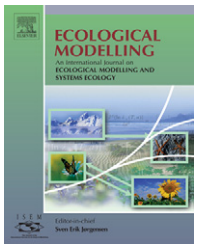
available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel



Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions

J.R. Leathwick[a,*], *J. Elith*[b], *T. Hastie*[c]

[a] *National Institute of Water and Atmospheric Research, P.O. Box 11115, Hamilton, New Zealand*
[b] *School of Botany, The University of Melbourne, Parkville, Victoria, Australia*
[c] *Department of Statistics, Stanford University, CA, USA*

ARTICLE INFO

Article history:
Published on line 18 July 2006

Keywords:
Distribution
Environment
Fish
Freshwater
Generalized additive model
Multivariate adaptive regression splines

ABSTRACT

Two statistical modelling techniques, generalized additive models (GAM) and multivariate adaptive regression splines (MARS), were used to analyse relationships between the distributions of 15 freshwater fish species and their environment. GAM and MARS models were fitted individually for each species, and a MARS multiresponse model was fitted in which the distributions of all species were analysed simultaneously. Model performance was evaluated using changes in deviance in the fitted models and the area under the receiver operating characteristic curve (ROC), calculated using a bootstrap assessment procedure that simulates predictive performance for independent data. Results indicate little difference between the performance of GAM and MARS models, even when MARS models included interaction terms between predictor variables. Results from MARS models are much more easily incorporated into other analyses than those from GAM models. The strong performance of a MARS multiresponse model, particularly for species of low prevalence, suggests that it may have distinct advantages for the analysis of large datasets. Its identification of a parsimonious set of environmental correlates of community composition, coupled with its ability to robustly model species distributions in relation to those variables, can be seen as converging strongly with the purposes of traditional ordination techniques.

1. Introduction

Over the last decade ongoing development of statistical modelling tools (e.g., Hastie et al., 2001) has led to a growing sophistication in the methods used to analyse relationships between the distributions of species and their environment (e.g., Guisan and Zimmermann, 2000). Such analyses are now widely used in terrestrial (e.g., Pereira and Itami, 1991; Ferrier et al., 2002), freshwater (e.g., Lek et al., 1996; Olden and Jackson, 2001) and marine settings (e.g., Ysebaert et al., 2002), motivated by purposes ranging from the testing of ecological hypotheses (e.g., Austin, 2002) or processes (Leathwick and Austin, 2001) to the prediction of species distributions across geographically extensive areas for conservation (e.g., Gregr and Trites, 2001; Elith and Burgman, 2002) and/or resource management (e.g., Borchers et al., 1997). While earlier techniques such as generalized linear models (GLM—McCullagh and Nelder, 1989) were found to be limited in their ability to fit the complex, non-linear relationships often occurring between species and environmental predictors (e.g., Austin et al., 1990), a range of techniques are now available that allow their more realistic description. Of these, generalized additive

* *Corresponding author.*
E-mail address: j.leathwick@niwa.co.nz (J.R. Leathwick).

0304-3800/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2006.05.022

models (GAM—Hastie and Tibshirani, 1990) are perhaps the most widely used, particularly in terrestrial (e.g., Leathwick, 1998) and marine studies (e.g., Gregr and Trites, 2001). However, although their use of non-parametric smoothing functions allows flexible description of complex species responses to environment (Yee and Mitchell, 1991), their computational complexity makes cumbersome the generation of predictions for independent datasets such as in a geographic information system (GIS).

Two other commonly used techniques capable of fitting non-linear relationships between species and environment are neural nets (Ripley, 1996) and classification and regression trees (Breiman et al., 1984). Of these, neural nets are prone to become computationally intractable with larger datasets (e.g., Moisen and Frescino, 2002; Friedman and Meulman, 2003), while misclassification can be problematic with classification and regression trees unless they are used in conjunction with boosting algorithms (Friedman and Meulman, 2003). A third alternative, multivariate adaptive regression splines (MARS—Friedman, 1991), has shown promise in recent comparative studies (Moisen and Frescino, 2002; Muñoz and Fellicísimo, 2004). This technique combines the strengths of regression trees and spline fitting by replacing the step functions normally associated with regression trees with piecewise linear basis functions (see Hastie and Tibshirani, 1990, Chapter 9). This allows the modelling of complex relationships between a response variable and its predictors. In practical terms, MARS has exceptional analytical speed, and its simple rule-based basis functions facilitate the prediction of species distributions using independent data (Muñoz and Fellicísimo, 2004), stored, for example, in a GIS.

In this study we compare the performance of GAM and MARS analyses of an extensive set of data describing the distributions of 15 fish species in New Zealand rivers and streams. While our past experience has mostly involved use of GAMs (e.g., Elith and Burgman, 2002; Leathwick and Austin, 2001), our increasing use of extensive databases has sometimes resulted in very slow model fitting, and at times we have been unable to fit GAM models to full datasets because of memory limitations imposed by the numerical complexities of this technique. This prompted our exploration of alternative methods capable of realistically analysing ecological data. A comprehensive description of the ecological insights derived from the MARS component of this analysis is contained in Leathwick et al. (2005).

2. Materials and methods

2.1. Distributional data

Both the species distribution data and the associated environmental data used in this study are described comprehensively by Leathwick et al. (2005) and only a brief summary is provided here. Fish distribution data comprised capture records from 9866 sites, extracted from the New Zealand Freshwater Fish Database (McDowall and Richardson, 1983; http://www.niwa.co.nz/services/nzffd/). As records of fish abundances were available for only a subset of sites, all data were converted to a common basis (presence–absence) for this analysis. Data describe the distributions of 15 diadromous species (Table 1), species that move between freshwater and marine habitats in completing their life cycles (McDowall, 1999). All species occurred in the dataset with a capture frequency of 0.5% or above, i.e., had a minimum of nearly 50 positive occurrences. Sixteen environmental predictors (Table 2) were selected for their functional relevance to the physiological and behavioural attributes of diadromous species. These include factors describing the character of the river segment within which the sampling site was located, downstream factors affecting the ability of diadromous fish to migrate from the sea to that river segment, and upstream/catchment-scale factors affecting environmental conditions at the sampling site. As regression methods are potentially sensitive to correlated variables, the final set of candidate variables was restricted to those with pairwise correlations of less than 0.7, with one variable normalised to reduce its correlation with other variables.

Table 1 – Fish species included in the analysis, and their prevalence, i.e., the proportion of sample sites at which they were recorded

Species code	Species name	Prevalence
Angaus	*Anguilla australis*	0.233
Angdie	*A. dieffenbachii*	0.577
Galarg	*Galaxias argenteus*	0.034
Galbre	*G. brevipinnis*	0.099
Galfas	*G. fasciatus*	0.137
Galmac	*G. maculatus*	0.118
Galpos	*G. postvectis*	0.025
Gobcot	*Gobiomorphus cotidianus*	0.183
Gobgob	*G. gobioides*	0.012
Gobhub	*G. hubbsi*	0.065
Gobhut	*G. huttoni*	0.211
Geoaus	*Geotria australis*	0.031
Chefos	*Cheimarrichthys fosteri*	0.121
Rhoret	*Rhombosolea retiaria*	0.007
Retret	*Retropinna retropinna*	0.042

2.2. Model fitting

2.2.1. Generalized additive models

Initially, we attempted to fit generalized additive models in S-Plus (v. 6.1, Insightful Corporation, Seattle) using a starting model that included all predictor variables as smooth terms, and which was then simplified as required using a backwards/forwards stepwise procedure to remove terms making a non-significant contribution. However, this procedure was not only very slow, but we were also unable to compare the statistical significance of fitting predictors as linear versus smooth terms because of the excessive memory demands with a dataset of this size. Similar problems were also encountered when this analysis was attempted using the 'gam' package in R (R Development Core Team, 2004).

As an alternative strategy, we used BRUTO (available in the 'mda' library for both S-Plus and R and documented by Hastie and Tibshirani, 1996), which fits a generalized additive model using an adaptive back-fitting procedure (Hastie and Tibshirani, 1990). In addition to identifying which variables to include in the final GAM model, BRUTO identifies the

Table 2 – Environmental predictors used to analyse fish capture

Variable	Mean and range
Segment scale predictors	
SegJanT—summer air temperature (°C)	16.6, 9.5–19.8
SegTSeas—winter air temperature (°C), normalised with respect to SegJanT, i.e., $\text{SegTSeas} = \left(\left(\frac{W-\overline{W}}{\sigma_W}\right) - \left(\frac{S-\overline{S}}{\sigma_S}\right)\right) \times \sigma_W$ where W is the winter temperature for a segment, $\overline{W}$ the average winter temperature for all segments, σ_W the standard deviation of winter temperature, S is the summer temperature and so on	0.75, −2.6 to 4.1
SegFlow—segment flow (m^3/s), fourth root transformed	0.82, 0.1–5.0
SegShade—riparian shade (proportion)	0.44, 0–0.8
SegSlope—segment slope (°), square-root transformed	2.2, 0–5.6
Downstream predictors	
DSDist—distance to coast (km)	51.5, 0.03–329.5
DSAveSlope—downstream average slope (°)	0.27, 0–14.5
DSMaxSlope—maximum downstream slope (°)	17.6, 0–56.5
Upstream/catchment scale predictors	
USAvgTNorm—average air temperature (°C)	Discarded
USRainDays—days/month with rain greater than 25 mm	1.29, 0.21–3.30
USSlope—average slope in the catchment (°)	13.9, 0–41.0
USIndigForest—area with indigenous forest (proportion)	0.334, 0–1
USPhos—average phosphorous concentration of underlying rocks, 1: very low to 5: very high	2.35, 1–5
USCalc—average calcium concentration of underlying rocks, 1: very low to 4: very high	1.46, 1–4
USHard—average hardness of underlying rocks, 1: very low to 5: very high	3.05, 1–5
USPeat—area of peat (proportion)	0.007, 0–1
USLake—area of lake (proportion)	0.002, 0–1

optimal degree of smoothing for each variable. BRUTO also allows specification of a penalty parameter that is applied to the addition of extra variables in the model, and we used 10-fold cross-validation to verify that the default value of 2 for the penalty parameter was appropriate for our data (Hastie et al., 2001). However, because BRUTO can only be used to fit models assuming Gaussian errors, model parameters describing the selected variables and their degree of smoothing were extracted and used to specify a model of identical form but allowing for binomial errors. This was then fitted using the standard GAM function ('gam'). Comparison of full backwards stepwise GAM models and BRUTO/GAM models for all species using k-fold cross-validation (e.g., Hastie et al., 2001) indicated that, while the full GAM models were better fitted to the training data, the BRUTO/GAM models delivered superior performance for independent sites. In addition, because the BRUTO/GAM models could be fitted in only 1–2% of the time taken for the full backwards GAM models, we were able to assess their performance more rigorously using the computationally intensive re-sampling techniques described below.

2.2.2. *Multivariate adaptive regression splines*

All MARS models were fitted in R using code available in the same 'mda' library used for fitting the BRUTO/GAM models. We also evaluated the closely similar 'polymars', available in the 'polspline' library for R, but found this to be much slower. It also differs in some key respects from the original formulation of MARS.

MARS is a procedure for fitting adaptive non-linear regression that uses piecewise basis functions to define relationships between a response variable and some set of predictors (Friedman, 1991). Basis functions are defined in pairs, using a knot or value of a variable that defines an inflection point along the range of a predictor, e.g.,

$$\text{bf}_n = \max(0, 1.0 - \text{SegFlow}) \quad \text{and}$$

$$\text{bf}_{n+1} = \max(0, \text{SegFlow} - 1.0).$$

In this example the knot takes a value of 1, and the values of bf_n can therefore be seen to have a value of 1 when SegFlow is 0, declining to 0 as SegFlow approaches 1. Values remain fixed at 0 at values of SegFlow greater than 1. By contrast, bf_{n+1} (the pair to bf_n) takes a value of SegFlow − 1 when SegFlow is greater than 1, but otherwise takes a value of 0. Within the model, coefficients applied to each of the basis functions define the slopes of the non-zero sections. Use of a single basis function allows the fitting of a non-zero slope within part of the range of a predictor variable (Fig. 1a), while the fitting of two basis functions for a predictor variable in a linear regression allows specification of different slopes within different parts of its range (Fig. 1b). More than one knot (i.e., more than one pair of basis functions) can be specified for a predictor variable, allowing complex non-linear relationships to be fitted. Alternatively, the basis functions can be envisaged as a new predictor matrix, in which one or more columns that are basis functions replace each predictor variable in the original data.

When fitting a MARS model, knots are chosen automatically in a forward stepwise manner (Hastie and Tibshirani, 1996). Candidate knots can be placed at any position within the range of each predictor variable to define a pair of basis functions. At each step, the model selects the knot and its corresponding pair of basis functions that give the greatest decrease in the residual sum of squares. Knot selection

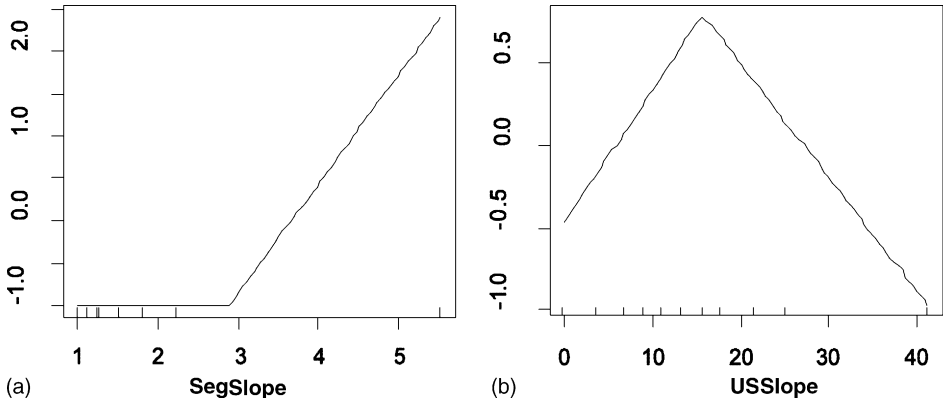


Fig. 1 – Responses of varying complexity fitted to different predictors by a MARS model. (a) A single knot was specified at a value of 2.9, but only the right-hand basis function was retained in the final model; (b) a single knot was specified at a value of 15.5, and both basis functions were retained.

proceeds until some maximum model size is reached, after which a backwards-pruning procedure is applied in which those basis functions that contribute least to model fit are progressively removed. At this stage, a predictor variable can be dropped from the model completely if none of its basis functions contribute meaningfully to predictive performance. The sequence of models generated from this process is then evaluated using generalized cross-validation, and the model with the best predictive fit is selected.

Two novel features are possible when using MARS. First, interactions between variables can be fitted, but rather than fitting a global interaction between a pair of variables, these are specified using basis functions. As each basis function only describes variation for part of the range of its variable, interactions are specified locally, i.e., the interaction effect is confined to the sub-ranges of the two variables described by the non-zero parts of the basis functions, rather than across the full range of both variables. The R implementation of MARS also allows for the fitting of multiple response variables, which allows a model to be fitted that simultaneously relates variation in the occurrence of all species to the environmental predictors in one analysis. In this case knots are selected based on their ability to reduce the residual sum of squares, averaged across all species. The final MARS model then uses a common set of basis functions for all response variables, but individual regressions are used to relate variation in each species to the final set of basis functions (i.e., to calculate unique coefficients for each basis function per species).

The current implementation of MARS in R uses least squares fitting appropriate for data with normally distributed errors. With binomial data this frequently results in the range of fitted values being erroneously expanded beyond their normal 0–1 range, e.g., from −0.2 to 1.2 or more. Rather than simply truncating these values, we used the procedure described by Friedman (1991) in which he proposes use of a GLM to constrain them within the correct range for presence–absence data. This was achieved by fitting a MARS model using the standard R code, extracting the basis functions from the MARS model, and computing a GLM that related these to the presence/absence of each species. Four sets of MARS models were fitted for this comparative analysis, i.e., two sets of 15 individual species models, one fitted without interactions and the second fitted with first order interactions. Two multiresponse models were then fitted, one without interactions and one with first order interactions.

2.3. *Model evaluation*

Comparison of the performance of the five sets of statistical models, i.e., individual GAM models, and individual and multiresponse MARS models fitted with and without interactions, was carried out using both the change in residual deviance as in conventional logistic regression, and the area under the receiver operating characteristic curve (ROC—e.g., Fielding and Bell, 1997). The latter indicates the ability of a model to discriminate between sites where a species is present versus those where it is absent. A score of 0.5 indicates that a model has no discriminatory ability, while a score of 1 indicates that presences and absences are perfectly discriminated. ROC areas were calculated for each of the models by evaluating the performance of a model against the species occurrence data used to define it (referred to as ROC^{train}). However, as these estimates are likely to be overly optimistic about model performance, we also used the 632+ bootstrap method (ROC^{boot}) to estimate model performance when predictions are made to independent data (Efron and Tibshirani, 1997). Bootstrapping gives similar results to a cross-validation, but is less prone to bias (Steyerberg et al., 2001).

3. Results

3.1. *Comparative performance of GAM and MARS models*

Comparison of the five sets of models relating fish presence/absence to environment indicates that the BRUTO/GAM models explained, on average, approximately 7% more units of deviance than both sets of non-interaction MARS models (Table 3). For MARS, the individual models fitted using interactions explained the greatest amounts of deviance, while the multiresponse model fitted with interactions explained

Table 3 – Summary of GAM and MARS models

Model	Deviance explained	Variables retained	ROC^{train}	ROC^{boot}
GAM individual	1505	13.2	0.863	0.847 (0.013)
MARS individual—non-interaction	1409	9.4	0.853	0.839 (0.016)
MARS multiresponse—non-interaction	1410	12	0.854	0.842 (0.016)
MARS individual with interactions	1541	9.7	0.861	0.838 (0.024)
MARS multiresponse with interactions	1473	10	0.859	0.845 (0.023)

Table values indicate: the average amount of deviance explained; the average number of predictor variables retained in the final models; area under the receiver operator characteristic curve statistics (ROC) averaged across 15 species and calculated using the training data (ROC^{train}); and ROC scores calculated using bootstrap re-sampling (ROC^{boot}) to assess performance when predicting to independent sites, with standard errors shown in brackets.

intermediate amounts of deviance. GAM models included the highest number of predictor variables (Table 3), and both the non-interaction and interaction MARS multiresponse models used more predictors than the average number used by the corresponding MARS individual models.

Average marginal changes in deviance when dropping individual predictors from the various final models (Table 4) suggest that a relatively small set of predictors plays a dominant role in explaining variation in the probability of capture for most diadromous fish species. These include correlates or drivers of key functional aspects of stream character, including accessibility from the sea (DSDist), summer temperature (SegJanT), stream size (SegFlow) and catchment-scale drivers of variation in stream hydrology, particularly those affecting variability in water flow (USSlope and USRainDays).

ROC areas calculated for the various models using their training data (ROC^{train} in Table 3; Appendix A) suggest that the GAM models marginally outperform both the individual MARS models and the MARS multiresponse model unless the MARS models are fitted with interactions. However, when the same statistic is calculated using bootstrap simulation to assess performance when predicting to independent data, standard errors on the adjusted ROC scores (ROC^{boot} in Table 3) indicate that the practical significance of any differences between the five sets of models is minimal. Comparison of ROC bootstrap scores from the non-interaction MARS individual and MARS multiresponse models indicates that the latter on average gives marginally better discrimination, particularly for species of lower prevalence (Fig. 2), such as *Galpos* and *Rhoret* (Appendix A).

Inspection of the response functions fitted by the MARS individual and multiresponse models indicates that the piecewise MARS functions generally approximated closely the more continuous curves fitted by the GAM models (Fig. 3). Most differences occurred in parts of the range of variables represented by few data points, and/or where there were wide standard errors about the GAM curves. Complex, non-linear responses were fitted for most species for a number of vari-

Table 4 – Summary of contributions of predictors to GAM individual models, non-interaction MARS individual models, and the non-interaction MARS multiresponse model

	GAM		MARS individual		MARS multiresponse		Average	
	Count	Δ-dev.	Count	Δ-dev.	Count	Δ-dev.	Δ-dev.	Rank
DSDist	15	139.8	14	177.9	15	136.7	151.4	1
SegJanT	15	106.2	14	140.8	15	126.4	124.5	2
SegFlow	14	66.2	13	108.2	15	69.5	81.3	3
USSlope	15	67.9	13	84.7	15	77.0	76.5	4
USRainDays	15	47.9	12	63.9	15	43.3	51.7	5
DSAveSlope	10	29.3	10	36.6	15	31.3	32.4	6
DSMaxSlope	15	31.2	11	30.2	15	30.5	30.6	7
SegTSeas	13	29.7	13	33.4	15	27.2	30.1	8
SegShade	14	28.0	9	29.8	15	25.1	27.6	9
USIndigForest	13	15.2	9	13.4	15	16.1	14.9	10
USPhos	9	14.1	4	8.1	15	14.6	12.3	11
USLake	8	10.6	2	4.6	15	11.3	8.9	12
USCalc	9	14.2	3	3.5	0	0.0	5.9	13
USHard	12	14.9	4	2.5	0	0.0	5.8	14
SegSlope	10	7.6	5	7.1	0	0.0	4.9	15
USPeat	11	7.2	5	3.0	0	0.0	3.4	16

Table entries indicate both the number of models for which each variable was retained as a significant predictor (Count), and the mean change in residual deviance when dropping that variable from final models (Δ-dev.). The two right-hand columns indicate changes in deviance averaged across all three modelling techniques, and their ranking, based on this average. Assessment of the contribution of environmental variables to MARS models fitted using interactions was not attempted.

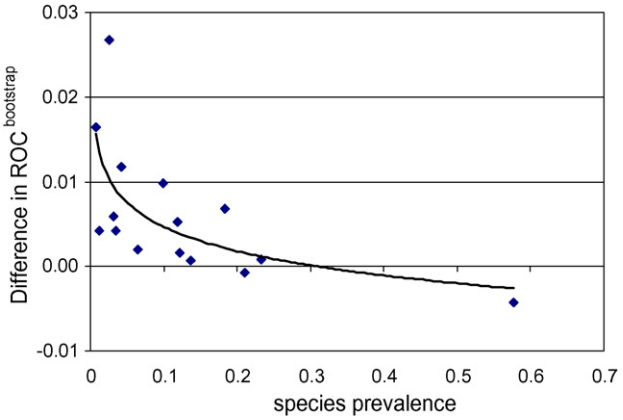


Fig. 2 – Relationship between species prevalence and the improvement in ROC score calculated using bootstrap re-sampling when fitting a multiresponse as opposed to an individual MARS model. The trend line indicates the best fit using an equation of the form δ-ROC = −0.004 ln(prevalence) − (−0.0049), and has an R^2 of 0.414.

ables, and particularly for SegJanT, along which there was particularly strong sorting of species. One variable (USAvgTNorm) was omitted from the analysis after it was fitted with overly complex curves with very high standard errors in a large proportion of the GAM models. Computation speed was similar for the BRUTO/GAM and MARS individual models, both fitting in around 1% of the time taken to fit a full backwards stepwise GAM. The MARS multiresponse models required approximately 60% of the time taken to fit all 15 MARS individual models.

4. Discussion

4.1. *Comparative performance of GAM and MARS models*

These results provide important insights into both the relative strengths of two readily available modelling methods, and the need for robust methods for assessing model performance. The value of using models capable of fitting non-linear relationships between species occurrence or abundance and environment identified in other studies (e.g., Moisen and Frescino, 2002; Olden and Jackson, 2002; Muñoz and Fellicísimo, 2004; Moisen et al., this issue) is emphasized again here by the frequency with which complex non-linear responses were fitted in our analyses. Many of these fitted responses were asymmetrical or skewed, and would therefore be difficult to fit with conventional parametric models. While we focus on the technical aspects of the modelling here, we have investigated the ecological relationships in detail elsewhere and are satisfied that they are sensible and provide insight into the ecology of these fish (Leathwick et al., 2005).

Comparison of GAM and MARS models confirms that the piecewise fitting of linear segments by the latter captures much of the information described by the more sophisticated scatter-plot smooth functions used in GAMs. Although the relative lack of performance difference between the two

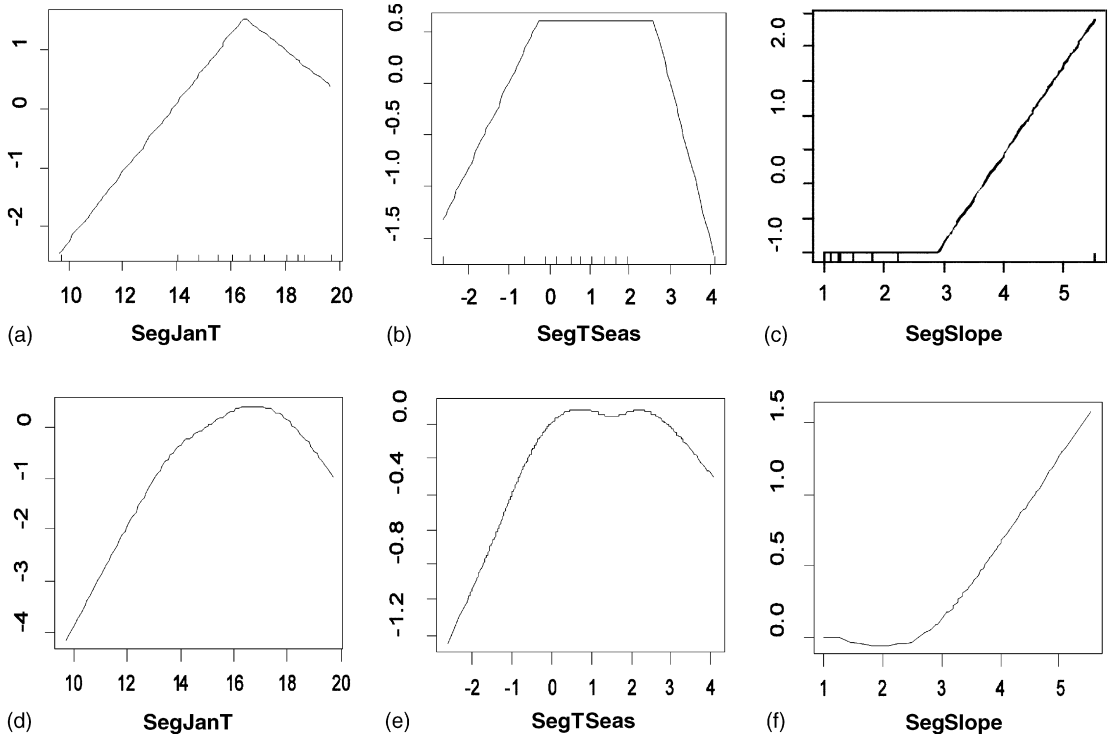


Fig. 3 – Examples of functions relating the presence/absence of Angdie to three predictors as fitted by MARS (a–c) and GAM (d–f) models.

types of MARS models and the more numerically complex GAM models was somewhat surprising, it is consistent with results from several other recent comparative studies (e.g., Elith, 2002; Moisen and Frescino, 2002; Muñoz and Fellicísimo, 2004). Together, these suggest that a number of non-linear techniques offer similar levels of performance for modelling species distributions, leaving questions of choice of technique to be influenced more by other considerations such as computational speed, ability to ignore predictors of marginal relevance, transparency of fitted relationships and the ease with which model results can be incorporated into other analyses.

While the superior speed of MARS had already been noted in other ecological studies (e.g., Moisen and Frescino, 2002), the computational efficiency of BRUTO when used as a tool to specify a GAM model was a more novel result. Although this algorithm has been available for approximately 15 years, we are not aware of its use in any ecological studies. Its speed results largely from its use of an adaptive back-fitting procedure similar to that used in MARS to guide both the selection of variables and to identify their optimal degree of smoothing. In addition, varying the penalty parameter used to specify model complexity using cross-validation to assess model performance on independent data allows a ready means to identify the most parsimonious model. In terms of speed, both these approaches clearly offer considerable advantages ($\approx$2 orders of magnitude) over more conventional stepwise fitting of GAMs, and are also likely to be considerably faster than other computationally intensive techniques that can become intractable with larger datasets (Moisen and Frescino, 2002). Both techniques readily allow inspection of the responses fitted between a species and its predictors to enable their consistency with ecological knowledge to be assessed (Austin, 2002; Austin et al., this issue). In addition, our experience confirms that the computationally more simple basis functions fitted by MARS offer advantages over GAM models when model results are used for subsequent prediction as in a GIS (Muñoz and Fellicísimo, 2004). Finally, the problems caused by our GAM models fitting overly complex curves, and which resulted in the omission of one variable from the analysis, were not encountered when that variable was included in models fitted with MARS.

The biggest disadvantage we identify in the immediate use of BRUTO and MARS for ecological modelling is that their current implementations are fitted assuming normally distributed errors, so they need to be coupled with a GAM or GLM model, respectively, to properly analyse presence/absence or count data, a procedure described for MARS by Friedman (1991). In both cases, this was achieved readily through the development of relatively simple scripts in R. Some caution may be required when the automated model selection procedure is used to specify degrees of freedom (BRUTO) or select knots (MARS) for correlated pairs of variables. Where two such variables are fitted with markedly different degrees of freedom, both the complexity of the responses that are fitted, and their relative contributions to the model outcome may vary depending on the order in which they are fitted. In practice, this means that care must be taken with choice of predictor variables, and preference given to variables that are not strongly correlated to others in the set. Predictions will still be reasonable if made to regions where the predictor variables have a similar correlation structure, but may be more problematic if predictions are made for new sites where these correlations change (Austin, 2002).

4.2. *Robust model evaluation*

The importance of robust model evaluation is clearly evident when the ROC area statistics we computed using the training data are compared with those from the more rigorous bootstrapping that assesses model performance when predictions are made for new data. For example, ROC area statistics computed using training data indicated that the GAM models provided better discrimination than either the individual MARS models or the MARS multiresponse model. However, results from the more robust bootstrap assessments indicate minimal difference in model discrimination, and instead suggest that there is a strong tendency for both the GAM and interaction MARS models to over-fit the data. That is, they have adapted to idiosyncrasies that, while occurring in the training data, had little relevance to an independent set of evaluation data. This issue is discussed further in Edwards et al. (this issue). *k*-Fold cross validation provides an alternative approach to model evaluation, and might be more feasible with some computationally demanding modelling approaches such as stepwise GAMs. However, its estimates of error rates with independent data are likely to be less precise than those derived from bootstrapping, which can be thought of as a smoothed version of cross-validation (Efron and Tibshirani, 1997). Other examples of the use of the bootstrap in evaluating modelled predictions can be found in Wintle et al. (2005) and Thomson et al. (2005).

4.3. *Simultaneous modelling of species*

Finally, we were surprised by the strong performance of the multiresponse MARS model, and this result has potentially major implications for the practicalities of analysing large datasets describing the distributions of numerous species. Our initial concern was that the distributions of species of low prevalence might be poorly analysed by such a model, because their specific relationships with environment would be submerged by information from more widespread species. However, ROC area scores for predictions of low prevalence species from the MARS multiresponse model were consistently higher than for equivalent scores from the MARS individual models. We interpret this as most likely reflecting the manner in which a MARS multiresponse model uses information across the full suite of species in selecting which predictors to use in forming basis functions, i.e., relevant predictors are included because of their strong community signal, whereas that signal might be insufficient to trigger inclusion of these predictors when fitting a single species model (Guisan et al., 1999). In addition, more prevalent species are likely to influence the selection of a larger set of relevant predictors than would be selected if rare species were analysed on their own. While this might introduce some risk that models for rare species are over-fitted, it has the advantage that the distributions of rare species are modelled within the same environmental framework as their more widespread counterparts. As a consequence, predicted distributions for rarer species are more

likely to coincide with the distributions of those common species with which they co-occur. While we performed our analyses with equal weights applied to all species, some benefit might be derived from altering weights to increase the influence of widespread species where these are of particular interest.

In conceptual terms, such a model can be seen as strongly convergent with the purposes of canonical correspondence analysis (CCA—ter Braak, 1987), an ordination technique that is widely used to relate community patterns to environment. However, this latter technique makes a number of important assumptions including that both species niche breadths and maximum abundances are equal, and that both the distributions of species responses to environment across plots and of species abundances within plots are Gaussian in shape. In addition, analysis is constrained so that the canonical axes are composed of linear combinations of environmental variables. While CCA is considered robust to departures from these assumptions, uncertainty remains about the degree to which the effects of such departures affect analysis results (see Austin, 2002). Given that the majority of ecological datasets are unlikely to meet these assumptions, we suggest that MARS offers an important alternative for the analysis of relationships between environment and community composition. As our results show, it is capable of both identifying the most parsimonious set of environmental predictors that explain variation in multiresponse composition, and robustly describing the distributions of species within the multi-dimensional space defined by these predictors. Most importantly, MARS achieves this using statistical techniques that accommodate robustly the widely varying and generally non-linear relationships that exist between species and their environment. This could also have important applications for the modelling of rare species, which although difficult to model well, are often the focus of conservation effort.

Acknowledgements

This project would not have been possible without the enormous effort that went into the assembly of the New Zealand Freshwater Fish Database, which was instigated by Bob McDowall and is maintained by Jody Richardson—particularly thanks are owed to numerous researchers who have contributed their data over several decades. Similar credit must be given to those who developed major databases describing environmental variation in New Zealand's rivers and streams, and in particular Ton Snelder, Mark Weatherhead and Helen Hurren. The inspiration for a comparative analysis of GAM and MARS models was generated by discussions at a workshop on statistical modelling of species distributions held in Riederalp, Switzerland, in August 2004. This work was funded by New Zealand's Foundation for Research, Science and Technology under contract C01X0305.

Appendix A

See Table A.1.

Table A.1 – Discrimination ability of GAM and MARS models by species

Species	GAM		MARS individual, non-interaction		MARS multiresponse, non-interaction		MARS individual interaction		MARS multiresponse interaction	
	ROC^{train}	ROC^{boot}	ROC^{train}	ROC^{boot}	ROC^{train}	ROC^{boot}	ROC^{train}	ROC^{boot}	ROC^{train}	ROC^{boot}
Angaus	0.832	0.826 ± 0.009	0.829	0.823 ± 0.005	0.827	0.823 ± 0.009	0.835	0.826 ± 0.009	0.830	0.825 ± 0.010
Angdie	0.725	0.714 ± 0.011	0.723	0.713 ± 0.012	0.716	0.708 ± 0.013	0.753	0.740 ± 0.011	0.740	0.732 ± 0.012
Galarg	0.886	0.862 ± 0.016	0.876	0.852 ± 0.021	0.871	0.855 ± 0.019	0.870	0.839 ± 0.032	0.873	0.854 ± 0.026
Galbre	0.873	0.862 ± 0.011	0.861	0.852 ± 0.012	0.864	0.858 ± 0.012	0.865	0.850 ± 0.013	0.858	0.850 ± 0.012
Galfas	0.900	0.893 ± 0.008	0.895	0.889 ± 0.008	0.893	0.888 ± 0.009	0.896	0.888 ± 0.009	0.892	0.886 ± 0.008
Galmac	0.841	0.831 ± 0.010	0.833	0.824 ± 0.011	0.836	0.829 ± 0.010	0.858	0.846 ± 0.007	0.846	0.838 ± 0.013
Galpos	0.887	0.861 ± 0.018	0.866	0.836 ± 0.022	0.871	0.850 ± 0.021	0.880	0.849 ± 0.024	0.854	0.833 ± 0.021
Gobcot	0.794	0.783 ± 0.011	0.787	0.777 ± 0.012	0.787	0.780 ± 0.012	0.798	0.785 ± 0.012	0.793	0.787 ± 0.012
Gobgob	0.931	0.911 ± 0.019	0.919	0.904 ± 0.018	0.933	0.911 ± 0.058	0.933	0.896 ± 0.056	0.934	0.910 ± 0.088
Gobhub	0.908	0.898 ± 0.010	0.895	0.886 ± 0.034	0.895	0.887 ± 0.011	0.903	0.893 ± 0.013	0.900	0.891 ± 0.013
Gobhut	0.860	0.852 ± 0.009	0.852	0.846 ± 0.009	0.850	0.845 ± 0.009	0.868	0.862 ± 0.009	0.865	0.859 ± 0.010
Geoaus	0.800	0.764 ± 0.025	0.796	0.764 ± 0.025	0.794	0.767 ± 0.025	0.826	0.776 ± 0.026	0.817	0.791 ± 0.023
Chefos	0.850	0.838 ± 0.011	0.845	0.835 ± 0.013	0.841	0.834 ± 0.011	0.851	0.839 ± 0.012	0.833	0.825 ± 0.017
Rhoret	0.976	0.948 ± 0.009	0.956	0.927 ± 0.015	0.965	0.939 ± 0.011	0.908	0.835 ± 0.107	0.967	0.933 ± 0.066
Retret	0.884	0.865 ± 0.016	0.865	0.850 ± 0.019	0.871	0.859 ± 0.016	0.871	0.849 ± 0.019	0.879	0.865 ± 0.015
Averages	0.863	0.847 ± 0.013	0.853	0.839 ± 0.016	0.854	0.842 ± 0.016	0.861	0.838 ± 0.024	0.859	0.845 ± 0.023

Table entries indicate receiver operator curve statistics computed using the training data (ROC^{train}) and bootstrap re-sampling using 300 repetitions (ROC^{boot}). Standard errors are shown for ROC^{boot}. Species names corresponding to each six-letter species code are shown in Table 1.

REFERENCES

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

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

Austin, M.P., Belbin, L., Meyers, J.A., Doherty, M.D., Luoto, M. Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory, this issue.

Borchers, D.L., Buckland, S.T., Priede, I.G., Ahmadi, S., 1997. Improving the precision of the daily egg production method using generalized additive models. Can. J. Fish. Aquat. Sci. 54, 2727–2742.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth and Brooks/Cole, Monterey, CA, 358 pp.

Edwards Jr., T.C., Cutler, D.R., Zimmermann, N.E., Geiser, L., Moisen, G.G. Effects of sample survey design on the accuracy of classification tree models in species distribution models, this issue.

Efron, B., Tibshirani, R.J., 1997. Improvements on cross-validation: the 632+ bootstrap method. J. Am. Stat. Assoc. 92, 548–560.

Elith, J., 2002. Predicting the distribution of plants. Ph.D. Thesis. The University of Melbourne, Melbourne, 304 pp.

Elith, J., Burgman, M.A., 2002. Predictions and their validation: rare plants in the Central Highlands, Victoria, Australia. In: Scott, J.M., et al. (Eds.), Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Covelo, CA, pp. 303–314.

Ferrier, S., Watson, G., Pearce, J., Drielsma, M., 2002. Extended statistical approaches to modelling spatial pattern in biodiversity: the north-east New South Wales experience. I. Species-level modelling. Biodivers. Conserv. 11, 2275–2307.

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.

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

Friedman, J.H., Meulman, J.J., 2003. Multiple additive regression trees with application in epidemiology. Stat. Med. 22, 1365–1381.

Gregr, E.J., Trites, A.W., 2001. Predictions of critical habitat for whale species in the waters of coastal British Colombia. Can. J. Fish. Aquat. Sci. 58, 1265–1285.

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

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

Hastie, T., Tibshirani, R.J., 1990. Generalized Additive Models. Monographs on Statistics and Applied Probability, vol. 43. Chapman and Hall, London, 335 pp.

Hastie, T., Tibshirani, R.J., 1996. Discriminant analysis by Gaussian mixtures. J. R. Stat. Soc. (Ser. B) 58, 155–176.

Hastie, T., Tibshirani, R., Friedman, J.H., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics. Springer-Verlag, New York, 533 pp.

Leathwick, J.R., 1998. Are New Zealand's *Nothofagus* species in equilibrium with their environment? J. Veg. Sci. 9, 719–732.

Leathwick, J.R., Austin, M.P., 2001. Competitive interactions between tree species in New Zealand's old-growth indigenous forests. Ecology 82, 2560–2573.

Leathwick, J.R., Rowe, D., Richardson, J., Elith, J., Hastie, T., 2005. Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. Freshwater Biol. 50, 2034–2052.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996. Applications of neural networks to modelling nonlinear relationships in ecology. Ecol. Model. 90, 39–52.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Monographs on Statistics and Applied Probability. Chapman and Hall, London, 261 pp.

McDowall, R.M., 1999. Driven by diadromy: its role in the historical and ecological biogeography of the New Zealand freshwater fish fauna. Ital. J. Zool. 65 (Suppl.), 73–85.

McDowall, R.M., Richardson, J., 1983. The New Zealand Freshwater Fish Survey, A Guide to Input and Output, vol. 12. New Zealand Ministry of Agriculture and Fisheries, Fisheries Research Division Information Leaflet, pp. 1–15.

Moisen, G.G., Frescino, T.S., 2002. Comparing five modelling techniques for predicting forest characteristics. Ecol. Model. 157, 209–225.

Moisen, G.G., Freeman, E., Blackard, J., Frescino, T., Zimmermann, N.E., Edwards Jr., T.C. Predicting tree species presence in Utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods, this issue.

Muñoz, J., Fellicísimo, Á.M., 2004. Comparison of statistical methods commonly used in predictive modelling. J. Veg. Sci. 15, 285–292.

Olden, J.D., Jackson, D.A., 2001. Fish–habitat relationships in lakes: gaining predictive and explanatory insight by using artificial neural networks. Trans. Am. Fish. Soc. 130, 878–897.

Olden, J.D., Jackson, D.A., 2002. A comparison of statistical approaches for modelling fish species distributions. Freshwater Biol. 47, 1976–1995.

Pereira, J.M.C., Itami, R.M., 1991. GIS-based modelling using logistic multiple regression: a case study of the Mt Graham Red Squirrel. Photogramm. Eng. Remote Sensing 57, 1475–1486.

R Development Core Team, 2004. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org.

Ripley, B.D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge.

Steyerberg, E.W., Harrell Jr., F.E., Borsboom, G.J.J.M., Eijkemans, M.J.C., Vergouwe, Y., Habbema, J.D.F., 2001. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J. Clin. Epidemiol. 54, 774–781.

ter Braak, C.J.F., 1987. The analysis of vegetation–environment relationships by canonical correspondence analysis. Vegetatio 69, 69–77.

Thomson, J.R., Fleishman, E., MacNally, R., Dobkin, D.S., 2005. Influence of the temporal resolution of data on the success of indicator species models of species richness across multiple taxonomic groups. Biol. Conserv. 124, 503–518.

Wintle, B.A., Elith, J., Potts, J., 2005. Fauna habitat modelling and mapping in an urbanising environment; a case study in the Lower Hunter Central Coast region of NSW. Aust. Ecol. 30, 729–748.

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

Ysebaert, T., Meire, P., Herman, P.M.J., Verbeek, H., 2002. Macrobenthic species response surfaces along estuarine gradients: prediction by logistic regression. Mar. Ecol. Prog. Ser. 225, 75–95.