ELSEVIER

Ecological Modelling 157 (2002) 89–100

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Generalized linear and generalized additive models in studies of species distributions: setting the scene

Antoine Guisan [a,b,*], Thomas C. Edwards, Jr [c], Trevor Hastie [d]

[a] *Swiss Center for Faunal Cartography (CSCF), Terreaux 14, CH-2000 Neuchâtel, Switzerland*
[b] *Institute of Ecology, University of Lausanne, BB, CH-1015 Lausanne, Switzerland*
[c] *USGS Biological Resources, Utah Cooperative Fish and Wildlife Research Unit, Utah State University, Logan, UT 84322-5210, USA*
[d] *Statistics Department, Stanford University, Sequoia Hall, Stanford, CA 94305, USA*

Abstract

An important statistical development of the last 30 years has been the advance in regression analysis provided by generalized linear models (GLMs) and generalized additive models (GAMs). Here we introduce a series of papers prepared within the framework of an international workshop entitled: *Advances in GLMs/GAMs modeling: from species distribution to environmental management*, held in Riederalp, Switzerland, 6–11 August 2001.We first discuss some general uses of statistical models in ecology, as well as provide a short review of several key examples of the use of GLMs and GAMs in ecological modeling efforts. We next present an overview of GLMs and GAMs, and discuss some of their related statistics used for predictor selection, model diagnostics, and evaluation. Included is a discussion of several new approaches applicable to GLMs and GAMs, such as ridge regression, an alternative to stepwise selection of predictors, and methods for the identification of interactions by a combined use of regression trees and several other approaches. We close with an overview of the papers and how we feel they advance our understanding of their application to ecological modeling.

Keywords: Statistical modeling; Generalized linear model; Generalized additive model; Species distribution; Predictive modeling

1. Introduction

An important statistical development of the last 30 years has been the advance in regression analysis provided by generalized linear models (GLM) and generalized additive models (GAM). Nowadays, both three-letter acronyms translate into a great potential for application in many fields of scientific research. Based on developments by Cox (1968) in the late sixties, the first seminal publications, also providing the link with practice (through software availability), were those of Nelder and Wedderburn (1972), McCullagh and Nelder (1983), Hastie and Tibshirani (1986, 1990). Since their development, both approaches have been extensively applied in ecological research, as evidenced by the growing number of published papers incorporating these modern regression

* Corresponding author. Tel.: +41-21-692-4254; fax: +41-21-692-4265
E-mail addresses: antoine.guisan@ie-bsg.unil.ch (A. Guisan), tce@nr.usu.edu (T.C. Edwards, Jr), hastie@stat.stanford.edu (T. Hastie).

0304-3800/02/$ - see front matter
PII: S0304-3800(02)00204-1

tools. This is due, in part, to their ability to deal with the multitude of distributions that define ecological data, and to the fact that they blend in well with traditional practices used in linear modeling and analysis of variance (ANOVA).

GLMs are mathematical extensions of linear models that do not force data into unnatural scales, and thereby allow for non-linearity and non-constant variance structures in the data (Hastie and Tibshirani, 1990). They are based on an assumed relationship (called a link function; see below) between the mean of the response variable and the linear combination of the explanatory variables. Data may be assumed to be from several families of probability distributions, including the normal, binomial, Poisson, negative binomial, or gamma distribution, many of which better fit the non-normal error structures of most ecological data. Thus, GLMs are more flexible and better suited for analyzing ecological relationships, which can be poorly represented by classical Gaussian distributions (see Austin, 1987).

GAMs (Hastie and Tibshirani, 1986, 1990) are semi-parametric extensions of GLMs; the only underlying assumption made is that the functions are additive and that the components are smooth. A GAM, like a GLM, uses a link function to establish a relationship between the mean of the response variable and a 'smoothed' function of the explanatory variable(s). The strength of GAMs is their ability to deal with highly non-linear and non-monotonic relationships between the response and the set of explanatory variables. GAMs are sometimes referred to as data- rather than model-driven. This is because the data determine the nature of the relationship between the response and the set of explanatory variables rather than assuming some form of parametric relationship (Yee and Mitchell, 1991). Like GLMs, the ability of this tool to handle non-linear data structures can aid in the development of ecological models that better represent the underlying data, and hence increase our understanding of ecological systems.

Few syntheses of GLMs and GAMs have been made since the first papers encouraged their use in ecological studies (Austin and Cunningham, 1981; Vincent and Haworth, 1983; Nicholls, 1989; Yee and Mitchell, 1991). As a first step in this direction, the series of papers included in this special issue all arose from a workshop (held in Riederalp, Switzerland, 6–10 August 2001) devoted to the use of GLMs and GAMs in ecology. Together, these papers constitute a valuable opportunity to report on the advances and insights derived from the application of these statistical tools to ecological questions over the last two decades. A series of more applied papers from the same workshop are found in a parallel special issue published in *Biodiversity and Conservation* (Guest Editors: Lehmann, A., Austin, M. and Overton, J.).

Our introductory review paper is necessarily restricted to GLMs and GAMs, and is intended to provide readers with some measure of the power of these statistical tools for modeling ecological systems. We first establish a context by discussing some general uses of statistical models in ecology, as well as providing a short review of several key studies that have advanced the use of GLMs and GAMs in ecological modeling efforts. We next present a general overview of GLMs and GAMs, and some of their related statistics that are used in predictor selection, diagnostics, and model evaluation. We close with an overview of the papers included in this volume and how we feel they advance our understanding of GLM and GAM applications to ecological modeling.

2. A framework for use of statistical models in ecological studies

We make a strong distinction here from general ecological models, speaking of statistical models as a subset distinct from conceptual or heuristic models. In most studies, some sort of conceptual or theoretical model (Austin, this volume) of the ecological system is already, and certainly should be, proposed (sensu Cale et al., 1983) before a statistical model is even considered (see also Guisan and Zimmermann, 2000). The purpose of the statistical model is to provide a mathematical basis for interpretation, examining such parameters as 'fit' (Do the measured predictors adequately explain the response?), 'strength' of

association (Is the relationship between the response and the predictors significant?), and to ascertain the contributions and roles of the different variables.

Reasons for the use of statistical models in ecology are as complex and varied as is the study of ecology (see Burnham and Anderson, 1998: Chapter 1). A complete overview is beyond the scope of this review, and readers are referred to more specialized literature (e.g. Ludwig and Reynolds, 1988; Jongman et al., 1995; Zar, 1996; Legendre and Legendre, 1998), all of which provide different and varying insights into the role of statistical modeling in ecology. From our perspective, one of the simplest and perhaps most widely understood characteristics of statistical models in ecology is the contrast between so-called explanatory and predictive models.

In general, explanatory models seek to provide insights into the ecological processes that produce patterns (e.g. Austin et al., 1990). Often, these relationships are determined from statistical models that ascertain the strength of the statistical relationship between a response (e.g. plant species presence) and a suite of one or more explanatory variables (e.g. precipitation, soil type, solar radiation). In contrast, predictive models typically seek to provide the user with a statistical relationship between the response and a series or predictor variables (hereafter simply called the predictors) for use in predicting the probability of species occurrence or estimating numbers of an organism at new, previously unsampled locations. These models often use variable reduction techniques in the analytical phase and have as their goal a model that predicts the ecological attribute(s) of interest from a restricted number of predictors. The concept of parsimony, that the simplest explanation is best, is inherent in such modeling efforts. The reduced model typically has lower variance, which will trade off with bias in optimizing prediction error.

Regression analyses have been broadly applied in ecology. However, one field where the use of modern regression approaches has proven particularly useful is the modeling of the spatial distribution of species and communities (Guisan and Zimmermann, 2000; Scott et al., 2002). Examples include the use of regression analyses to predict the distribution of tree and shrub species (Austin et al., 1983, 1990; Lenihan, 1993; Franklin, 1998; Guisan et al., 1999), of herbaceous species (Guisan et al., 1998; Guisan and Theurillat, 2000), of aquatic plant species (Lehmann, 1998), of terrestrial animal species (Pereira and Itami, 1991; Augustin et al., 1996; Manel et al., 1999; Guisan and Hofer, 2001; Jaberg and Guisan, 2001; Zimmermann and Breitenmoser, 2002), of birds (Manel et al., 1999, 2000), of aquatic animal species (invertebrates; Manel et al., 2000), of plant communities (Zimmermann and Kienast, 1999), or of structural vegetation types (Brown, 1994; Frescino et al., 2001). At a higher level of complexity, these approaches have also been used to investigate the distribution of plant (Currie and Paquin, 1987; Margules et al., 1987; Pausas, 1994; Heikkinen, 1996; Wohlgemuth, 1998) and animal diversity (Owen, 1989; Currie, 1991; Fraser, 1998).

Implicit in the application of regression tools for species modeling is a pseudo-equilibrium (Guisan and Theurillat, 2000, Austin this issue) between the organisms and their environments. Consequently, use of these tools to identify environmental factors responsible for the distribution of species that are, for example, still expanding their range in the study area can lead to biased results like truncated ecological response curves (Hirzel et al., 2001). GLMs and GAMs, the focus of this collection of papers, effectively model ecological (realized) rather than fundamental niches due to their intrinsic empirical nature. Thus, they implicitly incorporate biotic interactions and negative stochastic effects (Guisan and Zimmermann, 2000) that can change from one region to another. This can make models fitted for the same species, but in different areas and/or at different resolutions, difficult to compare (Guisan and Theurillat, 2000). Hence, the predictive capability of such models is frequently low (Roloff and Kernohan, 1999; Pearce and Ferrier, 2000), and most have limited success when applied to other sampling locations (Power, 1993). An exception is the tree species richness model of Currie (1991), developed in North America but which provided acceptable estimates when applied to UK.

Unfortunately, few studies using regression analyses for predictive purposes incorporate even simple statistical validation exercises (Fielding and Bell, 1997; Manel et al., 2002), even though numerous techniques exist (Manly, 1997). Even fewer perform field validation (Rykiel, 1996; Manel et al., 2002), calling into question the ultimate validity and application of the models (Guisan and Zimmermann, 2000). This lack of validation and uncertainty assessment remains a serious issue in ecological modeling (Fielding and Bell, 1997, Elith et al., this volume).

3. Regression models

3.1. Linear regression

Linear regression is one of the oldest statistical techniques, and has long been used in biological research. The basic linear regression model has the form:

$$Y = \alpha + X^T\beta + \varepsilon \quad (1)$$

where Y denotes the response variable, α is a constant called the intercept, $X = (X_1, \ldots, X_p)$ is a vector of p predictor variables, $\beta = \{\beta_1, \ldots, \beta_p\}$ is the vector of p regression coefficients (one for each predictor), and ε is the error. The error represents measurement error, as well as any variation unexplained by the linear model. When fitting a regression model, one tries to minimize this unexplained variation through the application of estimation techniques such as the least-squares (LS) algorithm.

Although a powerful approach in situations when appropriately applied, linear regression is limited by three main assumptions:

1) the errors ε_i are assumed to be identically and independently distributed; this includes the assumption that the variance of Y is constant across observations;
2) for testing purposes, the errors ε_i are assumed to follow a normal (Gaussian) distribution; and
3) the regression function is linear in the predictors.

Violation of assumption 1 constitutes a limitation to the application of most parametric statistical models, and is directly related to data sampling. Typically, many data in ecology are not Gaussian and do not have a constant variance. As an example, count data (e.g. number of individuals or species) follow a Poisson distribution (Vincent and Haworth, 1983; Jones et al., 2002; see also Barry and Welsh this volume), and their variance is proportional to their mean (Davison, 2001).

A common way of dealing with departures from assumptions 1 and 2 is to transform the response variable so that it meets the criteria of normality and constant variance. Several approaches for transforming data are available (e.g. Box–Cox approach), and the matter is still being discussed in the current literature (e.g. Marshall et al., 1995; Mateu, 1997). Violations of assumption 3 have traditionally been dealt with by augmenting the predictors with polynomial terms, interactions and other non-linear transformations of the original predictors, leading to a model non-linear in the X_j but linear in the parameters.

3.2. Generalized linear models

The assumptions above are implicit in LS regression. The advent of more flexible estimation techniques, such as maximum likelihood, was a major step forward in the development of GLMs (Nelder and Wedderburn, 1972, see McCullagh and Nelder, 1983 for the first comprehensive book). Because the mathematical rationale can be found in recent statistical textbooks (e.g. McCullagh and Nelder, 1989; Harrell, 2001; Hastie et al., 2001), we describe these models only briefly to provide context for the following papers.

In GLMs, the predictor variables X_j ($j = 1, \ldots, p$) are combined to produce a linear predictor LP which is related to the expected value $\mu = E(Y)$ of the response variable Y through a link function $g()$, such as:

$$g(E(Y)) = \mathrm{LP} = \alpha + X^T\beta \quad (2)$$

where α, X, β are those previously described in Eq. (1). We have written the model for generic variables X and Y; the corresponding terms for

the ith observation in the sample is:

$$g(\mu_i) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \quad (3)$$

Unlike classical linear models, which presuppose a Gaussian (i.e. normal) distribution and an identity link, the distribution of Y in a GLMs may be any of the exponential family distributions (e.g. Gaussian, Poisson or binomial) and the link function may be any monotonic differentiable function (like logarithm or logit). The variance of Y depends on $\mu = E(Y)$ through the variance function $V(\mu)$, giving $\mathrm{Var}(Y) = \phi V(\mu)$, where ϕ is a scale (also known as a dispersion) parameter. When the scale parameter is expected to be higher than the value anticipated under the chosen distribution (i.e. over-dispersion), the scale parameter can be estimated using quasi-likelihood; an extension of generalized least-squares (Davison, 2001).

The main improvements of GLMs over LS regression are hence:

1) the ability to handle a larger class of distributions for the response variable Y. Apart from the Gaussian, other distributions are the binomial, Poisson and Gamma; these are usually specified through their respective variance functions (see McCullagh and Nelder, 1989). GLMs can also accommodate more general qualitative (Davis and Goetz, 1990) and semi-quantitative (ordinal; Guisan and Harrell, 2000) response variables, usually based on a series of logistic binary GLMs;
2) the relationship of the response variable Y to the linear predictor (LP) through the link function $g(E(Y))$. In addition to ensuring linearity, this is an efficient way of constraining the predictions to be within a range of possible values for the response variable (e.g. between 0 and 1 for probabilities of presence). Guisan (2002) provides an illustration of this constraint, while Pregibon (1980), Breslow (1996) discuss some valuable tests for choosing the appropriate link function; and
3) it incorporates potential solutions (like quasi-likelihood) to deal with overdispersion (see Davison, 2001 for a more thorough discussion).

Fitting a GLM is much the same as fitting a multiple LS regression. Polynomial terms, or other parametric transformations, can be included in both cases in the set of predictors to account for non-linear and multi-modal responses (e.g. unimodal or bimodal). As in LS regression, the choice of the appropriate transformation can often be identified through scatterplots of partial residuals. Although several types of residuals are available for GLMs, partial residual plots based on the working residuals are most suitable for this purpose (Breslow, 1996). As in LS regression, influential observations (i.e. outliers) can be detected through standard diagnostics such as Cook's distance (see Breslow, 1996).

3.3. Generalized additive models

The identification of appropriate polynomial terms and transformations of the predictors to improve the fit of a linear model can be tedious and imprecise. The introduction of models that automatically identify appropriate transformations was a second important step forward in regression analyses. This led to a wider generalization of GLMs known as GAMs (Hastie and Tibshirani, 1990). One can envision the different regression models as being nested within each other, with simple and multiple LS linear regression (SLR and MLR) being the two most limiting cases, and GAMs the most general:

$$\mathrm{SLR} \subset \mathrm{MLR} \subset \mathrm{GLM} \subset \mathrm{GAM}$$

GAMs are parameterized just like GLMs, except that some predictors can be modeled non-parametrically in addition to linear and polynomial terms for other predictors. The probability distribution of the response variable must still be specified, and in this respect, a GAM is parametric. In this sense they are more aptly named semi-parametric models. A crucial step in applying GAMs is to select the appropriate level of the 'smoother' for a predictor. This is best achieved by specifying the level of smoothing using the concept of effective degrees of freedom. A reasonable balance must be maintained between the total number of observations and the total number of

degrees of freedom used when fitting the model (sum of levels of smoothness used for each predictor).

3.4. Variable selection methods and diagnostics

Variable selection is basically the same for all the described regression models, although evaluation criteria like the Akaike Information Criterion (AIC, Akaike, 1973; see also Sakamoto et al., 1988) can be used with GLMs and GAMs. In all models, one can use predefined rules such as deviance reduction as measured with the χ^2 statistic, or approaches that minimize AIC. More automatic procedures, such as stepwise regression or shrinkage rules, can also be used. However, stepwise procedures are considered to be high-variance operations because small perturbations of the response data can sometimes lead to vastly different subsets of the variables. They should be used with care (Guisan and Zimmermann, 2000). They can be improved if selection criteria based on permutations of the data, such as 5- or 10- fold cross-validation, are used (Hastie et al., 2001).

Shrinkage rules, such as ridge regression or lasso, are promising alternatives (Tibshirani, 1996; Harrell et al., 1996, 1998; Harrell, 2001; Hastie et al., 2001) to stepwise procedures when using GLMs or GAMs. Ridge regression keeps all the terms in the model, but shrinks their coefficients towards 0 using a quadratic penalty term, such as a bound on the sum-of-squares of the coefficients. This has the effect of reducing the variance of the fit of the model, while increasing the bias. By trading off these two quantities, a model that best predicts unseen observations can be identified. The lasso is similar, except that it imposes a bound on the sum of absolute values of the coefficients; this also shrinks the coefficients towards zero, but many of them are exactly set to 0 in the process. Hence, the lasso is a compromise between variable-subset selection and ridge regression. Both have a shrinkage parameter that needs to be selected, typically by cross-validation.

Collinearity in the predictors is another crucial problem associated with stepwise model selection (Brauner and Shacham, 1998). A common observation is that two highly correlated predictors can both appear non-significant even though each would explain a significant proportion of the deviance if considered individually. Various approaches can be used to detect harmful collinearity, such as condition number and variance inflation factor (VIF; Brauner and Shacham, 1998), although with careful model selection or regularization through application of ridge or lasso techniques collinearity becomes less of an issue.

The evaluation of interactions between two or more predictors is presently receiving more attention, particularly from an ecological perspective (see Austin, this volume). The failure to identify and incorporate ecologically meaningful interactions has constituted a major limitation of past ecological modeling exercises (Austin, this volume). A promising approach, suggested by T. Hastie during the workshop, is to use classification and regression tree (CART) techniques in a complementary way to GLMs and GAMs to identify these interactions. Another approach for considering species interactions might be to set up simultaneous GLM or GAM equations where each modelled species is incorporated as a predictor into the model of one or several other species (see Austin, 1971; Brzeziecki, 1987; Guisan, 2002).

Inference tests for the selection of predictors that explain a significant portion of the variance, or deviance in the case of maximum likelihood estimation techniques, are similar for all regression models, mainly the F-test in the case of LS regression and χ^2-tests in the case of GLMs and GAMs (Cantoni and Hastie, 2002). Several diagnostics can be applied to continuous response regression models. These can be used to assess the relevance of the chosen model (quantile–quantile plots, residual plots), identify outlying observations (Cook's distance), or to identify remaining trends in the data (partial residual plots). More specific diagnostic plots are needed in the case of GLMs or GAMs for nominal or ordinal responses (see e.g. Guisan and Harrell, 2000), or in the case of logistic binomial models (Davison, 1989a,b; Davison and Tsai, 1992; Hosmer and Lemeshow, 2000).

Where geographic space matters and local processes (e.g. climate) are known to occur, locally-weighted regressions can be fitted. Here, regression analyses are repeatedly applied within a moving window over the geographic range of interest (e.g. DAYMET climate modeling, Thornton et al., 1997). The study by Huntley et al. (1995) is one of the rare examples of this approach applied to species distribution modeling. This approach could be implemented with a two-dimensional smoother in a GAM, using the two geographic coordinates as a variable, along with other terms in the model. However, there are problems in making predictions from such a model, if the predictions are to be made at geographic locations vastly different from those encountered in the training data. Local regression models extrapolate poorly, if at all, depending on the type of kernel used.

4. What's in this issue

The papers presented in this volume provide a broad evaluation of GLMs and GAMs as applied to species distribution modeling. Many explore one or more issues, attempting to determine, in part, the utility of these tools for ecological modeling.

The first contribution by Mike Austin provides a major link between ecological theory and statistical modeling. Going further than simply reviewing the strengths and weaknesses of GLMs and GAMs, he proposes a useful framework for modeling species and community distributions, and testing ecological hypotheses. Austin makes a distinction among ecological models, data models and statistical models, and he warns that particular attention should be given to each type of model to ensure that relevant conclusions are drawn from model results. Particular attention is given to some of the major questions in species modeling, such as the shapes of response curves, causal versus indirect ecological predictors, modeling individual versus collective properties (communities, biodiversity), and incorporating ecological features in models, such as dispersion and competition.

Jari Oksanen and Peter Minchin expand the discussion on the shape of species response curves along continuous ecological gradients. Using data on vascular plant distribution along an elevation gradient, they test four main types of models for fitting such responses: (i) a hierarchical set of models (HOF) discussed by Huisman et al. (1993); (ii) binomial GLMs (logistic link); (iii) binomial GAMs (logistic link); and (iv) beta-functions (Austin et al., 1994). HOF models are the most effective method for their data, and GAMs provide very similar results in most cases.

The question of the shapes of species response curves also receives attention in Einar Heegaard's contribution. He shows how powerful GAMs are in this regard, principally through their fitting of non-parametric responses that more closely follow the data. His main emphasis is to provide the missing link between GAM-fitted responses and ecological theory. He does this by proposing two simple parameters, the outer and central borders, which can be easily calculated from the estimated values of the GAMs. Hence, these parameters are directly related to the response rather than to a parametric response function as in the case of GLMs.

Mark Boyce and co-authors introduce the use of GLMs for building resource selection functions (RSF) describing habitat use by animals—indeed a very similar approach to building predictive habitat distribution models—and the way they are commonly evaluated. Using two case studies of data distributed both in space and time, they show how the model evaluation process itself can be affected by ecological and behavioral variations that are specific to different sites and species history. They emphasize the importance of predictive capabilities of RSF models and propose a form of k-fold cross-validation for evaluating prediction success.

Thomas Yee and Monique Mackenzie introduce a broader class of linear and additive models, vector generalized linear models (VGLMs) and vector generalized additive models (VGAMs), and discuss their potential for ecological research. VGLMs and VGAMs comprise a large family of models and distributions, among which GLMs and GAMs are only a subset. As in GLMs and

GAMs, VGLMs are model-driven whereas VGAMs are more data-driven. A large variety of VGLMs and VGAMs are described and illustrated with examples of particular relevance for ecologists.

Simon Woods and Nicole Augustin give a tutorial introduction to smoothing splines and their computational details. They then propose an approximation that simplifies the computations considerably, and extend these to GAM models. The reduced computations allow fast automatic selection of the smoothing parameters using the GCV criterion, an approximation to cross-validation. They demonstrate their R implementation of this software on two environmental examples.

Simon Barry and Alan Welsh present alternative flexible GAMs models for predicting species distributions when observed count data include a larger proportion of zeros than expected (i.e. zero-inflated) in a Poisson distribution. In their approach, distribution patterns are modeled in two steps: (i) picking the main presence–absence pattern using a logistic model; and (ii) fitting the remaining variation in abundance where the species is present (i.e. conditional on the response being greater than 0) by using a second-abundance (Poisson)-model. An example is then provided using data on the distribution on stem counts of *Eucalyptus mannifera* in the South-East of Australia.

Patrick Osborne and Susana Suárez-Seoanes discuss the problem of spatial non-stationarity in the data and its influence on the quality of the model fit. Using data from three bird species in Spain, they build sub-models by partitioning the data set spatially into geographical quarters or rings based on the centroid of the modeling space. These sub-models are compared with each other and against the global model. They conclude that spatial partitioning is useful for detecting spatial non-stationarity, and hence alert the modeler to some particular eco-geographic patterns, but that random sampling should be preferred to build robust models. For future research, they propose interesting alternative modeling approaches for use when spatial non-stationarity is detected.

Jennifer Miller and Janet Franklin applied a different approach to evaluate spatial dependences, using indicator kriging that included neighborhood influences as a predictor in the model. Their modeling of the distribution of four vegetation alliances in the Mojave Desert in California shows that including such spatial autocorrelation in the model improves model fit and accuracy, although the resulting spatial predictions may look unrealistically smooth in some cases.

Gretchen Moisen and Tracey Frescino compare GAMs to four alternative modeling techniques of various levels of complexity: simple linear models (LM), CART, multivariate adaptive regression splines (MARS), and artificial neural networks (ANN). Models were applied to both nominal and continuous forest habitat response variables. MARS and ANN worked best when applied to simulated data, but less so when applied to real data, in which case a LM approach often provided comparable results. GAMs and MARS, however, were marginally best overall for modeling forest characteristics.

The contribution by Jane Elith and co-authors is based on the observation that too few studies of species distribution modeling incorporate maps of prediction uncertainties. Such information is often crucial for decision makers involved in conservation and management, and should be seen as an important complement to probability maps. In their review, they distinguish epistemic from linguistic uncertainties. The first category encompasses measurement of systematic errors as well as natural variation, model uncertainty and subjective uncertainty, whereas the second category encompasses vagueness, ambiguity and underspecificity. These concepts are discussed in the context of GLMs, and examples using logistic regression models are presented.

Richard Aspinall uses GLMs to evaluate and calibrate classification outputs from high spatial resolution hyperspectral imagery, focusing on the identification and mapping of coarse woody debris and *Populus* spp. Hyperspectral imagery is a promising remote sensing approach with application to the predictive mapping of finer habitat units and, potentially, of single species distributions. He used logistic regression to relate the imagery to the categorical field data, then built a predictive model identifying both the likelihood of

woody debris and *Populus* spp. being present, and the amount, in each specific spatial location. The measured goodness-of-fit of the model provides a simple but comprehensive measure of classification accuracy.

Anthony Lehmann and co-authors present a powerful automated tool they call GRASP, which formalizes an approach to species distribution modeling using GAMs. The approach is illustrated using two types of response variable: (i) presence–absence of the fern species *Cyathea dealbata* (binary data); and (ii) overall fern species richness (count data). Measures of model goodness-of-fit and of quality of predictions, as well as model interpretability, are discussed.

Elisabeth Zaniewski and co-authors test an approach for fitting GAMs when only presence data are available. Models are prepared for 43 native species of ferns in New Zealand, using presence-only data from a data set including presence–absence data gathered at nearly twenty thousand sites. They compared their approach with another, the ecological niche factor analysis (ENFA; Hirzel et al., 2001), which is also used for modeling presence—only data. Because logistic models require a binary response variable, 'pseudo' absences were generated according to different methods. The GAM-based models provided slightly better species' predictions than the ENFA-based approach, although ENFA provides more realistic predictions of collective properties such as species richness

Finally, Alexandre Hirzel and Antoine Guisan use a simulation approach to compare different sampling strategies and different sample sizes for building GLMs-based predictive habitat distribution models. The strategies compared included a: (i) systematic grid; (ii) pure random; (iii) random-stratified with an equal number of replicates per stratum; and (iv) random-stratified with the number of replicates per stratum proportional to the stratum area. Comparative evaluations show that, overall, equal random-stratified and grid perform best and are seemingly equivalent.

It is our belief that the papers in this volume provide a unique overview of the use of GLMs and GAMs in modeling species distributions. Each evaluates one or more aspects of these statistical tools and their use in ecological studies. Together, they serve as an excellent source of knowledge that should foster the continued and increased use of GLMs and GAMs in ecology. Enjoy!

5. List of workshop participants

Twenty-eight scientists from 11 countries attended the workshop. We wish to extend our warmest thanks to them for their involvement. In alphabetical order, the following persons were present: Richard Aspinall (USA), Nicole Augustin (D), Mike Austin (AUS), Simon Barry (AUS), Ana Bio (NL), Mark Boyce (CA), Margaret Cawsey (AUS), Thomas C. Edwards, Jr (USA), Jane Elith (AUS), Simon Ferrier (AUS), Antoine Guisan (CH), Trevor Hastie (USA), Einar Heegaard (NO), Alexandre Hirzel (CH), Christianne Ilg (CH), John Leathwick (NZ), Anthony Lehmann (CH), Monique Mackenzie (NZ), Ramona Maggini (CH), Jennifer Miller (USA), Jari Oksanen (FIN), Patrick Osborne (UK), Jacob Overton (NZ), Jennie Pearce (CA), Nicolas Ray (CH), José Texeira (P), Elisabeth Zaniewski (CH). Here, AUS = Australia, CA = Canada, CH = Switzerland, D = Germany, FIN = Finland, NZ = New-Zealand, P = Portugal, NL = The Netherlands, NO = Norway, UK = United Kingdom, USA = United States of America.

Acknowledgements

The workshop was jointly organized by the University of Geneva, the Swiss Center for Faunal Cartography (CSCF, Neuchâtel, Switzerland), the CSIRO in Canberra (Australia) and the Landcare Research Institute in Hamilton (NZ). The organizing committee was composed of five scientists from four countries: Anthony Lehmann (University of Geneva, Switzerland; present address: CSCF, Switzerland), Antoine Guisan (CSCF, Switzerland; present address: University of Lausanne), Mike Austin (CSIRO, AU), Jake Overton (Landcare, NZ), Thomas C. Edwards Jr (US Geological Survey, USA). The workshop benefited from generous donations from four sponsors: the

Swiss Academy of Natural Sciences (ASSN/SANW), the Swiss National Science Foundation (SNF), Mathsoft Ltd, through its software distribution branch Insightful (S-Plus), and the Swiss branch of the Environmental Science Research Institute Inc. (ESRI). We sincerely thank them all. Last, we thank Mike Austin, Anthony Davison, T.B. Murphy, and Thomas Yee for their valuable comments on this and earlier drafts of our manuscript.

References

Akaike, H., 1973. Information theory as an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Second International Symposium on Information Theory. Akademiai Kiado, Budapest, Hungary, pp. 267–281.

Augustin, N.H., Mugglestone, M.A., Buckland, S.T., 1996. An autologistic model for the spatial distribution of wildlife. J. Appl. Ecol. 33, 339–347.

Austin, M.P., 1971. Role of regression analysis in plant ecology. Proc. Ecol. Soc. Aust. 6, 63–75.

Austin, M.P., 1987. Models for the analysis of species response to environmental gradients. Vegetatio 69, 35–45.

Austin, M.P., Cunningham, R.B., 1981. Observational analysis of environmental gradients. Proc. Ecol. Soc. Aust. 11, 109–119.

Austin, M.P., Cunningham, R.B., Good, R.B., 1983. Altitudinal distribution in relation to other environmental factors of several Eucalypt species in southern New South Wales. Aust. J. Ecol. 8, 169–180.

Austin, M.P., Nicholls, A.O., Margules, C.R., 1990. Measurement of the realized qualitative niche: environmental niche of five Eucalyptus species. Ecol. Monogr. 60, 161–177.

Austin, M.P., Nicholls, A.O., Doherty, M.D., Meyers, J.A., 1994. Determining species response functions to an environmental gradient by means of a beta-function. J. Veg. Sci. 5, 215–228.

Brauner, N., Shacham, M., 1998. Role of range and precision of the independent variable in regression of data. Am. Inst. Chem. Eng. J. 44, 603–611.

Breslow, N.E., 1996. Generalized linear models: checking assumptions and strengthening conclusions. Stat. App. 8, 23–41.

Brown, D.G., 1994. Predicting vegetation types at treeline using topography and biophysical disturbance variables. J. Veg. Sci. 5, 641–656.

Brzeziecki, B., 1987. Analysis of vegetation–environment relationships using a simultaneous equations model. Vegetatio 71, 175–184.

Burnham, K.P., Anderson, D.R., 1998. Model Selection and Inference: a Practical Information Theoretic Approach. Springer, New York.

Cale, W.G., O'Neill, R.V., Shugart, H.H., 1983. Development and application of desirable ecological models. Ecol. Model. 18, 171–186.

Cantoni, E., Hastie, T., (in press). Degrees-of-Freedom Tests for Smoothing Splines. Biometrika.

Cox, D.R., 1968. Notes on some aspects of regression analysis (with Discussion). J. R. Stat. Soc. B49, 1–39.

Currie, D.J., 1991. Energy and large-scale patterns of animal- and plant-species richness. Am. Nat. 137, 27–49.

Currie, D.J., Paquin, V., 1987. Large-scale biogeographical patterns of species richness of trees. Nature 329, 326–327.

Davis, F.W., Goetz, S., 1990. Modeling vegetation pattern using digital terrain data. Landscape Ecol. 4, 69–80.

Davison, A.C., 1989a. Model-checking I: general regression models. Revista Brasileira de Probabilidade e Estatística 3, 77–86.

Davison, A.C., 1989b. Model-checking II: binary data. Revista Brasileira Probabilidade Estatística 3, 87–96.

Davison, A.C., 2001. Biometrika centenary: theory and general methodology. Biometrika 88, 13–52.

Davison, A.C., Tsai, C.-L., 1992. Regression model diagnostics. Int. Stat. Rev. 60, 337–353.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence–absence models. Environ. Conserv. 24, 38–49.

Franklin, J., 1998. Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. J. Veg. Sci. 9, 733–748.

Fraser, R.H., 1998. Vertebrate species richness at the mesoscale: relative roles of energy and heterogeneity. Glob. Ecol. Biogeogr. Lett. 7, 215–220.

Frescino, T.S., Edwards, T.C., Jr, Moisen, G.G., 2001. Modeling spatially explicit forest structural attributes using generalized additive models. J. Veg. Sci. 12, 15–26.

Guisan, A., 2002. Semi-quantitative models for predicting the spatial distribution of plant species. In: Scott, J.M., Heglund, P.J., Samson, F., Haufler, J., Morrison, M., Raphael, M., Wall, B. (Eds.), Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Covelo, California.

Guisan, A., Harrell, F.E., 2000. Ordinal response regression models in ecology. J. Veg. Sci. 11, 617–626.

Guisan, A., Theurillat, J.-P., 2000. Equilibrium modeling of alpine plant distribution: how far can we go. Phytocoenologia 30, 353–384.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186.

Guisan, A., Hofer, U., 2001. Modélisation du domaine de distribution potentielle des espèces. In: Hofer, U., Monney, J.-C., Dusej, G. (Eds.), Les Reptiles de Suisse (in French,

German and Italian). Birkhäuser Verlag, Basel, pp. 183–189.

Guisan, A., Theurillat, J.-P., Kienast, F., 1998. Predicting the potential distribution of plant species in an alpine environment. J. Veg. Sci. 9, 65–74.

Guisan, A., Weiss, S.B., Weiss, A.D., 1999. GLMs versus CCA spatial modeling of plant species distribution. Plant Ecol. 143, 107–122.

Harrell, F.E., Jr, 2001. Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer, New York.

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.

Harrell, F.E., Margolis, P.A., Gove, S., Mason, K.E., Mulholland, E.K., Lehmann, D., Muhe, L., Gatchalian, S., Eichenwald, H.F., 1998. Development of a clinical prediction model for an ordinal outcome. Stat. Med. 17, 909–944.

Hastie, T.J., Tibshirani, R.J., 1986. Generalized additive models. Stat. Sci. 1, 297–318.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall.

Hastie, T.J., Tibshirani, R.J., Friedman, J., 2001. Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, New York.

Heikkinen, R.K., 1996. Predicting patterns of vascular plant species richness with composite variables: a meso-scale study in Finnish Lapland. Vegetatio 126, 151–165.

Hirzel, A., Helfer, V., Métral, F., 2001. Assessing habitat-suitability models with a virtual species. Ecol. Model. 145, 111–121.

Hosmer, D.W., Jr, Lemeshow, S., 2000. Applied Logistic Regression, second ed.. Wiley, New York.

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

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

Jaberg, C., Guisan, A., 2001. Modeling the influence of landscape structure on bat species distribution and community composition in the Swiss Jura Mountains. J. Appl. Ecol. 38, 1169–1181.

Jones, M.T., Niemi, G.J., Hanowski, J.M., Regal, R.R., 2002. Poisson regression: a better approach to modeling abundance data. In: Scott, J.M., Heglund, P.J., Samson, F., Haufler, J., Morrison, M., Raphael, M., Wall, B. (Eds.), Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Covelo, CA.

Jongman, R.H.G., ter Braak, C.J.F., van Tongeren, O.F.R., 1995. Data analysis in community and landscape ecology. Cambridge University Press, Cambridge, UK.

Legendre, P., Legendre, L., 1998. Numerical ecology, second ed. (Englilsh). Elsevier, Amsterdam.

Lehmann, A., 1998. GIS modeling of submerged macrophyte distribution using generalized additive models. Plant Ecol. 139, 113–124.

Lenihan, J.M., 1993. Ecological responses surfaces for North American tree species and their use in forest classification. J. Veg. Sci. 4, 667–680.

Ludwig, J.A., Reynolds, J.F., 1988. Statistical Ecology. Wiley, New York.

Manel, S., Dias, J.-M., Ormerod, S.J., 1999. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. Ecol. Model. 120, 337–347.

Manel, S., Buckton, S.T., Ormerod, S.J., 2000. Testing large-scale hypotheses using surveys: the effects of land use on the habitats, invertebrates and birds of Himalayan rivers. J. Appl. Ecol. 37, 756–770.

Manel, S., Williams, H.C., Ormerod, S.J., (in press). Evaluating presence–absence models in ecology: the need to account for prevalence. J. Appl. Ecol.

Manly, B.F.J., 1997. Randomization, Bootstrap and Monte Carlo Methodology Methods in Biology, second ed.. Chapman & Hall, New York.

Margules, C.R., Nicholls, A.O., Austin, M.P., 1987. Diversity of Eucalyptus species predicted by a multi variables environmental gradient. Oecologia (Berlin) 71, 229–232.

Marshall, P., Szikszai, T., LeMay, V., Kozak, A., 1995. Testing the distributional assumptions of least squares linear regression. Forest. Chron. 71, 213–218.

Mateu, J., 1997. Methods of assessing and achieving normality applied to environmental data. Environ. Manage. 21, 767–777.

McCullagh, P., Nelder, J.A., 1983. Generalized Linear Models, First ed.. Chapman and Hall, London.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed.. Chapman & Hall, London.

Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized linear models. J. R. Stat. Soc. A135, 370–384.

Nicholls, A.O., 1989. How to make biological survey go further with generalized linear models. Biol. Conserv. 50, 51–75.

Owen, J.G., 1989. Patterns of herpetofaunal species richness: relation to temperature, precipitation and variance in elevation. J. Biogeogr. 16, 141–150.

Pausas, J.G., 1994. Species richness patterns in the understorey of Pyrenean Pinus sylvestris forest. J. Veg. Sci. 5, 517–524.

Pearce, J.L., Ferrier, S., 2000. Evaluating the predictive performance of models developed using logistic regression. Ecol. Model. 133, 225–245.

Pereira, J.M.C., Itami, R.M., 1991. GIS-based habitat modeling using logistic multiple regression: a study of the Mt. Graham Red Squirrel. Photogramm. Eng. Rem. Sens. 57, 1475–1486.

Power, M., 1993. The predictive validation of ecological and environmental models. Ecol. Model. 68, 33–50.

Pregibon, D., 1980. Goodness-of-link tests for generalized linear models. J. Appl. Stat. 29, 15–24.

Roloff, G.J., Kernohan, B.J., 1999. Evaluating the reliability of habitat suitability index models. Wildl. Soc. Bull. 27, 973–985.

Rykiel, E.J., Jr, 1996. Testing ecological models: the meaning of validation. Ecol. Model. 90, 229–244.

Sakamoto, Y., Ishiguro, M., Kitagawa, G., 1988. Akaike Information Criterion Statistics. KTK Scientific Publisher, Tokyo.

Scott, J.M., Heglund, P.J., Samson, F., Haufler, J., Morrison, M., Raphael, M., Wall, B., 2002. Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Covelo, California.

Thornton, P.E., Running, S.W., White, M.A., 1997. Generating surfaces of daily meteorological variables over large regions of complex terrain. J. Hydrol. 190, 214–251.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B58, 267–288.

Vincent, P.J., Haworth, J.M., 1983. Poisson regression models of species abundance. J. Biogeogr. 10, 153–160.

Wohlgemuth, T., 1998. Modeling floristic species richness on a regional scale: a case study in Switzerland. Biodivers. Conserv. 7, 159–177.

Yee, T.W., Mitchell, N.D., 1991. Generalized additive models in plant ecology. J. Veg. Sci. 2, 587–602.

Zar, J.H., 1996. Biostatistical Analysis, third ed.. Prentice Hall, Upper Saddle River, USA.

Zimmermann, N.E., Kienast, F., 1999. Predictive mapping of alpine grasslands in Switzerland: species versus community approach. J. Veg. Sci. 10, 469–482.

Zimmermann, F., Breitenmoser, U., 2002. A distribution model for the Eurasian Lynx (*Lynx lynx*) in the Jura mountains, Switzerland. In: Scott, M., et al. (Ed.), Predicting Species Occurrence: Issues in Scale and Accuracy. Island Press, Covelo, CA.