ELSEVIER

Ecological Modelling 157 (2002) 261–280

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns

A. Elizabeth Zaniewski [a,*], Anthony Lehmann [a], Jacob McC. Overton [b]

[a] *Laboratoire d'Ecologie et de Biologie Aquatique, University of Geneva, 18 ch. des Clochettes, CH-1206 Geneva, Switzerland*
[b] *Manaaki Whenua-Landcare Research, Private Bag 3127, Hamilton, New Zealand*

Abstract

Identification of areas containing high biological diversity ('hotspots') from species presence-only data has become increasingly important in species and ecosystem management when presence/absence data is unavailable. However, as presence-only data sets lack any information on absences and as they suffer from many biases associated with the ad hoc or non-stratified sampling, they are often assumed problematic and inadequate for most statistical modeling methods. In this paper, this supposition is investigated by comparing generalized additive models (GAM) fitted with 43 native New Zealand fern species presence/absence data, obtained from a survey of 19 875 forested plots, to GAM models and ecological niche factor analysis (ENFA) models fitted with identical presence data and, in the case of GAM models, computer generated 'pseudo' absences. By using the same presence data for all models, absence data is isolated as the varying factor allowing different techniques for generating 'pseudo' absences used in the GAM models to be analyzed and compared over three principal levels of investigation. GAM models fitted with an environmentally weighted distribution of 'pseudo' absences and ENFA models selected environmental variables more similar to the GAM presence/absence models than did the GAM models fitted with randomly distributed 'pseudo' absences. Average contributions for the GAM presence/absence model showed mean annual temperature and mean annual solar radiation as the most important factors followed by lithology. Comparisons of prediction results show GAM models incorporating an environmentally weighted distribution of 'pseudo' absences to be more closely correlated to the GAM presence/absence models than either the GAM models fitted with randomly selected 'pseudo' absences or the ENFA models. ENFA models were found to be the least correlated to the GAM presence/absence models. These latter models were also shown to give the most optimistic predictions overall, however, as ENFA predicts habitat suitability rather than probability of presence this was expected. Summation of species predictions used as a measure of species potential biodiversity 'hotspots' also shows ENFA models to give the highest and largest distribution of potential biodiversity. Additionally, GAM models incorporating 'pseudo' absences were more highly correlated to the GAM presence/absence model than was ENFA. However, ENFA identified more areas of potential biodiversity 'hotspots' similar to the GAM presence/absence model, than either GAM model incorporating 'pseudo' absences.

Keywords: Biodiversity hotspots; ENFA; GAM; New Zealand ferns; Presence-only data; Spatial predictions

* Corresponding author. Tel.: +41-22-705-7105; fax: +41-22-789-4989
E-mail address: ezaniewski21@hotmail.com (A.E. Zaniewski).

0304-3800/02/$ - see front matter
PII: S0304-3800(02)00199-0

1. Introduction

Over the past two decades, advances in computational capabilities have allowed for increasingly greater and more intense statistical computations than was previously possible. These advances have lead to the development and use of numerous statistical techniques that have been used to predict species potential distribution by relating known species distributions to the spatial distribution of environmental variables (Franklin, 1995; Guisan and Zimmermann, 2000). These techniques enable a probability of occurrence to be predicted in a location where no species information is known. This has been particularly helpful in the domain of species and ecosystem management where identification and protection of areas containing high biological diversity has become tantamount, but where species data sets are often limited or lacking. Among some of the more commonly employed statistical techniques, extensively reviewed in both Franklin (1995), Guisan and Zimmermann (2000), are multiple regression models—generalized linear modeling (GLM; McCullagh and Nelder, 1989) and generalized additive modeling (GAM; Hastie and Tibshirani, 1986). Multiple regression models used to predict the spatial distribution of species are, however, commonly limited to binary data regimes (e.g. presence/absence data sets) (Franklin, 1998) that have a specific and consistent sampling strategy, as they give the most interpretable and meaningful results. Unfortunately this type of data set, while suitable for most forms of analysis (Austin, 1994), is often unavailable or unfeasible to obtain since it is usually gathered from expensive, time consuming and labor intensive systematic field surveys (Austin et al., 1994; Ferrier and Watson, 1997; Franklin, 1998).

The vast majority of data that is available today consist of presence-only data sets (i.e. where there is no information on the absence of species) collected on an ad hoc or non-stratified basis. However, although these presence-only data sets are often the most abundant and, in many cases, the only data type currently available for predictive modeling (Araújo and Williams, 2000), they are also the most difficult to successfully incorporate into statistical modeling methods. Presence-only data sets suffer three fundamental drawbacks that limit both their use and validity in models. The most obvious of these is the intrinsic lack of accurate absence data, which is a necessary component in most modeling methods. Second, is the unknown sampling bias associated with ad hoc or non-systematic data samples (Austin, 1994), where the sample is often dependent upon factors such as distance to cities, accessibility and type of environment, rather than on a stratified or systematic strategy. A third drawback of these types of data sets lies in the unknown sampling bias of rare versus common species (Ferrier and Watson, 1997). In Herbarium data sets, typically composed of presence-only data, a disproportionately high number of occurrences for rare species as compared with common species is frequently observed. These among several other factors including unknown plot size, unknown precision in species identification as well as effects of habitat disturbance, species competition and species dispersal rates (Fig. 1), which also affect presence/absence data sets, make presence-only data sets more problematic to accurately model than most systematically gathered presence/absence data sets as they drastically increase the number of unverifiable assumptions, which in turn decreases the interpretability and significance of results.

However, despite these drawbacks, presence-only data sets are nevertheless frequently the only available data type containing species occurrence information, due in part to time and/or financial constraints as well as to data collection strategies aiming at inventories instead of statistical analysis, from which predictive modeling efforts—for ecosystem and/or species management—must be determined. In recent years, several modeling techniques have been developed and studied which incorporate this type of data, among them environmental envelopes (BIOCLIM; e.g. Austin, 1994), genetic algorithms (GARP; e.g. Peters and Thackway, 1998) and ecological niche factor analysis (ENFA; e.g. Hausser, 1995; Hirzel et al., 2002b). Although several of these modeling methods have been found to appropriately model data sets containing presence-only data (Austin, 1994; Austin et al., 1994; Ferrier and Watson,

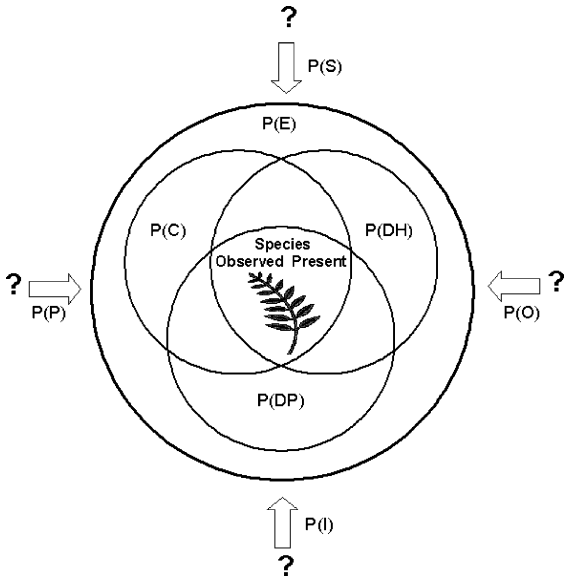

Fig. 1. A graphical representation of factors that may influence whether a species is observed present or absent in either presence/absence or presence-only data sets. Several factors that typically influence whether a species is observed present or absent in a data set are represented by circles: $P(E)$ = probability that the environment contains the species, $P(DH)$ = probability the species is absent due to effects of disturbance history, $P(DP)$ = probability species is absent due to insufficient dispersal time, and $P(C)$ = probability the species is absent due to effects of competition. Additional factors that are especially problematic in presence-only data sets are represented by four arrows: $P(S)$ = probability that the plot is sampled, $P(P)$ = probability the species is observed in that size of plot, $P(O)$ = probability the species is observed if there, and $P(I)$ = probability the species is correctly identified.

1997; Peters and Thackway, 1998; Hirzel et al., 2002a), generally they have not been found to surpass the performance of statistical modeling techniques which require data sets gathered from systematic surveys containing both presence and absence data (Austin, 1994; Austin et al., 1994; Ferrier and Watson, 1997).

GAM is one such modeling method found to have a particularly robust performance when modeling species presence/absence data (Yee et al., 1991; Austin et al., 1994, 1995; Austin and Meyers, 1996; Leathwick et al., 1996; Leathwick and Rogers, 1996; Ferrier and Watson, 1997; Bio et al., 1998; Franklin, 1998; Lehmann, 1998; Bio, 2000; Pearce and Ferrier, 2001; Lehmann et al., 2002). GAM models are a non-parametric extension of GLM models that fit predictor variables independently by smooth functions rather than by assumed linear or quadratic relationships, as is the case in GLM (Hastie and Tibshirani, 1986). In a pioneering study by Yee et al. (1991), several angiosperm species presence/absence data was fitted to GAM models to predict species distributions on the North Island of New Zealand. Since, GAM models have also been used to predict plant species distributions from presence-only data sets. In 1997, Ferrier and Watson modeled several different Australian plant groups using GAM with both presence/absence data and presence-only data supplemented with a sample of randomly generated 'pseudo-absences'. Their results revealed that although GAM models derived from presence-only data gave weaker prediction results than GAM models derived from presence/absence data, they did perform better than BIOCLIM—a heuristic model which incorporates only presence data (Ferrier and Watson, 1997).

Another heuristic modeling approach recently developed to predict species potential distribution from presence-only data is ecological niche factor analysis (ENFA; Hirzel et al., 2002b). This approach, based on Hutchinson (1957) ecological niche theory, creates habitat suitability maps that indirectly reveal species potential distribution (Hausser, 1995; Hirzel et al., 2002b). It was originally created to predict fauna distributions that are especially susceptible to erroneous or 'false' absences due to an animal's ability to disperse or hide during field surveys (Hausser, 1995; Hirzel et al., 2002b). Hirzel et al. (2002b) suggest that inclusion of these types of 'false' absences in predictive modeling could substantially bias analysis and thus propose ENFA as an alternative approach to modeling species potential distributions when there is no reliable absence data. As this approach does not incorporate species absence data, it also has potential for predicting plant species distributions from presence-only data sets.

Predicting species distributions from statistical models incorporating presence-only data sets and generated 'pseudo' absences has the potential to be a convenient and useful alternative when system-

atically gathered presence/absence data is unavailable or impossible to obtain. However, aside from the work of Ferrier and Watson (1997), few studies investigate the possibility of incorporating presence-only data sets into statistical induction methods usually reserved for presence/absence data sets. This study attempts to further examine this area of species modeling by analyzing and comparing GAM models fitted with true presence/absence data to GAM models fitted with identical presence data and computer generated 'pseudo' absences. Although the presence only data set modeled in this study does not represent an authentic herbarium or museum data set, as it is derived simply by removing absence data points from a presence/absence data set that was gathered in a known systematic manner, it does allow for an assessment to made in regards to the effectiveness of modeling species distributions from data sets lacking absence data.

The aim of this study is to assess the potential for modeling species spatial distributions on a continental scale in GAM using presence data and computer generated 'pseudo' absences, as well as, to investigate two methodologies for creating 'pseudo' absences. Results from these modeling efforts will further be compared with ENFA models derived solely from presence data. All models are fitted with identical presence data for 43 native New Zealand fern species and will be evaluated and compared at three levels of investigation: (I) model, (II) predictions, and (III) potential species richness (see Fig. 2). The latter level is used as a surrogate for predicting potential biodiversity 'hotspots' and is derived through summing of species predictions (Guisan and Theurillat, 2000; Lehmann et al., 2002). By using the same presence data for all models, absence data is isolated as the varying factor allowing different methodologies for modeling presence-only data as well as different techniques for generating 'pseudo' absences to be analyzed and compared. The main objectives are summarized as follows:

1) To compare GAM presence-only models to GAM models fitted with presence/absence data.
2) To investigate and compare different techniques of generating 'pseudo' absences.
3) To compare GAM presence-only models to ENFA models.
4) To investigate presence-only models' potential for predicting areas of increased species richness.

2. Study area

New Zealand is located in the Southern Hemisphere between approximately 35 and 47° S latitude. It has a total landmass of 270 000 km^2 spread out over three main landmasses—the North and South Islands, and Stewart Island—as well as several other smaller islands. The North Island contains the most varied landforms, which reflects both its tectonic and volcanic history (Leathwick and Rogers, 1996; Leathwick et al., 1998). While the North Island is generally characterized by a warm, moist climate, the Lake Taupo region, located in the center of the north island dominated by andesitic and basaltic volcanoes, generally experiences cooler mean annual temperatures (Leathwick and Mitchell, 1992; Leathwick et al., 1998). The South Island, divided along a Southwest–Northeast axis by the Southern Alps formed mostly from Paleozoic or Mesozoic greywacke and schist (Leathwick et al., 1998), typically experiences cooler mean annual temperatures and lower annual solar radiation than the North Island. As New Zealand contains a varied and extensively sampled indigenous forests, and an extremely varied climatic range (Wardle, 1991) it makes for an ideal study area for predictive modeling on a regional (Leathwick, 1995) or continental scale.

3. Dataset

3.1. Species data

Fern species presence/absence data was the response variable used in this study. 19 875 RECCE plots (Allen, 1992) of indigenous forests were

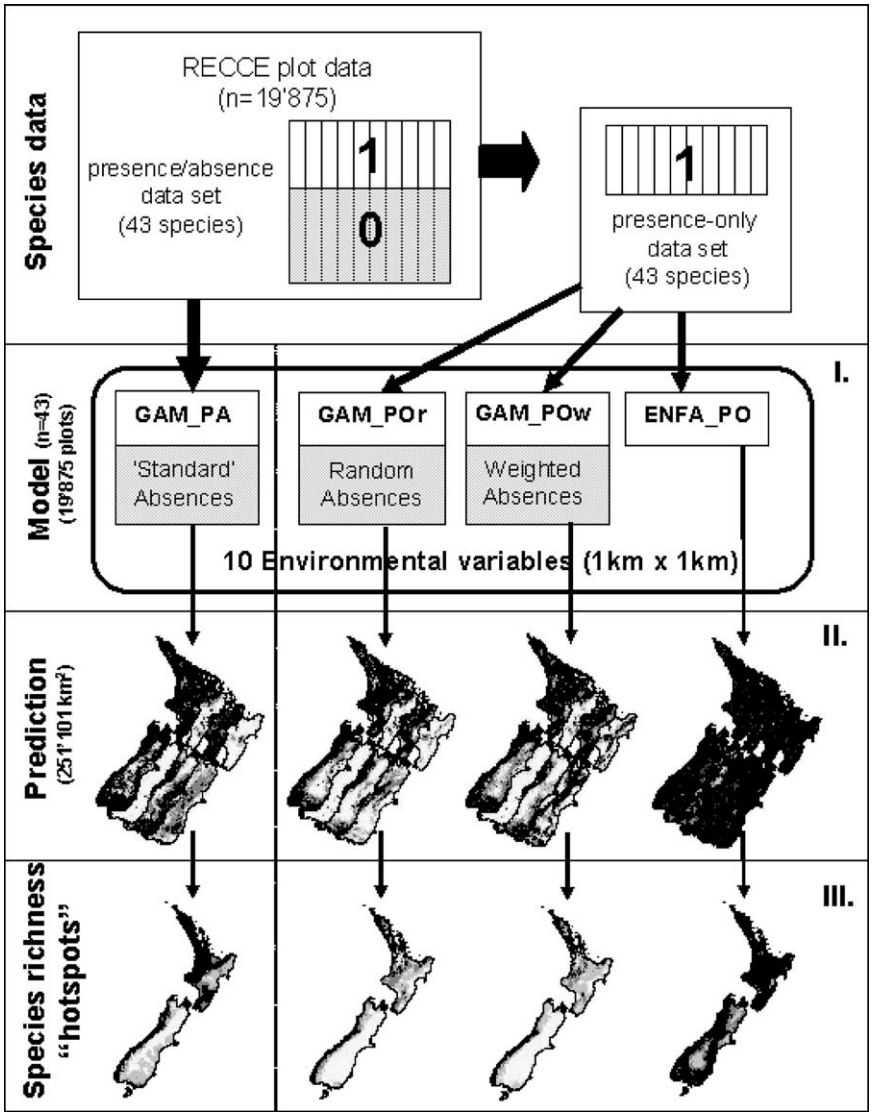

Fig. 2. Overview of the major components addressed in this study. Four modeling strategies are compared (GAM_PA, GAM_POr, GAM_POw, ENFA_PO) along three levels of investigation. (I) Models: GAM_PA models are fitted with species presence/absence data from the original RECCE data set (see Fig. 3a); GAM_POr models are fitted with the original presence data and randomly generated 'pseudo' absences (see Fig. 3b); GAM_POw models are fitted with the original presence data and randomly generated 'pseudo' absences weighted in favor of areas environmentally similar to RECCE plots (see Fig. 3c); and ENFA_PO models are fitted with the original presence data alone (see Fig. 3d). (II) Predictions: mean prediction results from the 43 species are compared. (III) Species richness: summations of species predictions are used to compare the potential for predicting areas of high biodiversity.

selected from the National Indigenous Vegetation Survey Database (NIVS) held at Landcare Research in Hamilton, New Zealand. These RECCE plots, which range from 200 to 400 m^2, were sampled over a 40-year period from 1950 to 1990 in order to inventory the remaining native New Zealand vegetation and to investigate the impact introduced browsing animals have had on native vegetation (Allen, 1992). As plots were located in vestige forest stands, the surveyed vegetation is

postulated to be at a climax succession with species occupying all available suitable habitats within the plots. Out of over 120 fern taxa observed in these selected forested plots, Lehmann et al., (2002) found 43 species to show high frequency of observation in plots (together accounting for at least 70% of observations) and likely accurate identification during sampling. To reduce the likelihood of inaccurate species identification, this same set of species was selected and included in this study. All response variable data used in this study originated from these initial 43 fern species presence/absence data. Species names and number of occurrences in plots are given in Table 1.

Table 1
Response variables

Abbreviation	Species name	Presence	Absence
ANALAN	*Anarthropteris lanceolata*	477	19 398
ASPBUL	*Asplenium bulbiferum*	4367	15 508
ASPFLA	*Asplenium flaccidum*	8973	10 902
ASPOBL	*Asplenium oblongifolium*	1164	18 711
ASPPOL	*Asplenium polydon*	2423	17 452
BLECHA	*Blechnum chambersii*	2265	17 610
BLECOL	*Blechnum colensoi*	787	19 088
BLEDIS	*Blechnum discolor*	9204	10 671
BLEFIL	*Blechnum filiforme*	1398	18 477
BLEFLU	*Blechnum fluviatile*	4865	15 010
BLEFRA	*Blechnum fraseri*	536	19 339
BLENIG	*Blechnum nigrum*	904	18 971
BLENOV	*Blechnum novae-zelandiae*	7595	12 280
CYADEA	*Cyathea dealbata*	3104	16 771
CYAMED	*Cyathea medullaris*	1204	18 671
CYASMI	*Cyathea smithii*	7319	12 556
DICFIB	*Dicksonia fibrosa*	237	19 638
DICLAN	*Dicksonia lanata*	573	19 302
DICSQU	*Dicksonia squarrosa*	6312	13 563
HISINC	*Histiopteris incisa*	3072	16 803
HYMDIL	*Hymenophyllum dilatatum*	1435	18 440
HYMFER	*Hymenophyllum ferrugineum*	1245	18 630
HYMLYA	*Hymenophyllum lyallii*	410	19 465
HYMMAL	*Hymenophyllum malingii*	95	19 780
HYPMIL	*Hypolepis millefolium*	861	19 014
HYPRUF	*Hypolepis rufobarbata*	237	19 638
LASHIS	*Lastreopsis hispida*	2194	17 681
LEPHYM	*Leptopteris hymenophylloides*	2080	17 795
LEPSUP	*Leptopteris superba*	2519	17 356
LYGART	*Lygodium articulatum*	1021	18 854
PAESCA	*Paesia scaberula*	546	19 329
PHYNOV	*Phymatosorus novae-zelandiae*	139	19 736
PHYPUS	*Phymatosorus pustulatus*	6675	13 200
PHYSCA	*Phymatosorus scandens*	1105	18 770
PNEPEN	*Pneumatopteris pennigera*	656	19 219
POLRIC	*Polystichum richardii*	389	19 486
PYRELE	*Pyrrosia eleagnifolia*	886	18 989
RUMADI	*Rumohra adiantiformis*	2756	17 119
STICUN	*Sticherus cunninghamii*	1127	18 748
TRIELO	*Trichomanes elongatum*	66	19 809
TRIREN	*Trichomanes reniforme*	3529	16 346
TRISTR	*Trichomanes strictum*	332	19 543
TRIVEN	*Trichomanes venosum*	975	18 900

Species names, abbreviation and number of occurrences in RECCE plots ($n = 19\,875$) for the 43 native New Zealand fern species modeled in GAM and ENFA.

3.2. Environmental data

Seven climatic estimates and three topographical characteristic were used (Table 2) to describe each 1×1 km of New Zealand ($n = 251\,101$), excluding areas devoid of vegetation, e.g. permanent glacial regions. Climatic variables and derived climate estimates used to indicate temperature regimes as well as light and water availability have been successfully used in previous studies to model and predict plant species spatial distributions (e.g. Leathwick, 1995; Leathwick and Rogers, 1996; Franklin, 1998; Lehmann et al., 2002). The climatic estimates used in this study were provided by Landcare Research NZ and the Department of Conservation, and were derived by fitting mathematical surfaces to long-term climatic data gathered by meteorological stations throughout the country (Leathwick and Stephens, 1998).

One of the dominating factors found to regulate plant growth and survival, and thus plant species potential spatial distribution in New Zealand, is temperature (Leathwick and Mitchell, 1992; Leathwick, 1995; Leathwick and Rogers, 1996). In this study, this was described using estimates for mean annual temperature (Ta) and temperature seasonality (Ts). The latter value is an index which represents the degree of deviation (Celsius) from the expected temperature at that site given its mean annual temperature and is calculated by subtracting the standard deviations (S.D.) for mean annual temperature from those for winter minimum (June) temperature (Leathwick, et al.,

Table 2
Predictor variables

Abbreviation	Quanitative variable name	(Minimum/maximum values)	Types
Ta*	Mean annual temperature (°C)	(−4.20/16.10)	–
Ts*	Temperature seasonality (°C)	(−3.911/5.066)	–
Sa*	Mean annual solar radiation ((MJ/m^2) per day)	(11.65/15.441)	–
Ss*	Solar radiation seasonality ((MJ/m^2) per day)	(−0.781/1.135)	–
VPD*	October VPD (kPa)	(0/0.58)	–
W*	Soil/water deficit (Mpa*day)	(0.0043/3.079)	–
R/E*	Precipitaion to potential evapo-transpiration (ratio)	(0.51/30.94)	–
S	Slope (°)	(1.5/40.0)	–
L	Lithology	–	(Metamorphic: gneiss, granite; schist. plutonic: diorite, gabbro, ultramafic quaternary: alluvium, loess, organic, sand sedimentary: limestone, strong, weak volcanic: andesite, basalt, rhyolite)
D	Drainage index	–	(Very poor, poor, impeded, moderate, good)

Names, abbreviations and values for the ten environmental variables used in this study to describe each 1 × 1 km of New Zealand. All ten predictor variables were entered into GAM-based models, however, in ENFA, soil/water deficit (W) and precipitation to

1996). This index conversion was performed in order to reduce the level of correlation inherent between both mean annual temperature and winter minimum temperature estimates, while still allowing for their influence on rate of growth and chance of survival, respectively, to be included in the model (Leathwick et al., 1996). As temperature seasonality is inversely related to continentality, positive values often indicate coastal areas having minor annual temperature fluctuations whereas negative values frequently identify areas subjected to extreme summer and winter temperature regimes.

Another major factor determining plant productivity is solar radiation (Leathwick and Mitchell, 1992; Leathwick, 1995; Leathwick et al., 1996; Lehmann et al., 2002). Here it is described by the climatic estimates for mean annual solar radiation (Sa) and the solar radiation seasonality index (Ss). This latter value was calculated in similar fashion to temperature seasonality, i.e. by subtracting the S.D.s for mean annual solar radiation from those for minimum winter (June) radiation, to reduce the level of correlation between estimates (Leathwick et al., 1996). While positive values for this index transformation often indicate high latitude sites with high winter solar radiation (i.e. south western side of the South Island), negative values tend to indicate areas with extremely low winter solar radiation (i.e. alpine region on the North and South Islands) (Leathwick et al., 1996).

Also included in models were climatic estimates indicating water availability—ratio of precipitation to potential evapo-transpiration (R/E), October vapor pressure deficit (VPD) and soil to water deficit (W)—that have been found to describe plant productivity more effectively than estimates concerning precipitation levels alone (Leathwick, 1995). Soil to water deficit (W), an annual integral of root zone water deficit, was calculated from monthly precipitation and solar radiation estimates and from soil rooting depth and texture (Leathwick and Stephens, 1998). Other environmental descriptions found to be important in vegetation modeling are those, which describe the geology and topology of an area (Franklin, 1998). For this study, the topographic situation for each plot was described by three landform variables–

slope (S), drainage (D) and lithology (L) (Leathwick, 1995). Drainage (D) was divided into five classes ranging from good to very poor and lithology (L) was grouped to create 15 soil parent type classifications. All topological and geographical data were supplied by the New Zealand Land Resource Inventory (NZLRI; Newsome, 1992).

4. Methods

4.1. GAM models

All GAM in this study were performed in SPLUS (v. 4.5, Mathsoft Inc., Seattle, WA, USA) using GRASP (Generalized Regression Analysis and Spatial Predictions; Lehmann et al., 1999), a set of S-PLUS functions developed to facilitate the modeling and analysis of species spatial distributions (Lehmann et al., 2002). Each species was individually modeled using a logistic link and a binomial error term. All models were fitted with the predictor variables listed in Table 1 using a both directional stepwise procedure to include only those variables found to be significant ($X^2 = P < 0.01$) to the model. Variable contribution was evaluated by assessing the variation in residual deviance as predictor variables were sequentially added and then dropped from the model.

GAM models were fitted with either presence/absence data or presence data supplemented with 'pseudo' absences (see Fig. 2). The presence/absence models (GAM_PA) were fitted with individual species' presence/absence data obtained from the 19 875 RECCE plots listed in Table 1 (see Fig. 3a). The two presence-only GAM models were fitted with the same species presence data as the presence/absence models, however, these models incorporated a newly derived set of computer generated 'pseudo' absences in place of those absences observed in RECCE plots (see Fig. 2). In this study each species 'pseudo' absences were created separately and in equal number to the 'true' absences found in the original RECCE presence/absence data set (Table 1). Although this methodology was based on a presence/absence data set, a similar methodology could be envisioned for a presence-only data set provided a group of species data was available. In this case, all plots or sites found to contain at least one species within a group of preselected species would be identified and counted. Then the number of presences for each individual species would be subtracted from this total number of plots to determine the number of 'pseudo' absences to create.

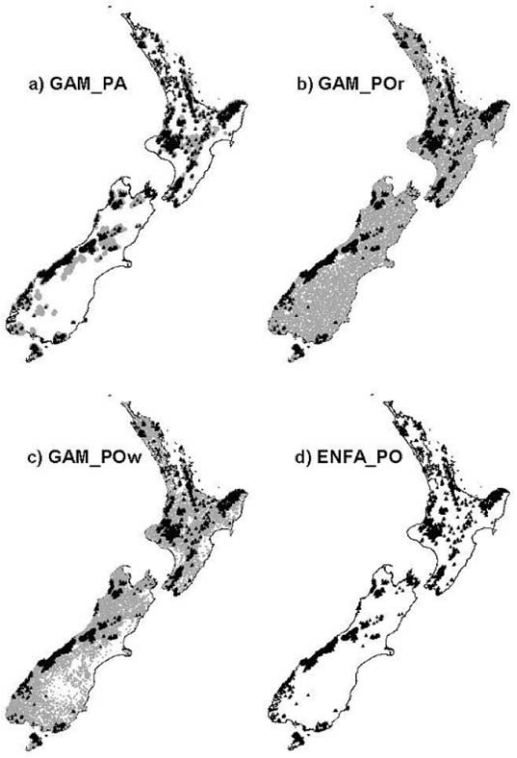

Fig. 3. Distribution of species data for ASPBUL (presence = 4367; absence = 15 508), derived from 19 875 selected RECCE plots of forested regions in New Zealand, in the four modeling methods: (a) GAM_PA: presence and absence data; (b) GAM_POr: presence data and random absences; (c) GAM_POw: presence data and weighted random absences; (d) ENFA_PO: presence data only. Black = presences; gray = absences.

'Pseudo' absence distributions were generated in SPLUS in two different ways. The first method (GAM_POr) aimed to replicate a random distribution of absences; therefore, these 'pseudo' absences were selected randomly in space from

all 1×1 km grid points ($n = 251\,101$) in New Zealand (see Fig. 3b). The second method (GAM_POw) aimed to weight a random distribution of 'pseudo' absences in favor of areas known to contain 'true' absences. However, as 'real' presence-only data sets do not contain any information on the absence of a species, and as 'true' absences in the RECCE data set are contained within the same 19 875 plots that also contain presence data, 'pseudo' absences were weighted in favor of 'true' absences by weighting a random distribution in favor of areas with environmental characteristics similar to the original 19 875 RECCE plots. This weighting factor was achieved by running a separate binomial GAM model to find the relative likelihood of a RECCE plot being sampled out of all 1×1 km grid points ($n = 251\,101$) in New Zealand. This was accomplished by assigning all 19 875 RECCE plot locations a value of '1' (presence) and an equal number ($n = 19\,875$) of randomly selected 1×1 km grid points from all of New Zealand, a value of '0' (absence). The resulting presence/absence data set ($n = 35\,760$) was then modeled in GRASP with the predictor variables presented in Table 2. The prediction results from this modeling effort (Fig. 4) were then used to weight a random selection of 'pseudo' absences for each species in favor of areas with environmental characteristics similar to those containing the original presence (and absence) recordings for any of the 43 species modeled (see Fig. 3c).

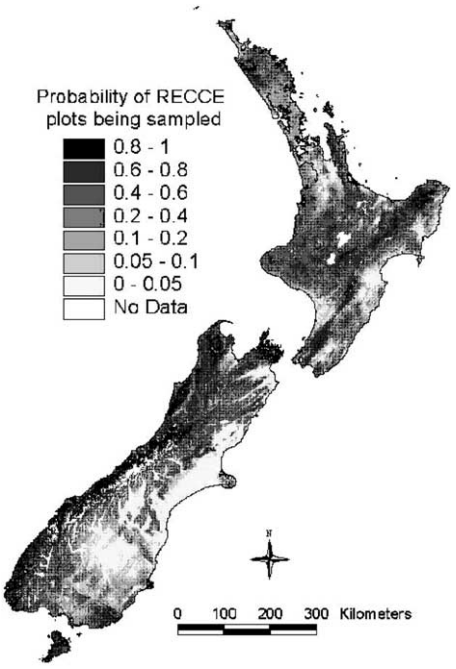

Fig. 4. Spatial prediction results from modeling the probability of a RECCE plot being sampled. Results were used to weight the selection of random absences used in GAM_POw models, in favor of areas with similar environmental characteristics to the RECCE plots.

4.2. ENFA models

All ENFA (Hausser, 1995; Hirzel et al., 2002b) models in this study (ENFA_PO) were performed in BIOMAPPER (Ver. 1.0; Hirzel et al., 2000), an autonomous program incorporating statistical tools and mapping functions created to facilitate the generation of habitat suitability models and maps (raster maps compatible with IDRISI 16 and 32) (Hirzel et al., 2000). ENFA, a modeling technique based on the ecological niche theory of Hutchinson (1957), computes habitat suitability indexes from environmental predictor variables and species presence-only data (Hirzel et al., 2002b). Similar to principal component analysis (PCA), environmental variables are summarized into two types of uncorrelated factors—marginality (representing the deviation + or − of a species' mean distribution from the global mean) and specialization (a ratio comparing the range of the global distribution to that of the species')—which together define a hyper-volume of space corresponding to the ecological niche of the species (Hirzel et al., 2002b). Habitat suitability indexes (scaled to range between 0 and 100) are computed by comparing these factors for observed species distribution to the distribution of the environmental variables in the whole area (Hirzel et al., 2002b). Habitat suitability maps produced in BIOMAPPER are a geographic representation of these suitability indexes calculated for each 1×1 km ($n = 251\,101$) of New Zealand.

Species presence data used in ENFA models was derived from the species data set listed in Table 1, by removing all absence data points for each individual species (see Fig. 2 and Fig. 3d). Once

imported into BIOMAPPER, each species map was randomly split into two equal data sets–one to calibrate the model and the other to validate it. Predictor variables listed in Table 2 were then introduced into BIOMAPPER and normalized using the Box–Cox transformation. All habitat suitability maps were created by incorporating the first four axes, representing marginality and three specialization factors, which in preliminary trials were found to explain 100% of the marginality and at least 70% of the specialization.

4.3. Comparison of GAM and ENFA

While both approaches use different modeling methodologies to predict species potential distributions from associations between environmental characteristics and species distributions, they also differ in how they select and incorporate predictor variables used to fit the model. In ENFA, all environmental predictor variables are included in the model. Hence, all ENFA species models are calculated with an identical set of differently weighted environmental variables. However, in GAM models, a stepwise procedure is used to select and incorporate only those predictor variables that explain a significant proportion of the null deviance for that species. Thus, the number and combination of environmental variables used to fit GAM models varies according to the results attained in each species model's stepwise procedure.

In our case, evaluation of model results also differs between these two approaches due to differences in the design of the modeling tools. In GAM models, all species data points are used to fit the model and evaluated through cross-validation ($n = 10$ groups). This method was chosen because it has been found to appropriately 'shake the data' (Hastie, 2001) while still allowing all available data points to be incorporated into the model. GAM models fitted with presence/absence data are evaluated by cross-validation on a ROC statistic (ranging from 0.5 to 1) (Fielding and Bell, 1997) and by examining the proportion of explained deviance (D^2). In ENFA models, however, a split-sample approach is used to randomly divide the data set into two groups–one to calibrate the model and the other to validate it. As ENFA models incorporate presence data alone, evaluation of these models is performed by analyzing the proportion of presences from the validation data set found in raster cells with a predicted suitability index (calculated from the calibration data set) greater than 50. While models ideally should be evaluated in the same manner evaluation measures such as ROC, which requires absences, are prohibitive to presence only modeling.

Presence-only models in this study are evaluated and compared relative to the 'standard' GAM presence/absence model, rather than by absolute comparison of values. In order to avoid spatial auto-correlation problems, prediction results are compared on a random sample of grid points rather than on all grid points. This is accomplished by randomly selecting 1000 1×1 km grid points from all of New Zealand ($n = 251\,101$) for each species and then comparing prediction results for each presence-only model (GAM and ENFA) to those of the GAM presence/absence model for that species. This is repeated 1000 times for each species model to obtain mean and S.D. values.

Protection of ecosystems and bio-geographic regions containing high species biodiversity—hotspots—has become an integral part of today's conservation strategies. While identification of these species rich areas is often accomplished by modeling species richness data directly, recent studies (Guisan and Theurillat, 2000; Lehmann et al., 2002) have found summation of species' predictions derived from models fitted with species presence/absence data to give comparable results. However, Herbarium presence-only data sets are often the only data source available from which to make these biodiversity assessments (Araújo and Williams, 2000). As this study aims to compare different techniques of incorporating presence-only data into statistical models and as all presence data used to fit models in this study is identical, the latter method was employed to allow for comparisons between presence-only (GAM_POr, GAM_POw and ENFA_PO) and presence/absence modeling methods (GAM_PA) (see Fig. 2).

All species plot data and environmental variables were stored in ARCVIEW (v. 3.1, ESRI, CA, USA) as shape and grid files, respectively. Queries

were performed in ARCVIEW and then exported as an ASCII file for importation into either SPLUS (v. 4.5, Mathsoft Inc., Seattle, WA, USA) or BIOMAPPER (v. 1.0, University of Lausanne, Switzerland). Results from modeling strategies were then re-imported into ARCVIEW and SPLUS where spatial predictions were evaluated.

5. Results

5.1. Observed and predicted species presence

A comparison between the observed mean probability of species presence ($n = 43$) in the sampled RECCE plots ($n = 19\,875$) and the predicted mean probability of species presence ($n = 43$) for GAM and ENFA models in randomly selected sites ($n = 10\,000$) across New Zealand is shown in Fig. 5. Mean prediction of species presence for the GAM_PA, GAM_POr and GAM_POw models, 0.093, 0.063 and 0.055, respectively, were found to be similar although lower than the observed mean probability of species presence in RECCE plots (0.115). ENFA_PO models, on the other hand, which predict habitat suitability rather than species presence, gave more elevated prediction results with an average prediction of habitat quality of 0.423. S.D. for these averages were greatest in RECCE plots (S.D. = 0.126) followed by ENFA_PO (S.D. = 0.102), GAM_PA (S.D. = 0.096), GAM_POr (S.D. = 0.066), and lastly GAM_POw (S.D. = 0.055). Although mean prediction results differ in amplitude among modeling methods, most notably ENFA_PO, they nevertheless appear to exhibit a similar pattern of prediction with individual species' predictions decreasing in concurrence with species' observed presence in RECCE plots (Fig. 6).

5.2. GAM_PA models

Cross-validation results for 43 GAM_PA models gave a mean ROC statistic of 0.864 with a S.D. of 0.076 when separated into ten random groups,

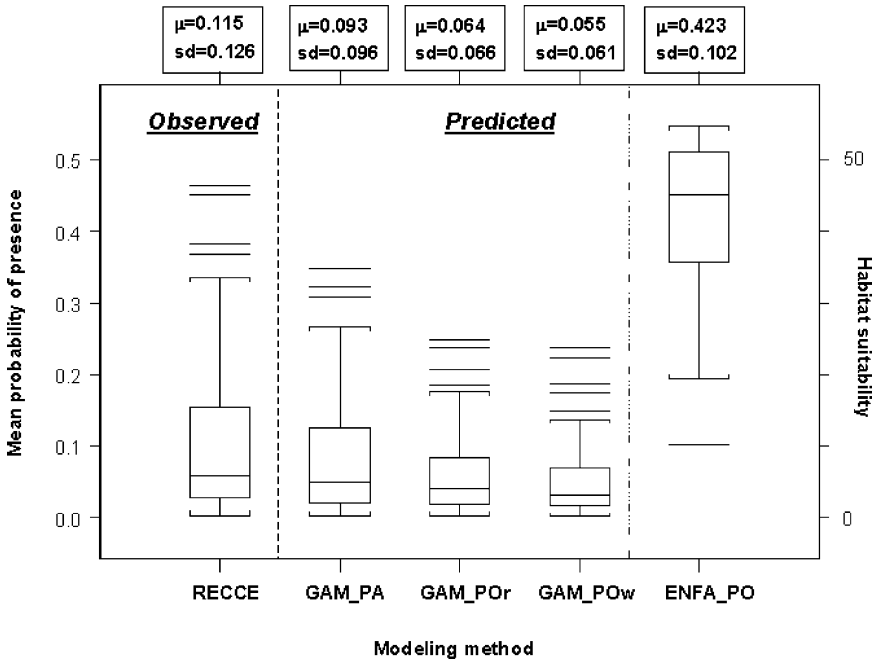

Fig. 5. Mean probability of observing fern species ($n = 43$) in sampled RECCE plots ($n = 19\,875$) compared with mean predictions of species presence for GAM_PA, GAM_POr, GAM_POw and habitat suitability for ENFA_PO models for 10 000 randomly selected 1×1 km sites across New Zealand.

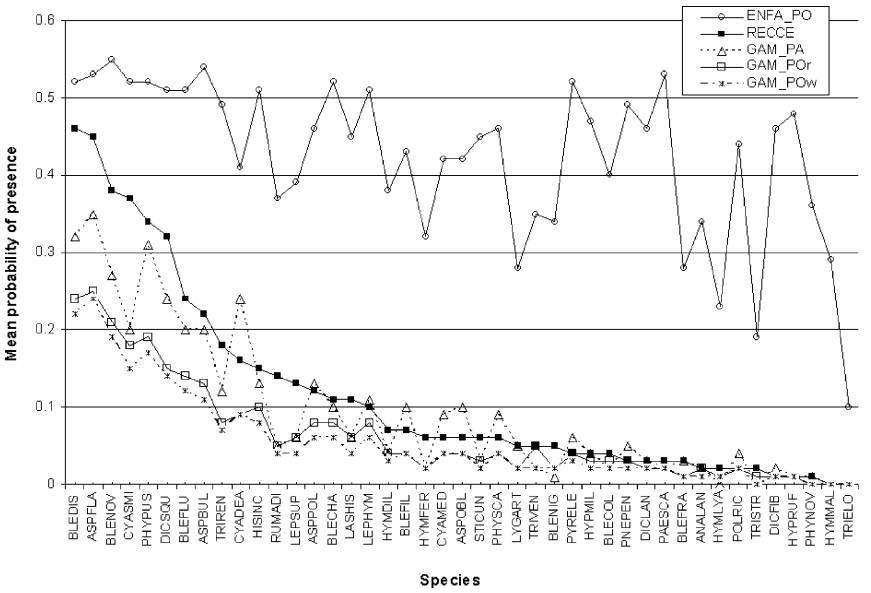

Fig. 6. Probability of observing each fern species ($n = 43$) in sampled RECCE plots ($n = 19\,875$) compared with predictions of species presence for GAM_PA, GAM_POr, GAM_POw and habitat suitability for ENFA_PO models for 10 000 randomly selected 1×1 km sites across New Zealand. Species are listed from left to right in descending order of occurrence in RECCE plots. Left = most common species; right = least common species.

and an average D^2 of 0.473 with a S.D. of 0.144 (see Fig. 7). Changes in the residual deviance resulting from the removal of an environmental

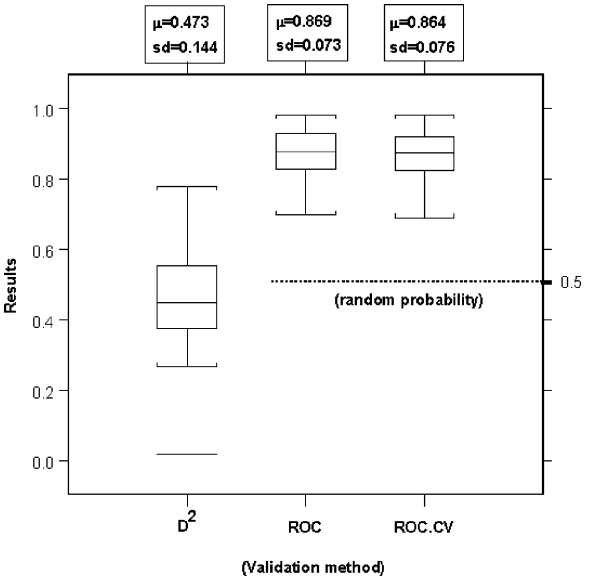

Fig. 7. Mean cross-validation results for D^2 and ROC for all GAM_PA models ($n = 43$).

variable showed mean annual temperature (Ta) to be the most significant contributing predictor variable overall to the presence/absence models, followed by mean annual solar radiation (Sa) and lithology (L). Contributions of individual predictor variables when introduced to the model alone again showed mean annual temperature (Ta) to be the greatest contributing variable to the presence/absence models, however, to a lesser degree than in the drop contribution. The next most contributing variables when introduced alone to models were mean annual solar radiation (Sa) and lithology (L). In these models, slope (S) and drainage (D) were found to give the least significant contribution when both removed and introduced alone to the model.

5.3. GAM_POr and GAM_POw models

Results for GAM models derived from presence-only data vary according to the method employed to create the 'pseudo' absences. Overall, GAM_POw models showed more similarities to the GAM_PA models than the GAM_POr mod-

els. In the GAM_POr models, changes in residual deviances when environmental variables were dropped from the model showed mean temperature (Ta) to be the most important predictor variable. However, in contrast to the GAM_PA models, both October VPD and slope (S) were the next most contributing variables. The potential contribution of variables when fitted alone to GAM_POr models also revealed differing results to those found for GAM_PA models. In these models precipitation to potential evapo-transpiration (R/E), October VPD, soil to water deficit (W) and solar seasonality (Ss) all gave greater individual contributions than did mean annual temperature (Ta). In GAM_POw models, changes in residual deviance when a variable was removed from the model showed mean annual temperature (Ta) to be the most important contributing variable followed by October VPD and precipitation to potential evapo-transpiration (R/E). When comparing the contribution of each variable alone in GAM_POw models, mean annual temperature (Ta) was again found to be the most significant contributor followed by mean annual solar radiation (Sa), precipitation to potential evapo-transpiration (R/E) and solar radiation seasonality (Ss). In both the GAM_POr and GAM_POw models, temperature seasonality (Ts) and drainage (D) were the least contributing variables when removed from or introduced alone in to the multivariate model.

5.4. ENFA_PO models

Of the ten predictor variables listed in Table 2, two variables—precipitation to potential evapo-transpiration (R/E) and soil to water deficit (W)—were found by this modeling technique to be too closely correlated, explaining similar portions of the variance which was more accurately explained by a third variable—October VPD—and were thus removed from all models. The remaining eight environmental variables, save solar radiation seasonality, were normalized using the Box-Cox transformation and incorporated in models. Solar seasonality could not be effectively normalized by this method as the resulting map gave one unique value for all raster cells.

Four explanatory axes, the first describing marginality and three others specialization, were retained for all models and accounted, on average, for 100% of the marginality and 82% of the specialization (ranging from 72.9 to 99.4% with a S.D. = 0.064) (Hirzel et al., 2002b). Marginality coefficients showed the majority of species (33 out of 43) to be primarily associated with regions with high solar radiation seasonality (Ss) (min. 0.49; max. 0.72) and low October VPD (min. −0.48; max. −0.60). On the contrary, drainage (D), lithology (L) and mean annual temperature (Ta) were found to have the least influence on species marginality. The following three axes, which reveal the degree of specialization of species, gave more mixed results and showed species to be limited within their species range by both mean annual temperature (Ta) and mean annual solar radiation (Sa) as well as by those variables indicating water availability—October VPD, drainage (D)—and lithology (L). Evaluation of species predicted habitat suitability maps showed that, on average, 89.2% (S.D. = 0.033) of presences from the validation data set were located in cells with a predicted suitability index greater than 50.

5.5. Spatial predictions

Mean correlation values comparing species' predictions ($n = 43$) from each presence-only model (GAM and ENFA) to those derived from the presence/absence GAM model for 1000 randomly selected 1 × 1 km sites in New Zealand are shown in Fig. 8. These results show GAM_POw models to be overall, more closely correlated to the GAM_PA models than either GAM_POr models or ENFA_PO models. The GAM_POw models had the highest mean correlation value of 0.725 with a S.D. of 0.122, followed by GAM_POr models, which had a mean correlation value of 0.611 and a S.D. of 0.170. ENFA_PO models gave the lowest mean correlation value of 0.410 with a S.D. of 0.152. Overall, GAM_POw models gave higher correlation values than either GAM_POr or ENFA_PO models for 38 species (Fig. 9). For 12 species, ENFA_PO models gave mean correlation values at least 0.40 lower than those given by GAM_POw models. Although these three differ-

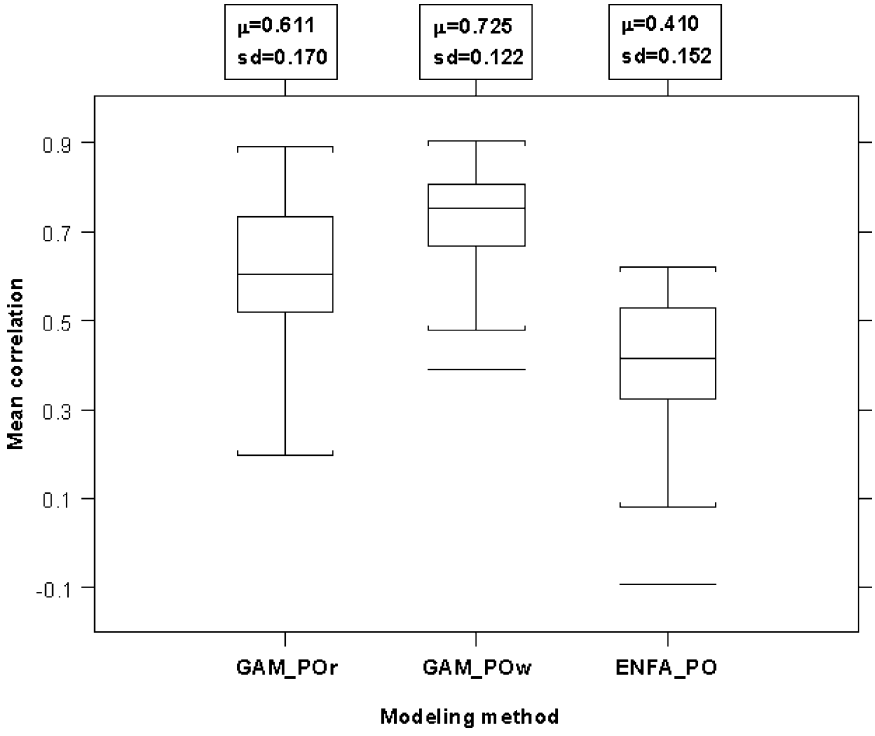

Fig. 8. Mean correlation of species predictions ($n = 43$) for each GAM_POr, GAM_POw and ENFA_PO models to the GAM_PA model for 1000 randomly selected 1×1 km grid points across New Zealand ($n = 251\,101$).

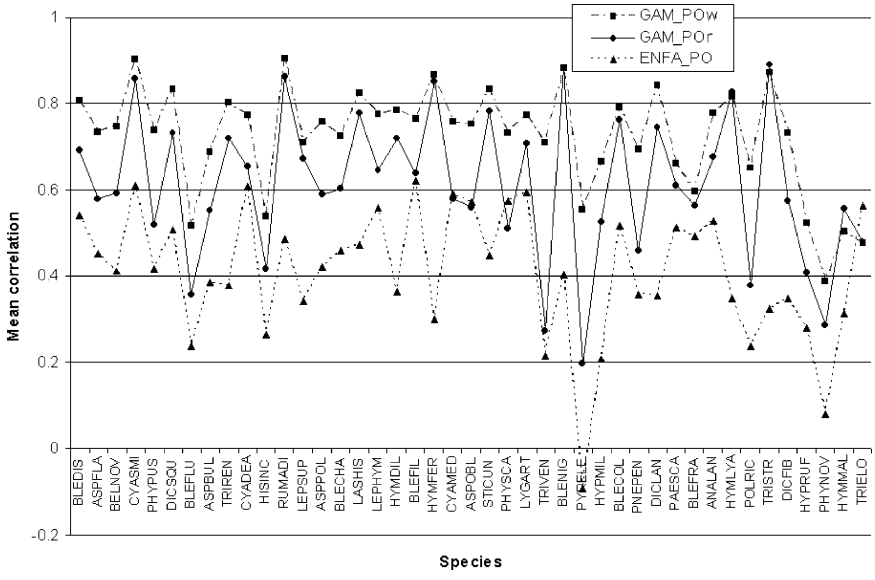

Fig. 9. Correlation of species predictions for each GAM_POr, GAM_POw and ENFA_PO models to the GAM_PA model for 1000 randomly selected 1×1 km grid points across New Zealand ($n = 251\,101$). Species are listed from left to right in descending order of occurrence in RECCE plots. Left = most common species; right = less common or rare species.

ent modeling methods were shown in an ANOVA test to have significantly ($P < 0.001$) different mean correlation results, they nevertheless appear to show similar patterns of correlation (see Fig. 9). Species predictions that are more highly correlated in one model tend to also be more highly correlated in the other modeling methods, without distinction between species frequently and rarely observed in the RECCE plots.

5.6. Summation of species predictions

Sums of the 43 species predictions—used as an indicator for biodiversity hotspots—for each of the four modeling methods across New Zealand is shown in Fig. 10. ENFA_PO gave by far the highest overall sum with a mean value of 18.139 as well as the highest S.D. of all models (6.371). GAM_PA gave the next highest mean sum of species probabilities (3.986) followed by GAM_POr (2.761) and GAM_POw (2.356). When comparing GAM_POr, GAM_POw and ENFA_PO to the GAM_PA model, used in this study as the 'standard' to which the other three modeling techniques are compared, the GAM_POw model gave the highest correlation value of 0.721, followed by GAM_POr (0.604) and ENFA_PO (0.581).

6. Discussion

6.1. Modeling methods

When comparing species observed frequency of presence in RECCE plots ($n = 19\,875$) to predicted probabilities for GAM_PA, GAM_POr, GAM_POw and ENFA_PO models for 10 000 randomly selected sites (1×1 km) across New Zealand ($n = 251\,101$) (Fig. 5), a large distinction between modeling approaches was observed. While, the three GAM-based models gave on average comparable but lower mean prediction values to the probability of species presence observed in the RECCE plots, ENFA_PO models gave, on the other hand, an overall prediction considerably more optimistic. A possible explanation for these divergent result lies in the way these two approaches differ in how and what they are actually predicting.

ENFA models, unlike GAM, do not predict probability of species presence, but rather how suitable a habitat may be for a species—habitat suitability index—from presence data alone. As ENFA aims to identify the most suitable habitat, indexes are scaled to range between 1 and 100 with the most optimal habitat having a value of 100 (Hirzel et al., 2002b). Predictions in GAM, however, are not scaled. Further, in GAM models both presence and absence data are necessary components, with absence data presumed to indicate areas where species are not present due to a negative species-environmental relationship. This unfortunately, is not always the case as species absences may be due to a variety of other reasons such as fire-related disturbance history (Guisan et al., 1999) and rate of dispersal (Hirzel et al., 2002a) among others (see Fig. 1), not related to the direct suitability of the habitat. Incorporation of this type of absence data in statistical modeling strategies can introduce too many unconfirmed assumptions and lead to less optimal models (Iverson and Prasad, 1998; Guisan et al., 1999), and in certain cases better prediction results can be obtained from models excluding this data entirely (Hirzel et al., 2002a). However, as RECCE plots were located in remnant forest stands presumed to be at climax succession and devoid of influential human-related disturbances, this is a less likely situation. In this case, exclusion of 'reliable' absence data from models such as ENFA could inversely raise model predictions in comparison to other modeling techniques that are capable of incorporating 'reliable' absence data, such as GAM, to fit the model. Thus pure presence-only modeling methods such as ENFA which exclude all absence data (reliable or not) are more likely to give potential distributions that more closely resemble the fundamental niche of the species, whereas alternative presence-only modeling methods which require the inclusion of 'pseudo' absences, as is the case with GAM, are more likely to reflect the natural distribution or realized niche of the species. Furthermore, the predictive scale of these two approaches differs. Presence data is

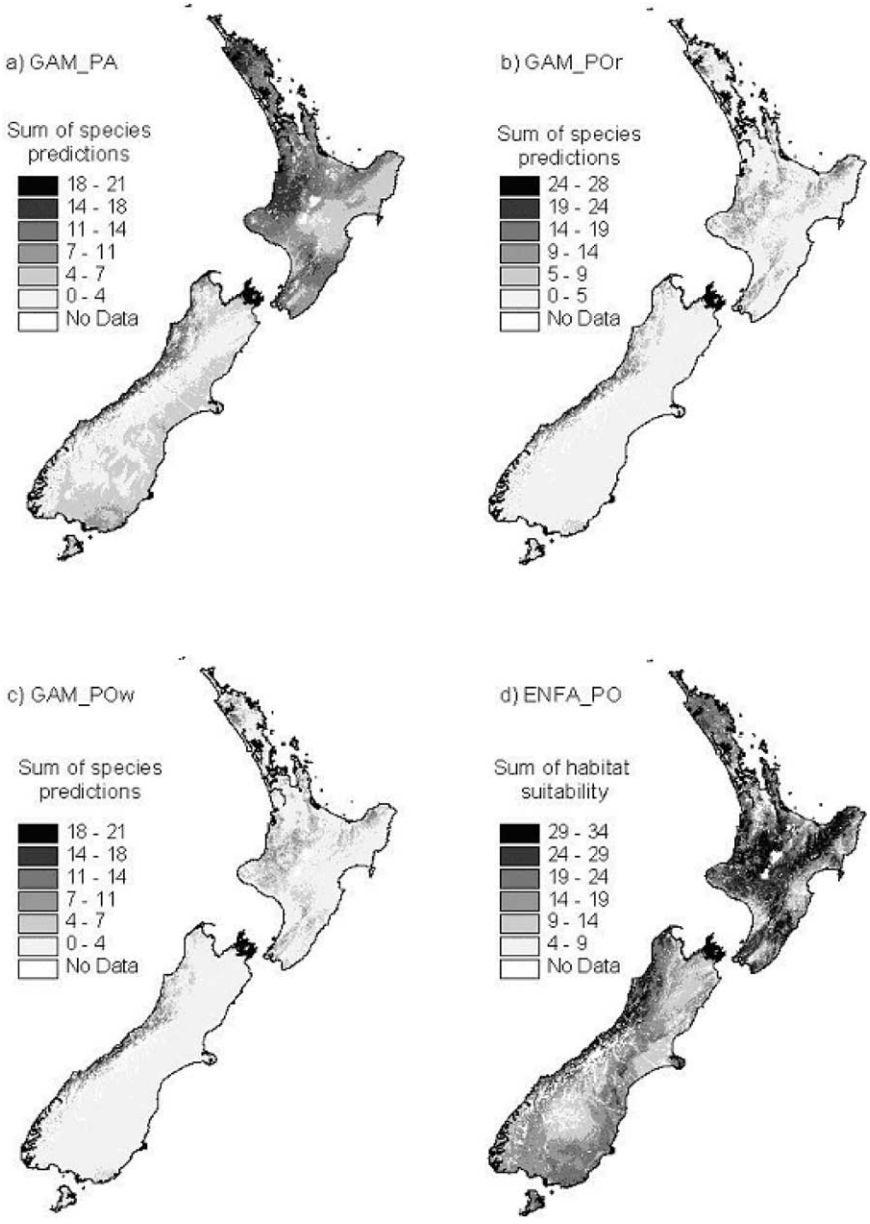

Fig. 10. Potential biodiversity 'hotspots' derived from summing species predictions ($n = 43$) for each of the four modeling methods: (a) GAM_PA; (b) GAM_Por ; (c) GAM_POw; (d) ENFA_PO.

obtained from 19 875 RECCE plots that range in size from 200 to 400 m^2, whereas predictions are made over a grid containing 251 101 1×1 km plots. Thus in ENFA, results express the likelihood of finding suitable habitat for a given species within a 200–400 m^2 plot within a given 1×1 km grid. In GAM models, however, computer generated 'pseudo' absences derived from much larger grid plots (1×1 km) than the original RECCE data plots are also included, thus further complicating comparison of GAM and ENFA model predictions.

Evaluation of GAM_PA models was determined through cross-validation on a ROC and D^2 statistic (see Fig. 7). This method was chosen to allow all available data points to help fit the model as well as to verify, through comparison with the ROC, that the model was not 'over fit' by an excess of predictor variables (Hastie, 2001). Results from this method of assessing model performance revealed deviance values, which are often very low (<0.2) in large binomial data sets (Lehmann et al., 2002), as well as both ROC values to be quite high thus confirming that there was no over fitting of the data. Therefore, GAM was determined as a robust modeling method for predicting fern species spatial distribution from presence/absence data, which is in agreement with previous studies by Leathwick (1995), Leathwick et al. (1996), Lehmann et al., 2002. This outcome thus justified using GAM_PA spatial prediction results as the 'standard' in this study, to which spatial prediction results from GAM_POr, GAM_POw and ENFA_PO models were compared.

6.2. Spatial Predictions

For most species, predictions for GAM_POw models were found to be more closely correlated to GAM_PA models than either GAM_POr models, or ENFA_PO models (see Fig. 8). A likely reason for this outcome, is that GAM_POw models are fitted with absence data located in areas having similar environmental attributes to the absence data used to fit the GAM_PA models (see Fig. 2a and c). In turn, this could explain why GAM_POw models were also found to incorporate similar numbers and combinations of environmental variables as GAM_PA models, which was not the case for GAM_POr models. Although ENFA and GAM methodologies differ, ENFA_PO models selected predictor variables more closely resembling those chosen by the GAM_PA model, than did the GAM_POr models. However, despite this apparent similarity ENFA_PO models proved the least correlated to the GAM_PA models. Although the three different modeling techniques (GAM_POr, GAM_POw, ENFA_PO) were shown to give significantly ($P<0.001$) different correlation results when compared with GAM_PA models, they did exhibit similar patterns of correlation on a per species basis (Fig. 9) with certain species being more weakly or strongly modeled by all methods.

While differences in model performance between abundant and rare species has been cited in the literature (Ferrier and Watson, 1997; Iverson and Prasad, 1998; Guisan et al., 1999; Guisan and Theurillat, 2000), no such distinction was observed when comparing these correlation values (Fig. 9) despite drastic differences in the quality and quantity of data included in models. In ENFA_PO models, only presence data is included in the model, thus those species not frequently observed in RECCE plots (i.e. rare species) had fewer data points fitting the model than more commonly observed species. However, for both GAM_POr and GAM_POw models the number of data points, presence and absence, used to fit each species' model was fixed to the total number of RECCE plots ($n=19\,875$) in the data set. This was a logistical restriction that simplified the modeling procedure and not an intrinsic limitation to either GAM or GRASP. Thus in these latter models, those species with higher frequencies of occurrence in the data set (i.e. common species) had fewer 'pseudo' absences incorporated into the model than those species with lower frequency of occurrence (i.e. rare species).

Although this study attempts to simulate presence-only models, it does not in fact incorporate true presence-only data into the models. Data used in this study was acquired from a systematic survey where rare species are those species having fewer records (presences) in the data set than common species. However, in typical presence-only data sets (e.g. Herbarium collections) rare species are not always those species recorded with the least frequency, as is the case in systematically gathered data sets. On the contrary, due to data gaps and biases (Peters and Thackway, 1998), these data sets often tend to record rare species with greater frequency than common species. This prejudice towards rare species observed in most Herbarium data sets may be attributed to the ad hoc nature with which the data is gathered (Austin, 1994) as well as to the natural inclination

of people to take more notice in what is rare over what is common (Ferrier and Watson, 1997). Ferrier and Watson (1997) cite this as an explanation for why they observed rare species to perform better in presence-only models than common species when often the opposite is the case (Iverson and Prasad, 1998; Guisan and Theurillat, 2000). As presence data used in this study was acquired from a systematic survey, these aforementioned biases, along with several other previously discussed disadvantages (e.g. inconsistent sampling strategy, plot size, and species identification) typically associated with presence-only data sets (see Fig. 1), are presumed to be minimal or non-existent.

6.3. Potential species richness

When comparing models' sums of predictions, ENFA_PO was shown to give not only the greatest overall values of summation, but also the largest distribution of potentially biodiverse areas over geographical space (see Fig. 10). This was not surprising, as ENFA_PO models have already been found to give overall higher and more generous predictions than other GAM-based models in this study due in part to differences in the type of data incorporated in the model (i.e. presence data alone) as well as to the type of predictions being made (i.e. habitat suitability). Likewise, correlation results comparing sums of species predictions for the three presence-only models (GAM_POr, GAM_POw, ENFA_PO) to the 'standard' GAM_PA model (Table 2), gave results similar to what was previously observed in the comparison of mean correlations of species predictions (Fig. 8), with GAM_POw being the most highly correlated followed by GAM_POr and ENFA_PO.

While the statistical validity of a model is important, so too are the accuracy and usefulness of its results (Fielding and Bell, 1997), particularly in species conservation efforts. Despite differences in methodology and low correlation results, visual interpretation of the summation map showed ENFA_PO to identify more areas of potential biodiversity 'hotspots' similar to the 'standard' GAM_PA model, than did any of the other models. GAM_POr and GAM_POw models were able to identify potential areas of elevated biodiversity along the southwestern coast of the South Island, but they did not signal any such areas along the northeastern coast of the North Island, well identified in the ENFA_PO model (see Fig. 10). If the objective is to protect rare or endangered species, overestimating areas of potentially elevated biodiversity might be more preferable to underestimating their existence (Fielding, 1999), however, one must be aware that optimistic predictions proved false could damage the level of confidence and support within political and non-scientific circles (Welsh, 2001). Additionally, these results further support a previous study (Lehmann et al., 2002) which found it unlikely that a few dominant species would erroneously elevate predictions of biodiversity to such a degree as to invalidate the results. While probability of occurrence and habitat suitability may be derived to predict the potential spatial distribution of species within their realized niche, the summing of these predictions do not (Lehmann et al., 2002). Biotic effects such as species competition (inter- and intraspecific) and rates of extinction and colonization (Araújo and Williams, 2000) are not incorporated into the summation, and thus the resulting potential of species richness may be inflated (Lehmann et al., 2002). In order to improve predictions of potential biodiversity 'hotspots', these biotic effects should also be incorporated into the assessment (Araújo and Williams, 2000).

7. Conclusion

In circumstances where species presence/absence data collected from a systematic stratified survey is unavailable and/or unattainable, GAM and ENFA show potential for predicting species spatial distributions from presence-only data sets, such as herbarium or museum collections. In GAM presence-only models, prediction results were improved by weighting 'pseudo' absences in favor of areas environmentally similar to the observed presences. However, while GAM presence-only models appear to predict species distributions from presence-only data more

accurately than ENFA, they appear less effective than ENFA in highlighting areas of potential biodiversity 'hotspots' from summing of species predictions.

However, despite these encouraging results, one must be aware of serious limitations due to such an approach. Although this study aimed to make better use of presence-only data, it did not in fact incorporate 'true' presence-only data. Presence data used in this study was obtained from a systematically gathered presence/absence data set and thus many of the unverified assumptions and biases typically associated with presence-only data sets (e.g. inconsistent sampling strategy, plot size, and species identification) that hinder interpretation of model results, did not pose problem. While GAM presence-only models appear to be improved by weighting the location of 'pseudo' absences in favor of known presences, they were nevertheless privilege to minimally biased presence data and thus spared from sampling biases that often lead to less optimal models. Furthermore, although GAM presence-only models appear to predict species distributions that more closely resemble the GAM presence/absence model than does ENFA, comparison of these two methodologies was limited, first by the models not reporting exactly the same type of predictions—probability of species occurrence versus habitat suitability—and second by not reporting or predicting them on the same scale.

The focal point of this study was to investigate and explore modeling methods that might make better use of the vast quantity of botanical presence-only data currently available. However, as the presence data used in this study was obtained from a presence/absence data set many of the unverified assumptions and biases as well as the intrinsic limitation (i.e. lack of absence data) that hinder interpretation of model results were not investigated. To fully examine the potential of presence-only data a 'true' presence-only data set should be modeled and compared with determine whether the methodologies employed in this study to formulate both the number and location of 'pseudo' absences still hold. Furthermore, although the presence-only models investigated in this study were developed and applied to ferns species, these techniques may also show potential in animal ecology where accurate absence data is particularly difficult to obtain as many animals, unlike plants, have the ability to move and/or hide (Hirzel et al., 2002b).

References

Allen, R.B., 1992. RECCE: an inventory method for describing New Zealand vegetation. Forest Research Institute, Christchurch, Bulletin Number 176.

Araújo, M.B., Williams, P.H., 2000. Selecting areas for species persistence using occurrence data. Biological Conservation 96, 331–345.

Austin, M.P., 1994. Data capability, Sub-project 3, Modeling of Landscape Patterns and Processes Using Biological Data. Division of Wildlife and Ecology, Commonwealth Scientific and Industrial Research Organization, Canberra.

Austin, M.P., Meyers, J.A., 1996. Current approaches to modeling the environmental niche of eucalypts: implication for management of forest biodiversity. Forest Ecology Management 85, 95–106.

Austin, M.P., Meyers, J.A., Doherty, M.D., 1994. Predictive Models for Landscape Patterns and Processes, Sub-project 2, Modeling of Landscape Patterns and Processes Using Biological Data. Division of Wildlife and Ecology, Commonwealth Scientific and Industrial Research Organisation, Canberra.

Austin, M.P., Meyers, J.A., Belbin, L., Doherty, M.D., 1995. Simulated data case study, Sub-project 5, Modeling of landscape patterns and processes using biological data. Division of Wildlife and Ecology, Commonwealth Scientific and Industrial Research Organisation.

Bio, A.M.F., Alkemade, R., Barendregt, A., 1998. Determining alternative models for vegetation response analysis: a non-parametric approach. Journal of Vegetation Science 9, 5–16.

Bio, A.M.F., 2000. Does vegetation suit our models? Thesis, Faculty of Geographical Sciences, Utrecht University.

Ferrier, S., Watson, G., 1997. An evaluation of the effectiveness of environmental surrogates and modeling techniques in predicting the distribution of biological diversity. NSW National Parks and Wildlife Service.

Fielding, A.H., 1999. Machine Learning Methods for Ecological Applications. Kluwer Academic Publishers.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24 (1), 38–49.

Franklin, J., 1995. Predictive vegetation mapping: geographic modeling of biospatial patterns in relation to environmental gradients. Progress in Physical Geography 19 (4), 474–499.

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

Guisan, A., Theurillat, J.P., 2000. Equilibrium modeling of alpine plant distribution: how far can we go? Phytocoenologia, special issue.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecological Modelling, in review.

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

Hastie, T., 2001. Personal Communication.

Hastie, T., Tibshirani, R., 1986. Generalized additive models. Statistical Science 3 (1), 297–318.

Hausser, J., 1995. Mammifères de la Suisse: Répartition * Biologie * Ecologie. Commission des Mémoires de l'Académie Suisse des Sciences Naturelles. Birkhäuser Verlag, Basel.

Hirzel, A.H., Hausser, J., Perrin, N., 2000. BIOMAPPER 1.0 beta—A New Software to Compute Habitat-Suitability Maps (URL: http://www.unil.ch/biomapper). Laboratory for Conservation Biology, University of Lausanne, Switzerland.

Hirzel, A.H., Helfer, V., Métral, F., 2002a. Assessing habitat-suitability models with a virtual species. Ecological Modelling.

Hirzel, A.H., Hausser, J., Chessel, D., Perrin, N., 2002b. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology.

Hutchinson, G.E., 1957. Concluding remarks. Cold Spring Harbor Symposium. Quantitative Biology 22, 415–427.

Iverson, L.R., Prasad, A.M., 1998. Predicting abundance of 80 tree species following climate change in the eastern United States. Ecological Monographs 68, 465–485.

Leathwick, J.R., Mitchell, N.D., 1992. Forest pattern, climate and vulcanism in central North Island, New Zealand. Journal of Vegetation Science 3, 603–616.

Leathwick, J.R., 1995. Climatic relationships of some of New Zealand forest tree species. Journal of Vegetation Science 6, 237–248.

Leathwick, J.R., Rogers, G.M., 1996. Modeling relationships between environmental and canopy composition in secondary vegetation in central North Island, New Zealand. New Zealand Journal of Ecology 20, 147–161.

Leathwick, J.R., Stephens, R.T.T., 1998. Climate Surfaces for New Zealand. Landcare Research Contract Report, LC9798/126.

Leathwick, J.R., Whitehead, D., McLeod, M., 1996. Predicting changes in the composition of New Zealand's indigenous forests in response to global warming: a modeling approach. Environmental Software 11, 81–90.

Leathwick, J.R., Burns, B.R., Clarkson, B.D., 1998. Environmental correlates of tree alpha-diversity in New Zealand primary forests. Ecography 21, 235–246 (Copenhagen).

Lehmann, A., Leathwick, J.R., Overton, J.McC., 2002. Assessing biodiversity from spatial predictions of species assemblages: A case study of New Zealand ferns. Ecological Modelling.

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

Lehmann, A., Leathwick, J.R., Overton, J., 1999. GRASP User's Manuel. Landcare Research, Hamilton.

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

Newsome, P., 1992. The Vegetative Cover Map of New Zealand. Ministry of Works and Development, Wellington.

Pearce, J., Ferrier, S., 2001. The practical value of modelling relative abundance of species for regional conservation planning: a case study. Biological Conservation 98, 33–43.

Peters, D., Thackway, R., 1998. A new biogeographic regionalisation for Tasmania. Tasmanian Parks and Wildlife Service GIS Section. Report for the National Reserve System Program Component of the National Heritage Trust. Project NR002, Undertake biophysical regionalism for Tasmania.

Wardle, P., 1991. Vegetation of New Zealand. Cambridge University Press, Cambridge.

Welsh, A.H., 2001. Personal Communication.

Yee, T.W., Mitchell, N.D., 1991. Generalized additive models in plant ecology. Journal of Vegetation Science 2, 587–602.