Journal of Vegetation Science 17: 819-830, 2006
© IAVS; Opulus Press Uppsala.

Non-parametric habitat models with automatic interactions

McCune, Bruce

Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331 USA;
E-mail Bruce.McCune@science.oregonstate.edu

Abstract

Questions: Can a statistical model be designed to represent more directly the nature of organismal response to multiple interacting factors? Can multiplicative kernel smoothers be used for this purpose? What advantages does this approach have over more traditional habitat modelling methods?

Methods: Non-parametric multiplicative regression (NPMR) was developed from the premises that: the response variable has a minimum of zero and a physiologically-determined maximum, species respond simultaneously to multiple ecological factors, the response to any one factor is conditioned by the values of other factors, and that if any of the factors is intolerable then the response is zero. Key features of NPMR are interactive effects of predictors, no need to specify an overall model form in advance, and built-in controls on overfitting. The effectiveness of the method is demonstrated with simulated and real data sets.

Results: Empirical and theoretical relationships of species response to multiple interacting predictors can be represented effectively by multiplicative kernel smoothers. NPMR allows us to abandon simplistic assumptions about overall model form, while embracing the ecological truism that habitat factors interact.

Keywords: Habitat model; Kernel smoothing; *Larix occidentalis*; *Lobaria*; Local model; Non-parametric multiplicative regression; NPMR; *Picea glauca*; *Picea mariana*; Regression; Species response surface; Spruce.

Abbreviations: GAM = Generalized additive model; GLM = Generalized linear model; NPMR = Non-parametric multiplicative regression.

Introduction

Eugene Odum (1971) stated Shelford's 'law' of tolerance as follows: "The presence and success of an organism depend upon the completeness of a complex of conditions. Absence or failure of an organism can be controlled by the qualitative or quantitative deficiency or excess with respect to any one of several factors which may approach the limits of tolerance for that organism."

This straightforward and manifest statement has defied a correspondingly simple and general statistical representation with the traditional tools used by ecologists. These tools are unnecessarily constrained by additivity of model terms and a limited array of functional forms. The problems in applying simple linear and logistic models to species responses to multiple interacting predictors have been clearly stated (Kaiser et al. 1994; Huston 2002; Cade & Noon 2003). This paper demonstrates how non-parametric multiplicative regression (NPMR) provides a simple, effective solution to the problem of representing empirical species response surfaces in a multidimensional niche space. Interactions among predictors are accommodated automatically and the overall form of the response surface need not be specified.

Habitat models describe how variation in species performance relates to one or more predictors. Measures of species performance include presence-absence, abundance, physiological rates, and demographic parameters (e.g. nesting success). Commonly used predictors include environmental variables (including biotic variables), site characteristics, time since disturbance, and other descriptors of disturbance regime.

Habitat models can take both conceptual and statistical forms. Conceptual habitat models have been formative in ecological theory, for example, the Hutchinsonian niche, an *n*-dimensional hypervolume (Hutchinson 1957), and Whittaker's diagrams of species responses to environmental gradients (e.g. Whittaker 1956). Statistical habitat models have been made for many species of particular concern. These models describe the important factors underlying a species' distribution and abundance, inform management decisions for rare or threatened species, and allow comparison of probable outcomes of alternative management strategies.

Ecologists readily grant complex species response functions in theory, but often use simplistic forms (linear or logistic; Rushton et al. 2004) that cannot hope to capture the complexity of a species in relationship to its habitat (Kaiser et al. 1994; Heglund 2002; Huston 2002). These simple forms do not accommodate the widely-accepted view that species have hump-shaped responses to environmental gradients. These hump-shaped response curves are implicit in Shelford's law of tolerance. Of course non-linear transformations of environmental gradients, such as Gaussian logistic regression (Huisman et al. 1993), can render hump-shaped response functions within the framework of linear models, but still constrain the model to particular functional forms. Generalized additive models (GAMs, e.g. Maravelias & Reid 1997; Mysterud et al. 2001; Begg & Marteinsdottir 2002) offer a more flexible approach by combining smoothing functions, though interactions must still be modelled explicitly (e.g. Ciannelli et al. 2004). Likewise, multivariate adaptive regression splines (MARS) provide a flexible way to fit complex surfaces, including the possibility of multiplicative basis functions that should readily accommodate interactions (Friedman 1991; Hastie et al. 2001).

Multiplicative kernel smoothers can also be used to represent the complexity of species responses to multiple interacting predictors. This kind of model provides two important advantages over other approaches: it automatically represents predictor interactions by combining predictors multiplicatively, such that the effect of one predictor can covary in a complex way with other predictors, and it requires no assumptions about the overall shape of the response surface. It accepts that complex interactions may result in the response surface in one part of the predictor space having no simple functional relationship to responses elsewhere in the predictor space. The chief disadvantages of multiplicative kernel smoothers are that the response surface must be fitted with a computationally intensive trial-and-error method and the results do not include an equation relating the response to the predictors. Instead, interpretation must rely on graphical visualization, measures of fit, and sensitivity analysis for individual predictors. Using these models in an explorative way informs, but does not preclude, parametric modeling. Having explored the response surface in a multidimensional space, one can then sensibly choose an appropriate functional form and proceed with non-linear regression or a generalized linear model, if desired.

The extension of kernel smoothing multiplicatively into many dimensions, and its combination with cross-validation, provides an easy, intuitive way to fit parsimoniously a species response surface to multiple predictors. A few papers in the literature have used similar smoothers, but differ in important ways from the method proposed here. Gignac et al. (1991a, b) generated 3D response surfaces for species abundance from gridded abundance data along environmental gradients, using distance-weighted means. Limitations to that approach included an arbitrarily selected (rather than optimized) search radius, arbitrary treatment of zeros and outliers, and no cross-validation. Locally-weighted smoothing (or regression) using the LOWESS method has been applied to habitat models in two and three dimensions (Huntley et al. 1989, 1995). These models were effective but had several drawbacks: (1) restriction to two or three pre-selected predictors instead of a conducting a free search for the best model using an indefinite number of variables from a pool of available predictors, (2) choosing the smoothing parameter for each predictor arbitrarily, rather than simultaneously optimizing it in all dimensions, and (3) fitting the model at fixed intervals within the plane of predictors, with linear interpolation between intervals, rather than fitting the model for each data point.

Multiplicative models

Habitat models and other species response models are most often additive, including those created by the usual least-squares multiple linear regression, generalized linear models (GLMs, including logistic regression), and generalized additive models (GAMs). In the latter two, using a log link function makes the log(mean) an additive function of the predictors, but the mean is a multiplicative function of the predictors. In GLMs and GAMs, interactions are accommodated by terms including more than one predictor. This paper proposes a form of non-parametric multiple regression in which the response is estimated from a multiplicative combination of all predictors.

Consider an extreme hypothetical example where we place an experimental population in the Antarctic, with some amount of food and shelter. Our species response variable, y, is the reproductive rate, x_1 is food, and x_2 is shelter. The model $y = b_1x_1 + b_2\mathrm{x}_2 + b_0$ says that the reproductive rate is increased by the availability of food and shelter, and that increasing either of these alone can increase the reproductive rate. The simple additive model comes to the erroneous conclusion that reproductive rates will be high if the population has abundant food but no shelter. Likewise, the model says the population will reproduce if given lots of shelter but no food. Interaction terms in linear models represent the curved response surfaces that we expect from interacting predictors, but these surfaces are relatively simple functions (in the simplest case, hyperbolic paraboloids) applied over the whole predictor space. We have no reason to expect that surfaces representing interactions will take this limited range in form.

The simplest GAM for this problem would be: $g[\mu(x)] = \alpha + f_1(x_1) + f_2(x_2)$, where f is an unspecified smooth function and g is a link function of μ (for example $g(\mu) = \mu$ is the identity link, or $g(\mu) = \log(\mu)$ is a log link). With the latter, the model is in one sense multiplicative, in that $\mu = e^{\alpha} \cdot e^{f_1(x_1)} \sum e^{f_2(x_2)}$, but it is still additive in that the effects of food and shelter are added independently to estimate the log of the reproductive rate. In biological terms this means that the effect of food on $\log(\mu(x))$ does not depend on shelter and vice-versa. It makes only slightly more sense for the log of the mean response to be an additive result of food and shelter than for the mean to be an additive result. On the other hand, if our interest is in μ rather than $\log(\mu)$, then in the multiplicative form, if x_1 is highly unfavourable, differences in x_2 will have little effect on μ. A more direct GAM for the Antarctic example is $g[\mu(x)] = f(x_1, x_2)$, with the constant set to zero (no reproduction) and using a non-parametric smoothing function combining x_1 and x_2 interactively and multiplicatively. GAM with interaction terms is challenging (e.g. see Ciannelli et al. 2004), out of reach for most of the grassroots practitioners of habitat modeling. A simpler, more intuitive approach uses multiplicative kernel smoothers to allow the effect of each predictor to depend on the value of other predictors, without needing to specify those interactions. This is non-parametric multiplicative regression (NPMR). The method is implemented in easy-to-use software in HyperNiche (McCune & Mefford 2004).

Non-parametric models

Ecological theory does not reliably inform us as to how a particular species responds to a particular habitat factor, much less to combinations of interacting factors. In general, however, we expect species responses to multiple factors to be complex, including non-linear, asymmetric, and multimodal responses. While these challenges are readily addressed in models of species response to a single factor, the problem becomes exponentially more difficult as the dimensionality of the data increases. This is the 'curse of dimensionality' (Bellman 1961).

Yet assumptions about the shape of a species response to environmental variables are central to any predictive parametric model (Austin 2002). NPMR circumvents this; predictive modeling can be effective without making any assumptions about the shapes of species responses to ecological factors or to their interactions.

Non-parametric multiplicative regression

This section derives the logic for using a multiplicative kernel smoother from basic biological principles. The goal is to provide an intuitive biological basis for the statistical approach. The method is then compared with additive models in three examples.

We seek to describe or predict an organism's performance in relationship to its environment. An organism's 'environment' includes not just abiotic factors, but also its biotic environment, including competition, predation, and disease. Assume further that we are measuring organismal or population performance in relation to environmental variables, and that performance has a minimum of zero and increasing into the positive real numbers. Some examples are population density, areal cover, rate of reproduction, and rate of respiration. The following treatment pertains to this class of variables representing organismal performance.

Axioms

1. Performance of a species has a maximum set by physiology and morphology. The maximum is fuzzy, rather than a set value, because of genetic variation among individuals. This axiom refers to short, ecological time scales, excluding the possibility of evolution.
2. As environmental factors weaken performance, a population collapses (its organisms die) and performance is minimized at zero.
3. Organisms respond simultaneously to many environmental factors.
4. If any single factor or combination of factors is intolerable, then the organisms in a population die and performance is zero, regardless of the values of other environmental variables.

Definitions

Environmental space: A multidimensional space, the m coordinates of the space defined by the m measured environmental variables, including both biotic and abiotic variables. A particular point $\mathbf{v}$ in this environmental space is thus the vector defining the value of each of the m variables: $[v_1, v_2 \ldots v_m]$, where v can be any real number. In habitat modeling, the environmental space is used as the predictor space. It can be used as an operational definition of a niche space.

Species performance (y_v): The performance of an individual or population of individuals at point $\mathbf{v}$ in environmental space, where $y_v \geq 0$.

Environmental measurement (x_{ij}): The value of predictor j for data point i.

Interaction: Response to predictor j depends on values of other predictors.

Describing species performance in relationship to environment

Usually we are interested in species performance over a range of multiple environmental factors. We collect a sample of species performance, along with measured environmental factors (predictors) for n data points in the environmental space. We build a response surface of $\mathbf{y}$ from its relationship to m predictors in $\mathbf{X}$.

$$y_i = f\left(x_1, x_2, \ldots, x_m\right) \tag{1}$$

A multiplicative kernel smoother allows all predictors to interact: the effect of each x_j can vary with the others, and the response surface in one region of environmental space need not bear any relationship to the response surface in other regions of that space.

We need to use data from near target point $\mathbf{v}$ to help estimate the response at that point because (1) all of our measurements have error, (2) we need to interpolate between data points, and (3) we wish to borrow information from nearby points to help estimate a response at a particular target point, because in most data sets no two cases will occupy the same point in environmental space.

The tolerance (s_j) of a species to a continuous predictor j defines how broadly we borrow information from nearby areas in the predictor space, while attempting to estimate the value at a target point. If a species is broadly tolerant to that factor, then we use information from a large neighbourhood of data points, while a species with narrow tolerance to that factor is better represented by using only data points that are close to the target point in the predictor space. A smooth way of representing the neighbourhood of the target point is to use a Gaussian weighting function centred on the target point $\mathbf{v}$, the function expressing the weight (w) given to each sample point i in estimating the response at point $\mathbf{v}$, based on the difference between x_i and $\mathbf{v}$, scaled by the standard deviation (tolerance) to that predictor:

$$w_{ij} = e^{-\frac{1}{2}\left[\left(x_{ij}-v_j\right)/s_j\right]^2} \tag{2}$$

This univariate weighting function (the kernel; Bowman & Azzalini 1997; Hastie et al. 2001) specifies the weight (w_{ij}) for an observation of a single predictor j at sample unit i, drawn from the matrix $\mathbf{X}$ of n observations for m predictors. This weighting can take various forms, but this Gaussian function is simple and intuitive. The standard Gaussian probability density function was modified so that the peak height is always one and the area under the curve varies. Note that use of Gaussian weights does not limit the global model to Gaussian forms; the weighting function is largely independent of the global model. Other weightings can be used, for example, a uniform weight of one within an observational window of a specific optimized width, and zero weight outside the window (McCune et al. 2003), but this gives a relatively rough response surface.

With categorical predictors a different approach is needed. The simplest method is to apply binary weights: an observation is given full weight for a given predictor if x_{ij} and the target point were assigned to the same category (w_{ij} = 1 if x_{ij} = v_j); otherwise, that observation is given zero weight in estimating the response at point $\mathbf{v}$ (w_{ij} = 0 if $x_{ij} \neq v_j$). An intriguing refinement would be to allow fuzzy categories, such that weights between zero and one are allowed for categorical predictors.

The multiplicative local mean estimator

We can then estimate the response y at target point $\mathbf{v}$ as:

$$\hat{y}_v = \frac{\sum_{i=1}^{n} y_i \left(\prod_{j=1}^{m} w_{ij} \right)}{\sum_{i=1}^{n} \left(\prod_{j=1}^{m} w_{ij} \right)} \tag{3}$$

This is a local mean estimator extended multiplicatively to m dimensions. In words, the estimate of the response is an average of the observed values, each value weighted by its proximity to the target point in the predictor space, the weights being the product of weights for individual predictors. The model allows interactions, because weights for individual predictors are combined by multiplication rather than addition. A key biological feature of the model is that failure of a population with respect to any single dimension j of the predictor space results in failure at point i, because the product of the weights for point i is zero if any $w_{ij} = 0$.

If point $\mathbf{v}$ is one of the n sample points, from which the response is estimated, we can reduce overfitting by excluding point i when it is the same as point $\mathbf{v}$:

$$\hat{y}_v = \frac{\sum_{i=1, i \neq v}^{n} y_i \left(\prod_{j=1}^{m} w_{ij} \right)}{\sum_{i=1, i \neq v}^{n} \left(\prod_{j=1}^{m} w_{ij} \right)} \qquad (4)$$

The notation $i \neq v$ indicates that if the target point v is one of the calibration data points, then it is excluded from the basis for the estimate of y_v. This is the fundamental equation for cross-validated local mean NPMR. Local linear NPMR is the same except that it is based on a locally weighted linear relationship (App. 1), rather than a local mean. The local model can take other forms, such as logistic or more complex polynomials.

With multiple categorical predictors, an observation is given full weight only if the target point matches all categorical predictors, otherwise the observation is given zero weight. With mixed categorical and quantitative predictors, the weights are multiplied as usual.

Multiplicative kernel smoothers are not new (e.g. Bowman & Azzalini 1997; Hastie et al. 2001), but this special form is noteworthy for ecologists modeling abundance or other performance variables with zero as a natural lower bound. It simply represents the axiom that organisms must simultaneously meet all environmental challenges or die.

NPMR is, therefore, a particular class of smoothing functions, in which an estimate for a particular target point in predictor space is made by combining information from observations nearby in the predictor space. The closer a data point is to the target point, the more weight is given to information from that point. How rapidly weights diminish with distance can be tuned for each predictor with a smoothing parameter, in this case the standard deviation of the Gaussian curve ('tolerance' s_j to a predictor j). Selecting a large standard deviation is comparable to having a broad window; conversely a small standard deviation gives appreciable weight only to observations that are very close to the target point in the predictor space. The Gaussian function is scaled to the predictor by arbitrarily setting one standard deviation equal to one sixth of the range of the predictor, thus representing a Gaussian curve with ± 3 standard deviations over the range of the predictor. Then for each predictor, s_j is set to maximize fit, subject to model fitting constraints described below.

For every point estimate of the response variable one can calculate a neighbourhood size (n_v*), the amount of data bearing on that particular estimate:

$$n_v^* = \sum_{i=1, i \neq v}^{n} \left(\prod_{j=1}^{m} w_{ij}^* \right) \qquad (5)$$

where $0 < n^* \leq n$ for a Gaussian kernel. If $n^* = 0$, as is possible with some other kinds of kernels, then no estimate is possible for that point because no data apply to it. Setting a minimum n^* required for an estimate protects against estimating a response in a region of the predictor space with insufficient data.

Model fitting and evaluation

Fitting an NPMR model requires simultaneous selection of predictors and their tolerances from a pool of available predictors, so as to maximize a measure of fit while satisfying criteria for parsimony. This demands an iterative search through a potentially enormous number of possible models, which is accomplished with the software HyperNiche (McCune & Mefford 2004). Variables are added in forward stepwise fashion, at each step making a grid search of variables and their tolerances. Variables already in the model are simultaneously evaluated for removal or change in tolerances with the addition of a new variable.

Using leave-one-out cross-validated statistics for fit reduces overfitting and results in more realistic error estimates. Incorporating the cross-validation into not just model evaluation but also model fitting expedites the search for a parsimonious model. For quantitative data, model fit can be evaluated by the size of the residual sum of squares (*RSS*) in relationship to the total sum of squares (*TSS*):

$$\mathrm{x}R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}_i \right)^2} \qquad (6)$$

This 'cross R^2' differs from the traditional R^2 because data point i is excluded from the basis for estimating $\hat{y}_i$. Consequently, with a weak model, it is not uncommon for RSS > TSS and x$R^2 < 0$. This method is similar to the use of *G*-values (Agterberg 1984; Gotway et al. 1996; Guisan & Zimmerman 2000).

For binary response data (presence-absence) a measure of fit was sought that could be applied to any method of estimating likelihood of occurrence and would avoid the arbitrary conversion of continuous estimates of probability of occurrence into a statement of 'present' or 'absent' (Fleishman et al. 2003). The xR^2 is inappropriate in this case because the goal is to estimate probabilities from presence-absence data, rather than producing estimates that exactly match the data. Log likelihood ratios met these criteria, expressing model improvement over a 'naïve model'. A naïve model is simply that our best estimate of the probability of encountering a species in a study area is the average frequency of occurrence of that species in the data. The ratio of the likelihood of the observed values ($\mathbf{y} = y_1, y_2, \ldots y_n$) under the fitted model (M_1) to the likelihood of the result under the naïve model (M_2) is given by:

$$B_{12} = \frac{p\left(\mathbf{y}|M_1\right)}{p\left(\mathbf{y}|M_2\right)} \tag{7}$$

where

$$p\left(\mathbf{y}|M\right) = \prod_{i=1}^{n} \hat{y}_i^{y_i}\left(1 - \hat{y}_i\right)^{1-y_i} \tag{8}$$

and $\hat{y}_i$ corresponds to the fitted values for the likelihood of occurrence under each model, M_j, $j = 1,2$.

Formal hypothesis testing with log-likelihood ratios requires that the parameters for one model be nested within the other and incorporates the difference in degrees of freedom between the two models. $\mathrm{Log}_{10}B$ is applied here, however, as a descriptive statistic in the sense of 'weight of evidence', similar to a Bayes factor (Kass & Raftery 1995), rather than a formal hypothesis test. Log*B* differs from a Bayes factor in that a Bayes factor is based on the marginal distribution of **y** given the prior model (the naïve model in this case), while log*B* is a simple log likelihood ratio for the two models, inverted so that as the weight of evidence increases, log*B* increases. Values of log*B* reported here from NPMR models are based on cross-validated estimates from the M_1 using a leave-one-out strategy. Log*B* can be interpreted as the ratio of the likelihood of cross-validated estimates from the fitted model to estimates from the naïve model expressed in powers of ten. Log*B* is negative when cross-validated estimates from the fitted model are worse than the naïve model. The same rationale can be applied to the difference between log*B* values calculated for each of two competing models of interest. Because log*B* is unbounded, it can be quite large when a strong relationship is modelled with a very large data set. The average contribution of a sample unit to log*B*, $10^{(\log B)/n}$, can be used to describe the strength of relationship, independent of sample size.

Controlling flexibility and parsimony

Selecting the best model from the multitude of models with many predictors can lead to overfitting the data. Overfitting is particularly problematic with small data sets, a large number of predictors relative to the sample size, or clumped sampling from the predictor space. The NPMR models presented here control overfitting in several ways simultaneously: built-in cross validation during model fitting, a flexibility control, and setting parsimony criteria to control inclusion of predictors. Each control restricts a different aspect of overfitting. Flexibility of the response surface can be controlled by setting a minimum acceptable average neighbourhood size, N^*, where N^* is the average of n_i^*. Stiff curves (large N^*) are needed with small data sets or clumped data distribution along important habitat dimensions. More flexibility is allowable with large high-density data sets. A reasonable starting point is to set minimum $N^* = 0.05(n)$.

Parsimony in number of predictors is partly controlled by cross-validation and partly by the minimum N^*. It can be further controlled by setting an improvement criterion, expressed as a percentage improvement in model fit when a new predictor is added. This criterion is particularly important for parsimonious models with large data sets. One can also set a minimum data:predictor ratio, an effective criterion for small data sets. For quantitative responses, the data:predictor ratio is the number of sample units divided by the number of predictors in the model. For binary responses, the data:predictor ratio is the number of observations in the least represented category (presences or absences) divided by the number of predictors in the model. Some suggest a minimum ratio of 10 for binary data (Harrell et al. 1996). Using all four parsimony criteria simultaneously is effective because each controls different aspects of overfitting. With a forward stepwise search, as soon as any one of the criteria cannot be met, the search for additional predictors stops.

Sensitivity analysis

Here 'sensitivity analysis' evaluates the relative importance of particular predictors within a model. This is particularly important in non-parametric regression, because we have no fixed coefficients or slopes to compare.

A general way to evaluate the importance of individual predictors is to nudge up and down observed values for individual predictors, and measure the resulting change in the estimated response for that point. By accumulating those sensitivities across all data points, one can evaluate the sensitivity of the model to each predictor. The greater the sensitivity, the more influence that variable has in the model.

The change in the response can be measured as a fraction of the observed range of the response variable, $|y_{\max} - y_{\min}|$. Scaling the differences in response and differences in predictors to their respective ranges allows a sensitivity measure (Q_j for predictor j) that is a ratio, independent of the units of the variables:

$$Q_j = \frac{\sum_{i=1}^{n}\left|\hat{y}_{i+} - \hat{y}_i\right| + \sum_{i=1}^{n}\left|\hat{y}_{i-} - \hat{y}\right|}{2n\left|y_{\max} - y_{\min}\right|\Delta} \qquad (9)$$

where $\hat{y}_{i+}$ and $\hat{y}_{i-}$ are the estimates of the response variable for case i, having increased or decreased, respectively, the predictor by an arbitrarily small value Δ (here $\Delta = 0.05$, i.e. 5% of the range of predictor j). With the formula above, a $Q = 1.0$ means that on average, nudging a predictor results in a change in response of equal magnitude; $Q = 0.0$ means that nudging a predictor has no detectable effect on the response. Sensitivity can also be calculated in a similar way from the root mean square of the differences, rather than the absolute differences.

Statistical significance

Statistical significance of a model derived by NPMR can be evaluated by a randomization test obtained when the vector of response values is shuffled, randomly reassigning their relationships to the predictor matrix. This is a simple, readily justified approach for multiple regression, though other kinds of randomizations are needed for particular experimental designs (Manly 1997, pp. 156-157). Bootstrap methods could also be applied, but more assumptions are made and caution is needed with small sample sizes (Westfall & Young 1992, p. 142-143; Manly 1997). The following procedure tests the null hypothesis that the fit of the selected model is no better than could be obtained by chance alone, given an equal number of predictors selected from the same pool of variables. The relationship between predictors and the response variable is destroyed by shuffling the values of the response variable, then repeating the same model fitting procedure as used with the unshuffled data, then calculating the fit. The same pool of predictors is used, but with the additional constraint that the model with the randomized data should have no more predictors than the model being evaluated. After repeating this many times, the proportion of randomization runs that results in an equal or better fit is used as the p-value for the test.

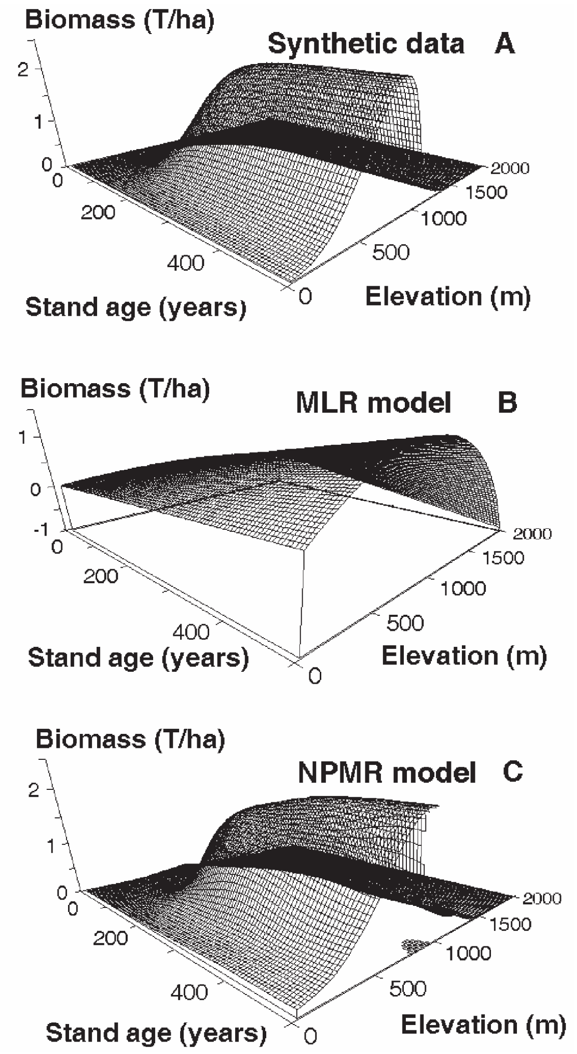

Fig. 1.A. Hypothetical response surface from combined Gaussian and sigmoid functions (Eq. 10), representing biomass of epiphytic nitrogen-fixing lichens in relation to elevation and stand age on the west slope of the Cascade Range. **B.** Multiple linear regression (MLR) model fit to a random sample of the hypothetical response surface. The model has two main predictors, a quadratic elevation term, and interactions between stand age and the two elevation terms (adj. $R^2 = 0.43$). **C.** Local mean NPMR model fit to the same random sample (x$R^2 = 0.97$). The small dropout from the surface had insufficient local data to fit the model.

Example with synthetic data

Fitting models to data sets with known underlying structure provides insights into the performance of different modeling approaches. The following example combines simple responses to two predictors. NPMR models were fit with HyperNiche (McCune & Mefford 2004); logistic and linear models were fit with SPSS ver. 11.5; GAM with S-PLUS ver. 6.2. The response function simulates the known response of biomass of nitrogen-fixing lichens (primarily *Lobaria oregana*) in western Oregon to elevation and stand age.

Biomass of *Lobaria* has a sigmoid response to stand age. Slow to establish in clear-cut or burned forests, *Lobaria* increases on optimum sites to a plateau averaging about 1.5 T/ha by about age 200 years (McCune 1993; Berryman 2002). *Lobaria* has the classic hump-shaped response to elevation, being rare at low elevations and dropping out completely at high elevations, but abundant between 500-1000 m (McCune et al. 2003).

A noiseless response surface was designed that incorporates these two predictors, elevation and stand age. The surface multiplies a sigmoid function by a Gaussian function (Fig. 1A):

$$y = \left(\frac{e^{-a_1 + b_1 x_1}}{1 + e^{-a_1 + b_1 x_1}} \right) \left(e^{-a_2 (x_2 - b_2)^2} \right) \qquad (10)$$

The first term is the sigmoid response to stand age (x_1) and the second term is the humped response to elevation (x_2). I then selected parameters to give a reasonable surface for elevation in meters and stand age in years ($a_1 = 5$, $b_1 = 0.03$, $a_2 = .00001$, $b_2 = 700$), and multiplied by 1.5^2 to set $y_{\max} = 1.5$ T/ha. This surface was randomly sampled with 200 points and the response at each point calculated with this equation. With this model, if either stand age or elevation is unfavourable, then the species is absent or nearly so.

A series of least-squares multiple linear regression (MLR) models illustrates the difficulty of representing even this simple two-predictor system with some traditional statistical tools, while the surface is easily fit by GAM and NPMR. The hazards of the more commonly used linear models are well known by statisticians but often not appreciated by ecologists. The simplest and most naïve model relating biomass to elevation and stand age is a MLR with two terms and a constant: $y = b_1x_1 + b_2x_2 + b_0$. The fit of this model, a tilted plane, to the data is expectedly poor (adj. $R^2 = 0.210$), yet the terms for both stand age and elevation differ significantly from zero. With an interaction term, $y = b_1x_1 + b_2x_2 + b_3x_1\,x_2 + b_0$, but the fit is still poor (adj. $R^2 = 0.245$). A residual plot reveals the hump-shaped relationship with elevation that is not being fit. Accommodating this with a quadratic term for elevation and the interaction of the quadratic term with stand age, we have
$y = b_1x_1 + b_2x_2 + b_3x_2^2 + b_4x_1\,x_2 + b_5x_1\,x_2^2 + b_0$,
yielding an improved, yet weak model (adj. $R^2 = 0.428$). The resulting surface (Fig. 1B) starts to resemble our known underlying model, but still leaves much to be desired, as shown by the residuals plotted against stand age and elevation (App. 2).

In contrast, both GAM and non-parametric multiplicative regression (NPMR) with a local mean and Gaussian kernel easily captured almost all of the variation in the response variable using the two predictors, elevation and stand age. The GAM (Poisson family, log link, spline smoother) readily captures the response surface ($R^2 > 0.99$), because the log link effectively decomposes the two multiplied underlying functions and the smoothing splines capture their shapes. NPMR closely reproduced the original response surface (Fig. 1C; xR^2 = 0.975) and had no major problems in the residuals (App. 2). Using other local models with NPMR made little difference in fit (local mean with rectangular kernel, xR^2= 0.983; local linear with Gaussian kernel, xR^2= 0.988). The local linear model resulted in slightly higher sensitivity to stand age ($Q = 0.37$ vs. $Q = 0.24$) and slightly lower sensitivity to elevation ($Q = 1.14$ vs. $Q = 1.35$) than did the local mean.

Example with real data

A second example illustrates a model fitting binary (presence-absence) data for the tree *Larix occidentalis* (Anon. 1999) to a suite of climate variables (Daly et al. 1994): mean January, July, and annual temperatures; mean January, July, and annual precipitation; mean relative humidity in January and July; and 'wetdays', the mean number of wet days in a year. *Larix occidentalis*, endemic to western North America, has a fairly small geographic range (Fig. 2A). Presumably its range would be more vulnerable to climate change than many other tree species, as it appears to have relatively tight climatic tolerances. The grids of climatic data and associated distribution data were randomly sampled with 2500 points between 43-49 °N and between 112-122 °W, which includes the entire distribution of *L. occidentalis* in the U.S. Similar climatic data were not available for the Canadian portion of the range of the species.

NPMR using a local mean, Gaussian weights, and a minimum $N^* = 100$ found a two-predictor model with log$B$ = 261. The strongest predictors were wetdays (s = 9.1 days/year, $Q = 1.54$) and mean annual temperature (s = 0.67 °C, $Q = 1.10$). The need for a hump-shaped model appears in the distribution of occupied points in this 2D slice through the sample space (Fig. 2B). The best three-predictor model improved the fit to log$B$ = 314. The two predictors remained with their previously stated tolerances and average relative humidity in July (s = 1.9%, $Q = 0.65$) was added. Including relative humidity decreased the sensitivity to wetdays and temperature to 1.22 and 1.06, respectively.

Because the sample size was so large, the cross-validation penalty to log$B$ was tiny. This resulted in an asymptotic log$B$ as predictors were added, rather than the ultimate decline in cross-validated fit expected of small data sets.

The response surface for the two-predictor model showed a distinct interaction between wetdays and temperature, as indicated by the diagonal ridge in the response surface (Fig. 2D). The ridge was asymmetric, with slopes varying from steep on the warm-wet side to

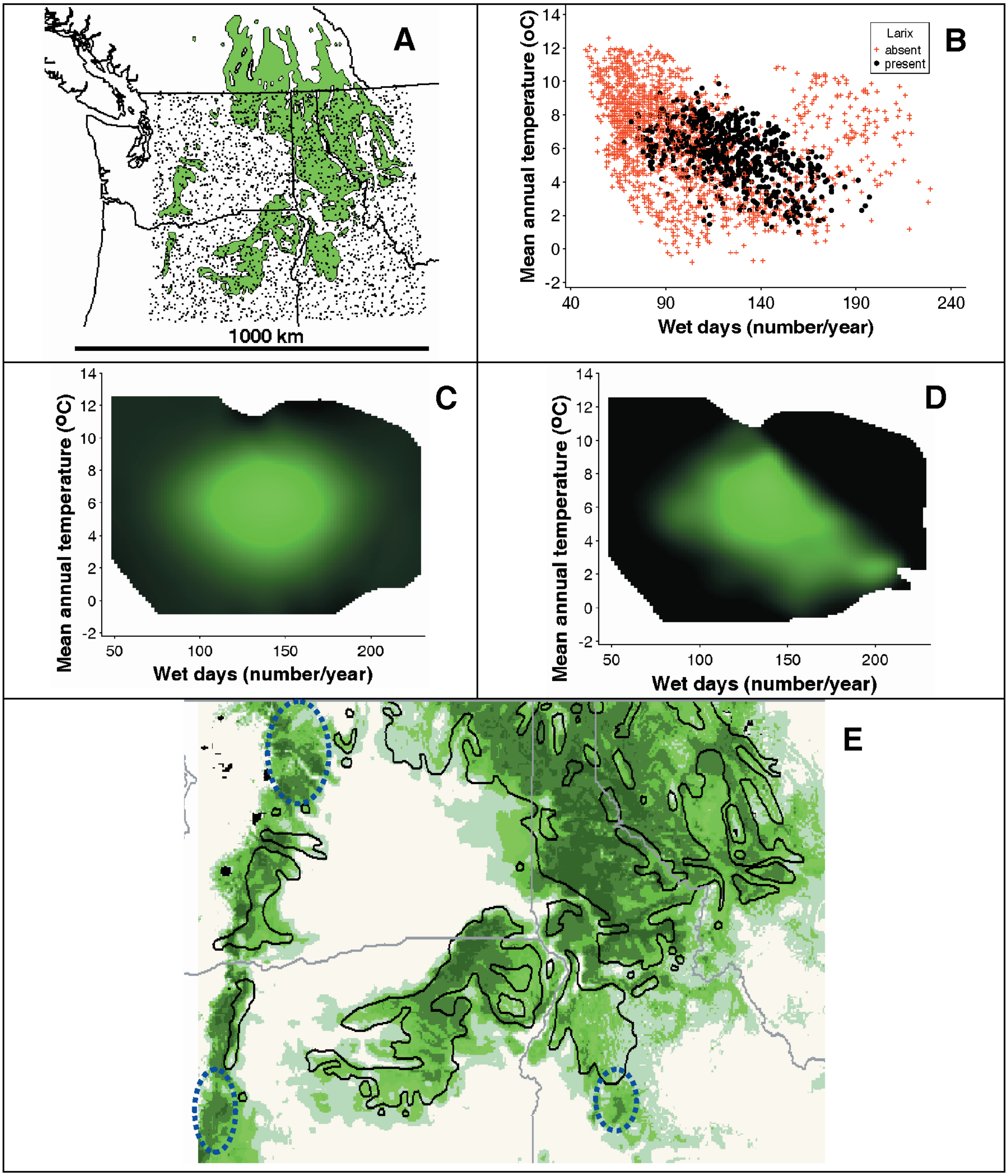

Fig. 2. **A.** Distribution of *Larix occidentalis* (green) in western North America and points used in the random sample. **B.** 2D slice through the predictor space; *Larix* was present at solid points (black), absent at + (red). **C.** Response surface from two-predictor GAM; gradient from black to green indicates likelihood of occurrence, with the greenest shade indicating the most favorable habitat. **D.** Response surface from two-predictor NPMR model. **E.** Estimated probability of occurrence of *L. occidentalis* superimposed on distribution map (black lines show range; gray lines are state boundaries). Deeper green indicates a higher probability of occurrence. Blue ellipses indicate areas where *L. occidentalis* is absent but potentially present, based on recent climate.

gentle on the warm-dry and cool sides.

A GAM (binomial family, logit link, spline smoother) with the two best predictors fit worse than NPMR with the same predictors (log*B* for GAM vs. naïve model = 241, log*B* for NPMR vs. GAM = 20). Adding an interaction term to the GAM gave only slight improvement.

Logistic regression (LR) of the same data set produced markedly poorer fits for a given number of predictors than did NPMR or GAM. Local mean NPMR with three predictors yielded log*B* = 314. In contrast, LR with forward stepwise selection from a pool of the nine untransformed climatic variables yielded a three-predictor model with log*B* = 169 (calculated from omnibus χ^2 statistic in SPSS as $\log_{10}B = \chi^2 / 4.60517$). Including two-way interactions among the four best predictors improved this to log*B* = 210. Adding quadratic terms for all predictors to the pool of predictors improved the best three-predictor LR model to log*B* = 211. Further improvements might be had with LR, but the point is not that a parametric model is impossible; rather that NPMR provides an effective, rapid, model-free assessment of the response surface, automatically allowing for interactions. NPMR can be used as the final model, or it can help to guide the design of an appropriate non-linear model, GLM, or GAM by studying 2D or 3D slices of the response surface.

Although the estimated likelihood of occurrence from NPMR and the actual distribution of *Larix* corresponded well (Fig. 2E), some differences emerged, suggesting disequilibrium between its current distribution and modern climates. One can readily identify where *Larix* is missing from areas that appear climatically favourable for it – for example the east slope of the Cascades in northern Washington. Most likely some factors other than modern climate, perhaps historical climates or disturbance regimes or both, have excluded the species.

Example with non-linear dynamics

Even a simple deterministic simulation model for two species and one environmental factor can produce a response surface that cannot be easily represented by standard habitat modelling tools. This is illustrated with a dynamic model of *Picea mariana* (Black spruce) and *Picea glauca* (White spruce) along a moisture gradient (Apps. 3 and 4). In the model, *P. mariana* and *P. glauca* increase in logistic fashion, following stand-replacing disturbance. *P. mariana* has a broader tolerance to moisture than does *P. glauca*, but *P. glauca* outcompetes *P. mariana* on mesic sites. This results in a bimodal realized niche for *P. mariana*, dominating on very dry and very wet sites (Curtis 1959; Loucks 1962). I used difference equations (App. 4) to generate a response surface on a grid of 10 dates × 11 steps on a moisture gradient, then used NPMR and GAM to fit statistical models to the noiseless response surface.

In this case GAM performs worse than NPMR (details in App. 3) because no single curve shape permeates either of the dimensions in the predictor space. In other words, GAM appears to fall short when parallel slices of the response surface along a given predictor have fundamentally different shapes, for example sigmoid on wet sites and hump-shaped on mesic sites. With NPMR, on the other hand, the curve shapes in one part of the multidimensional response surface need not bear any relationship to the shapes in other parts of the response surface.

Extensions to community analysis

So far NPMR has been applied only to problems with a single response variable. NPMR opens a door, however, to future improvements in multivariate analyses of ecological communities. Species abundance data have three properties in relationship to environmental gradients that challenge multivariate statistical analysis (Beals 1984; McCune & Grace 2002, pp. 35-43): the zero truncation problem, 'solid' response curves, and complex response shapes (including polymodality and asymmetry). Together, these properties produce the 'dust bunny' distribution of sample units in multidimensional species space (McCune & Grace 2002) rather than a multivariate normal distribution, demanding multivariate tools capable of effectively representing grossly nonlinear relationships. An ordination axis derived from community data can be considered a synthetic gradient through the dust bunny of sample units in species space. The collective relationships of species to these gradients, fit with NPMR, could be an improved basis for an optimization principle, replacing stress minimization in nonmetric multidimensional scaling.

The immediate utility of NPMR, however, will be improved empirical models of single species in relation to the factors that influence them. NPMR allows us to abandon simplistic assumptions about overall model form, embracing the ecological truism that habitat factors interact.

Acknowledgements. I thank my colleagues, especially Eric B. Peterson for collaborating on early work; Michael J. Mefford for collaboration with programming; Philip Dixon, Emily Holt, Erin Martin, Patricia Muir, Sarah Jovan, and Nigel Yoccoz for suggestions on the manuscript; and Alix Gitelman for her advice.

References

Anon. 1999. *Digital representation of "Atlas of United States Trees"*. by Elbert L. Little, Jr. Digital Version 1.0. USGS. URL http://climchange.cr.usgs.gov/info/veg-clim/

Agterberg, F.P. 1984. Trend surface analysis. In: Gaile, G.L. & Willmott, C.J. (eds.) *Spatial statistics and models*. Reidel, Dordrecht, NL.

Austin, M.P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. *Ecol. Model*. 157: 101-118.

Beals, E.W. 1984. Bray-Curtis ordination: an effective strategy for analysis of multivariate ecological data. *Adv. Ecol. Res*. 14: 1-55.

Begg, G.A. & Marteinsdottir, G. 1997. Environmental and stock effects on spatial distribution and abundance of mature cod *Gadus morhua*. *Marine Ecol. Progr. Ser*. 229: 245-262.

Bellman, R.E. 1961. *Adaptive control processes*. Princeton University Press, Princeton, NJ, US.

Berryman, S.D. 2002. *Epiphytic macrolichens in relation to forest management and topography in a western Oregon watershed*. Ph.D. Dissertation, Oregon State University, Corvallis, OR, US.

Bowman, A.W. & Azzalini, A. 1997. *Applied smoothing techniques for data analysis*. Clarendon Press, Oxford, UK.

Cade, B.S. & Noon, B.R. 2003. A gentle introduction to quantile regression for ecologists. *Frontiers Ecol. Environ*. 1: 412-420.

Ciannelli, L., Chan, K.-S., Bailey, K.M. & Stenseth, N.C. 2004. Nonadditive effects of the environment on the survival of a large marine fish population. *Ecology* 85: 3418-3427.

Curtis, J.T. 1959. *The vegetation of Wisconsin*. University of Wisconsin Press, Madison, WI, US.

Daly, C., Neilson, R.P. & Phillips, D.L. 1994. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteorol*. 33: 140-158.

Fleishman, E., MacNally, R. & Fay, J.P. 2003. Validation tests of predictive models of butterfly occurrence based on environmental variables. *Conserv. Biol*. 17:806-817.

Friedman, J.H. 1991. Multivariate adaptive regression splines (with discussion). *Ann. Stat*. 19: 1-141.

Gignac, L.D., Vitt, D.H., Zoltai, S.C. & Bayley, S.E. 1991a. Bryophyte response surfaces along climatic, chemical, and physical gradients in peatlands of western Canada. *Nova Hedw*. 53: 27-71.

Gignac, L.D., Vitt, D.H. & Bayley, S.E. 1991b. Bryophyte response surfaces along ecological and climatic gradients. *Vegetatio* 93: 29-45.

Gotway, C.A., Ferguson, R.B., Hergert, G.W. & Peterson, T.A. 1996. Comparison of kriging and inverse-distance methods for mapping soil parameters. *Soil Sci. Soc. Am. J*. 60: 1237-1247.

Guisan, A. & Zimmermann, N.E. 2000. Predictive habitat distribution models in ecology. *Ecol. Model*. 135: 147-186.

Harrell, F.E., Lee, K.L. & Mark, D.B. 1996. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med*. 15: 361-387.

Hastie, T.J. & Tibshirani, R.J. 1990. *Generalized Additive Models*. Chapman and Hall, London, UK.

Hastie, T., Tibshirani, R. & Friedman, J. 2001. *The elements of statistical learning*. Springer-Verlag, New York, NY, US.

Heglund, P.J. 2002. Foundations of species-environment relations. In: Scott, J.M., Heglund, P.J., Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.A. & Samson, F.B. (eds.) *Predicting species occurrences: Issues of accuracy and scale,* pp. 35-41. Island Press, Washington, DC, US.

Huisman, J., Olff, H. & Fresco, L.F.M. 1993. A hierarchical set of models for species response analysis. *J. Veg. Sci*. 4: 37-46.

Huntley, B., Bartlein, P.J. & Prentice, I.C. 1989. Climatic control of the distribution and abundance of beech (*Fagus* L.) in Europe and North America. *J. Biogeogr*. 16: 551-560.

Huntley, B., Berry, P.M., Cramer, W. & McDonald, A.P. 1995. Modelling present and potential future ranges of some European higher plants using climate response surfaces. *J. Biogeogr*. 22: 967-1001.

Huston, M.A. 2002. Critical issues for improving predictions. In: Scott, J.M., Heglund, P.J., Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.A. &Samson, F.B. (eds.) *Predicting species occurrences: Issues of accuracy and scale,* pp 7-21. Island Press, Washington, DC, US.

Hutchinson, G.E. 1957. Concluding remarks. *Cold Spring Harbor Symp. Quant. Biol*. 22: 415-427.

Kaiser, M.S., Speckman, P.L. & Jones, J.R. 1994. Statistical models for limiting nutrient relations in inland waters. *J. Am. Stat. Ass*. 89: 410-423.

Kass, R.E., & Raftery, A.E. 1995. Bayes factors. *J. Am. Stat. Assoc*. 90: 773-795.

Loucks, O.L. 1962. Ordinating forest communities by means of environmental scalars and phytosociological indices. *Ecol. Monogr*. 32: 137-166.

Manly, B.F.J. 1997. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. 2nd. ed. Chapman and Hall/CRC, Boca Raton, FL, US.

Maravelias, C.D. & Reid, D.G. 1997. Identifying the effects of oceanographic features and zooplankton on prespawning herring abundance using generalized additive models. *Marine Ecol. Progr. Ser*. 147: 1-9.

Martinez-Taberner, A., Ruiz-Perez, M., Mestre, I. & Forteza, V. 1992. Prediction of potential submerged vegetation in a silted coastal marsh, Albufera of Majorca, Balearic Islands. *J. Environ. Manage*. 35: 1-12.

McCune, B. 1993. Gradients in epiphyte biomass in three *Pseudotsuga-Tsuga* forests of different ages in western Oregon and Washington. *Bryol*. 96: 405-411.

McCune, B. & Grace, J.B. 2002. *Analysis of ecological communities*. MjM Software, Gleneden Beach, OR, US.

McCune, B. & Mefford, M.J. 2004. *HyperNiche. Non-parametric multiplicative habitat modeling*. Version 1.0. MjM Software, Gleneden Beach, OR, US.

McCune, B., Berryman, S.D., Cissel, J.H. & Gitelman, A.I..

2003. Use of a smoother to forecast occurrence of epiphytic lichens under alternative forest management plans. *Ecol. Appl.* 13: 1110-1123.

Mysterud, A., Stenseth, N.C., Yoccoz, N.G., Langvatn, R. & Steinhelm, G. 2001. Nonlinear effects of large-scale climatic variability on wild and domestic herbivores. *Nature* 410: 1096-1099.

Odum, E.P. 1971. *Fundamentals of ecology*. Saunders, Philadelphia, PA, US.

Rushton, S.P., Ormerod, S.J. & Kerby, G. 2004. New paradigms for modelling species distributions? *J. Appl. Ecol.* 41: 193-200.

Westfall, P.H. & Young, S.S. 1992. *Resampling-based multiple testing*. John Wiley and Sons, New York, NY, US.

Whittaker, R.H. 1956. Vegetation of the Great Smoky Mountains. *Ecol. Monogr.* 26: 1-80.

Received 27 September 2005;
Accepted 25 July 2006;
Co-ordinating Editor: P. Dixon.

For Apps. 1-4, see JVS/AVS Electronic Archives;
www.opuluspress.se/