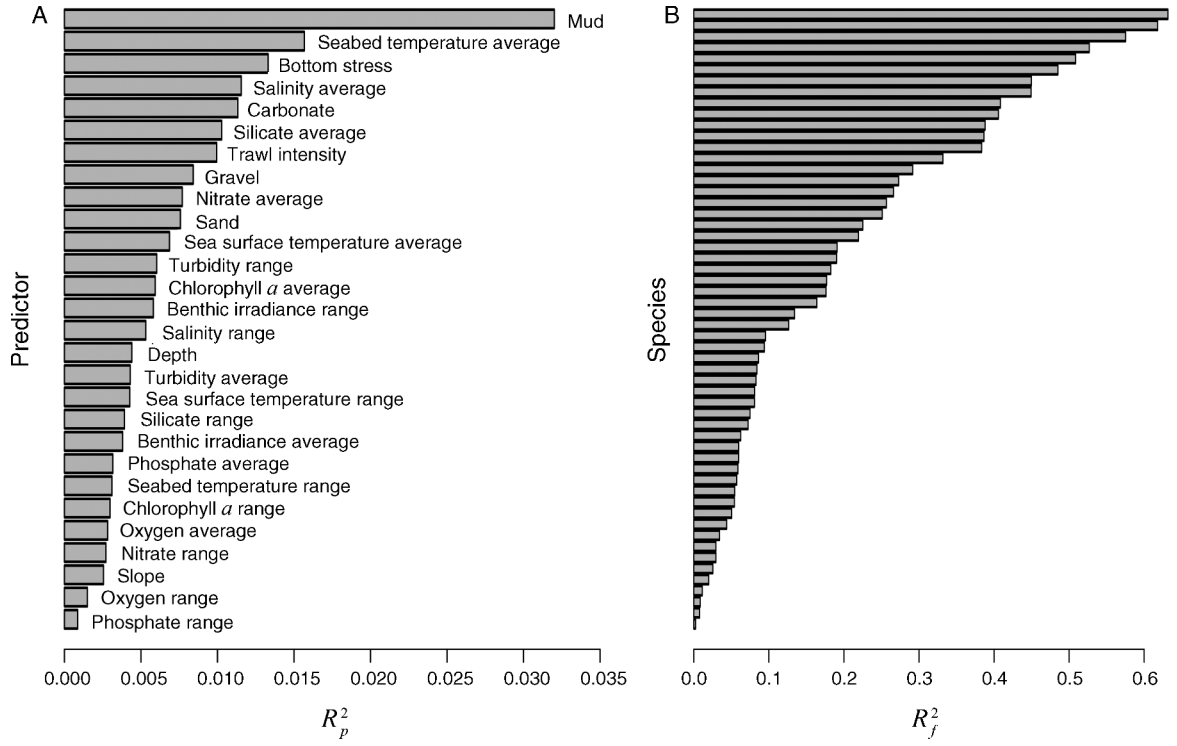

FIG. 4. Results for the Great Barrier Reef (GBR) data set. (A) Overall importance, R^2_p, for each predictor p of the physical drivers and (B) ranked goodness-of-fit, R^2_f, for each species f.

ment. Although such relationships, being correlative in nature, do not necessarily imply a causal link, they may inform as to the possible drivers of community patterns and what might be targeted for causal investigation. Gradient forests, like other modeling methods, are not guaranteed to find the true underlying gradients (which may be related to unmeasured or unmeasurable factors), nor are they guaranteed to explain a large portion of variation in the data. Nevertheless, searching for relationships between measured variables and species composition, and quantifying the extent to which the former may drive the latter, are useful goals in ecological analysis.

Applications

The functions F_p provide a transformation from the arbitrary anthropogenic measurement units of each predictor p to common biological units of compositional turnover. By applying the appropriate function to each predictor in turn, one can transform multidimensional environment space into the corresponding multidimensional biological space. If the environmental surrogates are available over a geographic region, the transformed biological space can be mapped to this region. Such a mapping represents expected patterns of biodiversity composition, without directly predicting compositional patterns. There are several ways to represent the biological space on a map, one of which is to use a color key on the first two or three principal components of the biological space (e.g., Pitcher et al. 2007; see also Ferrier et al. 2007). Alternatively, the biological space can be clustered, and the map colored by cluster label, which would represent the expected constrained assemblage patterns. Such maps can be used to inform regional marine planning and conservation, for instance in the placement of marine protected areas. Existing or proposed MPAs can also be assessed for coverage of biological space, and hence also their representation of biodiversity composition. See Supplement 2 for an example of these mapping techniques applied to the Great Barrier Reef data set.

Similarly, the transformed biological space can be used for gap analysis of the original survey data. The most extreme form of gap analysis is where the intention is to carry out a full-scale survey based on very limited data from a pilot survey. In this case the pilot survey can be used to provide the data for a gradient forest analysis. If deemed appropriate, data from other sources can also be used, for instance from nearby regions or from older historical surveys. The turnover functions from the resulting gradient forest analyses can be combined, suitably weighted, with those from the pilot survey. The resulting transformations are then used to convert the environmental data on the full region to biological space. The design of the full-scale survey can then be guided by stratifying on this biological space (Pitcher et al. 2007).

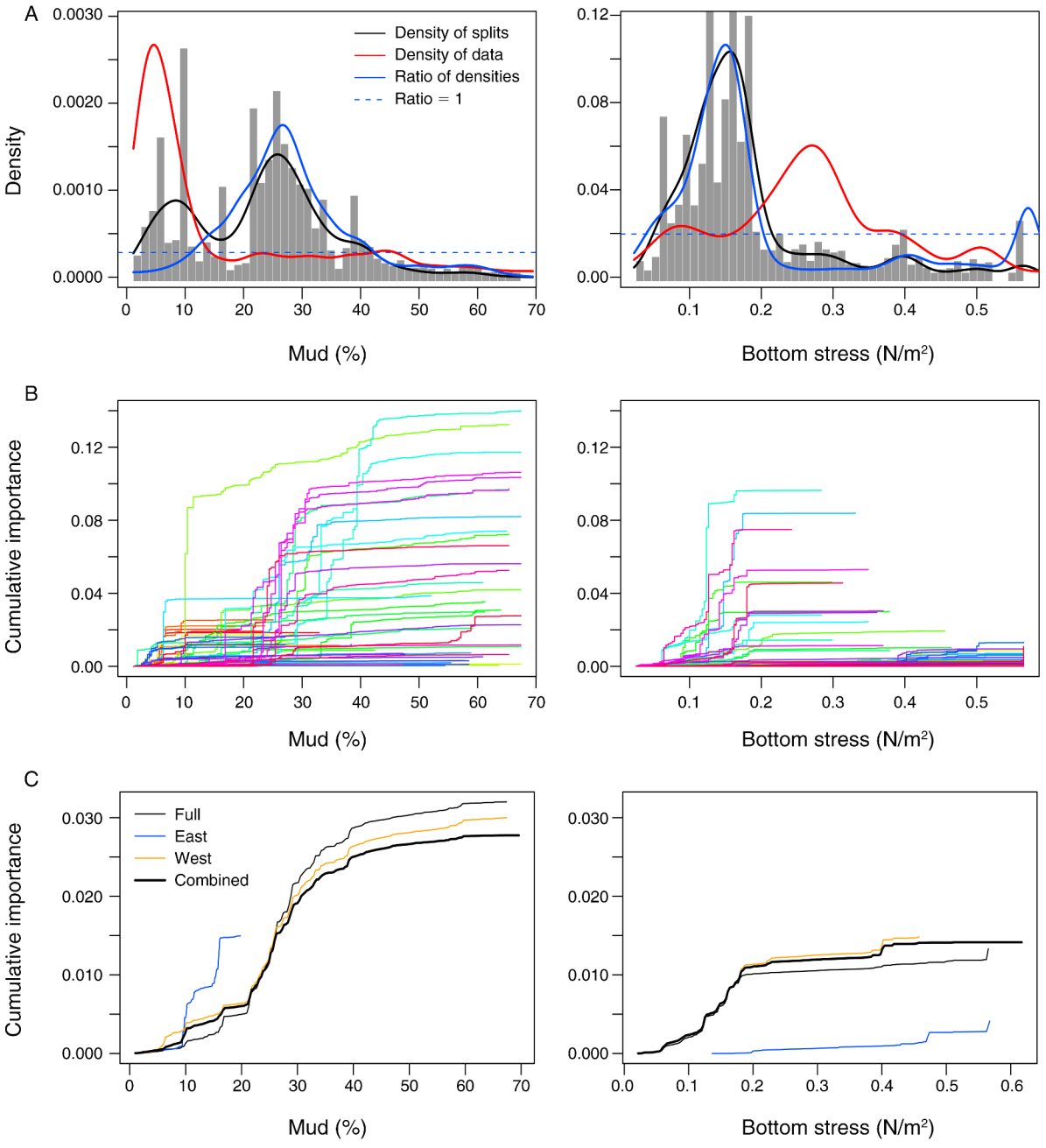

FIG. 5. Results for the Great Barrier Reef (GBR) data set. (A) Results for two important predictors: mud and bottom stress. The panel shows raw importance density $I(x)$ computed by kernel density estimation of split points weighted by importance (black line); binned raw importance density (gray bars); density $d(x)$ of observed predictor values (red line); and estimated importance $\hat{f}(x)$ computed as the ratio of importance density to predictor value density (blue line), with the horizontal dashed line indicating where the ratio is 1. Each curve is normalized to integrate to R^2_p. (B) Compositional turnover functions $\hat{F}_{fp}(x)$ for these two predictors for all 93 species. (C) Overall compositional turnover function $\hat{F}_{fp}(x)$ for these two predictors estimated from gradient forests of the original full data set, from the data set split into east and west sites, and from combining the east and west gradient forests.

Gradient forests can isolate important thresholds in environmental gradients. Changes in the environment, such as global warming, that push the surrogates over these thresholds could have important effects on the biological composition. The degree of change expected at a location can be quantified by the Euclidean distance between the current and future points in biological space. When rendered on a map, such distances could indicate the spatial distribution of potential vulnerability to environmental change.

Strengths and weaknesses

The gradient forest method has several strengths that distinguish it from other methods. First, because it is based on random forests, this machine learning approach can capture very complex relationships between the predictors and the species response with better predictive power than most other methods. Furthermore, model selection is implicit in the fitting process, there is no need to transform the predictors, and complex interactions between predictors are accounted for. In contrast, model selection using parametric methods requires considerable skill in choosing transformations of the predictors and it is usually difficult to anticipate what interactions should be accommodated.

Secondly, the use of the dimensionless R^2 to quantify compositional turnover enables abundance information from multiple species to be combined, even if that information comes from different sampling devices, surveys or regions and is of different type, such as counts or weight. Regression trees require the response variable to be transformed to have approximately normal errors. When the response is binary (i.e., presence–absence) it is more appropriate to use random forests based on classification trees. A proposed analogue of R^2 for classification trees is $R_c^2 = 1 - \text{OOB}$ misclassification rate/base error rate, where base error rate $= 2p(1 - p)$ and p is the prevalence of the species. R_c^2 is analogous to R^2 because it is 0 when the model has no predictive power and 1 when the model predicts perfectly. The analogue of the raw importances is the decrease is Gini impurity. The method for calculating F_p then follows analogously to the regression case.

Thirdly, unlike dissimilarity measures such as Bray-Curtis, gradient forests do not artificially constrain the degree of compositional change. If composition continues to change along an environmental gradient, the turnover function F_p continues to increase, whereas the Bray-Curtis dissimilarity reaches its maximum value (1) at some point along the gradient. Hence the patterns of compositional change will be less subject to the distortion effects inherent in methods based on Bray-Curtis dissimilarity.

Gradient forests can account for the inflation of the variable importance measures for correlated predictors using a conditional permutation strategy. While the extent to which this is successful depends on the degree of independent signal in the data, potentially spurious predictors should have reduced importance assigned to them compared to the results from an unconditional approach. Correlated predictors represent no major obstacle for gradient forest and its outputs; they are unlikely to affect the results much within a data region, and their correlations are obvious in a biplot of the final mapped product. However, caution is needed applying results to another region, especially where the available sets of predictor variables differ.

It is usually difficult to disentangle and interpret the independent effects of correlated predictors, regardless of numerical method employed. There is some controversy as to whether some correlated predictors should be dropped or whether all predictors should be retained (e.g., Cutler et al. 2007, Murray and Conner 2009, Knudby et al. 2010). We prefer to retain all predictors, partly because most predictors are correlated to some extent, so that any particular choice of exclusion is hard to justify, and partly because we do not know a priori which are the most important predictors that are truly related to the response.

Gradient forests account for spatial effects indirectly through the environmental predictors. Large scale spatial patterns unrelated to or correlated with environmental variables can also be modeled by including geographic position. However, gradient forests do not readily account for small-scale spatial autocorrelation, because of the assumption of independent normal errors. In principle it would be possible to account for spatial autocorrelation by generalizing tree node impurity to $\mathbf{r}^\top \mathbf{C}^{-1} \mathbf{r}$, where $\mathbf{C}$ is the correlation matrix and $\mathbf{r}$ is the vector of residuals within a node. The optimization would then be based on generalized least squares instead of ordinary least squares. However such a procedure would be very slow to run because it requires inverting two submatrices of $\mathbf{C}$ for every candidate split.

The extent to which not accounting for spatial autocorrelation is an issue will depend on the scale of sampling in the original surveys. For instance, in the Great Barrier Reef, Pitcher et al. (2007) found that in the majority of analyses the local correlation dissipated after about 5–6 km, hence the survey sites were chosen to be mostly 10–15 km apart (with $<1\%$ 3–5 km apart). Furthermore, gradient forest is an exploratory method, with the advantages outlined above and may be used in complement with other methods, such as generalized dissimilarity modeling (GDM; Ferrier et al. 2007), that can accommodate spatial distance between sites.

Comparison with other methods

There appear to be rather few available methods for modeling constrained community patterns in relation to environmental variables: canonical correspondence analysis (CCA; Ter Braak 1994), multivariate regression trees (MRT; De'ath 2002) and multivariate adaptive regression splines (MARS; Friedman 1991, see also Leathwick et al. 2006), and GDM. Of these, only GDM shares the property of gradient forest of providing information on the rate of community change along each environmental gradient. As such, these two methods lie on the continuum between CCA, having linear change along an environmental gradient, and MRT, having jumps along the gradient. MARS, being a smoother form of tree, also lies somewhere between MRT and CCA.

GDM is a very different method to gradient forest. It models a dissimilarity response within a GLM-like framework as a function of a linear combination of absolute differences of nonlinear monotonic functions f_p

of each predictor p. Nevertheless, this function is analogous to the gradient forest turnover function F_p. In GDM, the f_p functions are estimated as linear combinations of I-splines with positive coefficients, a constraint that satisfies the monotonicity condition, and the derivatives of these linear combinations are positive functions. The choice of how many degrees of freedom to use in the I-splines is analogous to the choice of bandwidth for density estimation in gradient forests. In gradient forests, the turnover rate (which is analogous to these derivatives) is estimated using a form of density estimation, either simple binning or kernel density estimation. In GDM, the turnover function is found by finding a function that fits many distances (site pairs). In gradient forests, the turnover function is constructed directly by integrating many instances of local distances on the predictor gradient (split points). It would be interesting to compare these complementary approaches on the same data sets.

Future development

The overall importance of predictors is derived from permutation on the OOB sample and thus can be considered robust. However the importance of the individual splits is derived from the in-bag split impurities, which may not accurately reflect the true importance of the split. Gradient forest relies on the averaging effect over many trees and species in combination with the robust overall importance estimates to derive importance density measures f_p. These measures could be improved by replacing the in-bag impurities by an alternative measure based on loss of prediction accuracy on permuting the predictor at the node where the split occurs.

Often in ecology the sampling distribution from which the biological data are drawn is ignored. For instance, when dissimilarity is calculated from presence–absence data, the absences may be due to sampling bias rather than to true absence of the biota. Bias can be due to diurnal, monthly or seasonal behavior or differing catchability by sampling device, which, especially for a comprehensive survey, cannot be anticipated in the design phase. Generalized linear models offer one way to adjust for such bias by including the nuisance variables in the linear predictor. There is currently no way to do this for random forests, but there has been recent interest in hybridizing tree models with GLMs (Chen et al. 2007, Zeileis et al. 2008, Yu et al. 2010). Further extension of such hybrid models to the random forest paradigm might allow adjustment for nuisance variables.

Random forests allow for estimation of uncertainty in the species response, and indeed this is essential for combining information across species according to R^2. However, it is less obvious how to assess uncertainty in the compositional turnover functions, or indeed in the resulting patterns of biological composition. A pragmatic approach could be to compare the pairwise Euclidean site distances in the biologically transformed space with Bray-Curtis distances, using a standard measure such as the stress metric resulting from monotone regression (as in nonmetric MDS). Alternatively, the sites in biologically transformed space could be mapped by Procrustes rotation to their locations in a nonmetric ordination of the Bray-Curtis distances. These measures would allow any methods able to produce them to be compared objectively.

Gradient forests were developed as a natural extension of variable importance analysis using random forests on marine benthic survey data. The extension allows one to quantify the degree of community change along the predictor gradient. Moreover, because the units of change are not survey dependent, the method allows disparate surveys to be synthesized. Gradient forests inherit from random forests their strengths (flexibility, implicit model selection, accuracy) and their weaknesses (assumption of Gaussian errors). Gradient forests should therefore be considered as complementary to other methods such as GDM and constrained CCA. The method has applicability outside of ecology in other domains, such as epidemiology, where the aim is to characterize multivariate data in terms of environmental predictors and to explore where along these gradients important changes are occurring.

Acknowledgments

The authors thank Matthew C. Fitzpatrick and Falk Huettmann for useful reviews that improved the manuscript, and Craig Brown and Bill Venables for reviewing an earlier draft. This work was supported by the Sloan Foundation under the Census of Marine Life; by the Science Sector of the Department of Fisheries and Oceans Canada through its Ecosystem Research Initiative (S. J. Smith); and by the CSIRO Wealth from Oceans Flagship and the Commonwealth Environment Research Facilities program Marine Biodiversity Hub (N. Ellis and C. R. Pitcher).

Literature Cited

Austin, M. P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling 157:101–118.

Breiman, L. 2001. Random forests. Machine Learning 45:5–32.

Chen, J., K. Yu, A. Hsing, and T. Therneau. 2007. A partially linear tree-based regression model for assessing complex joint gene-gene and gene-environment effects. Genetic Epidemiology 31:238–251.

Cutler, D. R., T. C. Edwards, Jr., K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. 2007. Random forests for classification in ecology. Ecology 88:2783–2792.

De'ath, G. 2002. Multivariate regression trees: a new technique for modeling species–environment relationships. Ecology 83:1105–1117.

Elith, J., and J. Leathwick. 2009. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics 40:677–697.

Evans, J. S., and S. A. Cushman. 2009. Gradient modeling of conifer species using random forests. Landscape Ecology 24:673–683.

Ferrier, S., and A. Guisan. 2006. Spatial modelling of biodiversity at the community level. Journal of Applied Ecology 43:393–404.

Ferrier, S., G. Manion, J. Elith, and K. Richardson. 2007. Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. Diversity and Distributions 13:252–264.

Friedman, J. 1991. Multivariate adaptive regression splines. Annals of Statistics 19:1–67.

Grömping, U. 2009. Variable importance assessment in regression: linear regression versus random forest. American Statistician 63:308–319.

Guisan, A., and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8:993–1009.

Guisan, A., and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135:147–186.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning: data mining, inference, and prediction. Second edition. Springer, New York, New York, USA.

Kendall, M. 1938. A new measure of rank correlation. Biometrika 30:81.

Knudby, A., A. Brenning, and E. LeDrew. 2010. New approaches to modelling fish–habitat relationships. Ecological Modelling 221:503–511.

Lawler, J. J., D. White, R. P. Neilson, and A. R. Blaustein. 2006. Predicting climate-induced range shifts: model differences and model reliability. Global Change Biology 12:1568–1584.

Lawler, J. J., Y. F. Wiersma, and F. Huettmann. 2011. Using species distribution models for conservation planning and ecological forecasting. Pages 271–290 *in* C. A. Drew, Y. F. Wiersma, and F. Huettmann, editors. Predictive species and habitat modeling in landscape ecology. Springer, New York, New York, USA.

Leathwick, J., J. Elith, M. Francis, T. Hastie, and P. Taylor. 2006. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. Marine Ecology Progress Series 321:267–281.

Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. R News 2:18–22.

Murray, K., and M. Conner. 2009. Methods to quantify variable importance: implications for the analysis of noisy ecological data. Ecology 90:348–355.

Nicodemus, K. K., J. D. Malley, C. Strobl, and A. Ziegler. 2010. The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinformatics 11:110.

Peters, J., B. Baets, N. Verhoest, R. Samson, S. Degroeve, P. Becker, and W. Huybrechts. 2007. Random forests as a tool for ecohydrological distribution modelling. Ecological Modelling 207:304–318.

Pitcher, C., et al. 2007. Seabed biodiversity on the continental shelf of the Great Barrier Reef World Heritage Area. Technical report, AIMS/CSIRO/QM/QDPI Final Report to CRC Reef Research Centre. AIMS/CSIRO/QM/QDPI CRC Reef Research Task Final Report. CSIRO Marine and Atmospheric Research, Cleveland, Australia. http://www.reef.crc.org.au/resprogram/programC/seabed/final-report.htm

Pitcher, C., W. Venables, N. Ellis, I. McLeod, M. Cappo, F. Pantus, M. Austin, P. Doherty, and N. Gribble. 2002. GBR seabed biodiversity mapping project: Phase 1 report to CRC-Reef. Technical report, CSIRO/AIMS/QDPI Report. CSIRO Marine and Atmospheric Research, Cleveland, Australia. http://www.reef.crc.org.au/resprogram/programC/seabed/Seabedphase1rpt.htm

Poiner, I., J. Glaister, C. Pitcher, C. Burridge, T. Wassenberg, N. Gribble, B. Hill, S. Blaber, D. Milton, D. Brewer, and N. Ellis. 1998. Final report on effects of trawling in the far northern section of the Great Barrier Reef: 1991–1996. CSIRO Division of Marine Research, Cleveland, Australia.

Prasad, A. M., L. R. Iverson, and A. Liaw. 2006. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems 9:181–199.

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

Silverman, B. W. 1986. Density estimation. Chapman and Hall, London, UK.

Strobl, C., A. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. 2008. Conditional variable importance for random forests. BMC Bioinformatics 9:307–317.

Ter Braak, C. 1994. Canonical community ordination. Part I: Basic theory and linear methods. Ecoscience 1:127–140.

Yu, K., W. Wheeler, Q. Li, A. Bergen, N. Caporaso, N. Chatterjee, and J. Chen. 2010. A partially linear tree-based regression model for multivariate outcomes. Biometrics 66:89–96.

Zeileis, A., T. Hothorn, and K. Hornik. 2008. Model-based recursive partitioning. Journal of Computational and Graphical Statistics 17:492–514.

Supplemental Material

Appendix A

Implementation of conditional permutation (*Ecological Archives* E093-015-A1).

Appendix B

Derivation of Eq. 6 (*Ecological Archives* E093-015-A2).

Supplement 1

Exploration of conditional importance using a synthetic data set. This document is provided as part of the documentation of the R package `extendedForest` (*Ecological Archives* E093-015-S1).

Supplement 2

Extended analysis of the Great Barrier Reef data set. This document is provided as part of the documentation of the R package `gradientForest` (*Ecological Archives* E093-015-S2).