ELSEVIER

Ecological Modelling 157 (2002) 227–247

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence☆

Jennifer Miller*, Janet Franklin

Department of Geography, San Diego State University, San Diego, CA 92182-4493, USA

Received 15 June 2001; received in revised form 25 February 2002; accepted 16 April 2002

Abstract

Generalized linear models (GLMs) and classification trees were developed to predict the presence of four vegetation alliances in a section of the Mojave Desert in California. Generalized additive models were used to provide response shapes for parameterizing GLMs. Environmental variables used to model the distribution of the alliances included temperature, precipitation, elevation, elevation-derived terrain variables (slope, transformed aspect, topographic moisture index, solar radiation, and landscape position), and categorical landform/surface composition variables. Vegetation distributions exhibit spatial dependence and therefore we used indicator kriging to derive neighborhood values of “presence” also used as predictors in the models. The models were developed using 2859 observations coded present or absent for each of the four alliances, and assessed using 960 observations. In general, all of the models were improved with the addition of the kriged dependence term. However, models that relied heavily on the kriged dependence term were less generalizable for predictive purposes. Classification tree models had higher classification accuracy with the training data, but were less robust when used for predictions with the test data. Each of the models was used to generate a map of predictions for each alliance and the results were often quite different. The predicted maps with the kriged dependence terms looked unrealistically smooth, particularly in the classification tree models where they were often selected as the most important variables, and therefore heavily influenced the spatial pattern of the resulting map predictions.

Keywords: Classification tree; Generalized linear model; Indicator kriging; Mojave Desert; Predictive vegetation modeling

1. Introduction

Recent developments in remote sensing and geographic information science have produced alternatives for mapping vegetation beyond traditional field survey and photointerpretation. One of

☆ Prepared for Workshop in Conservation Ecology. Advances in GLM/GAM Modeling: From Species' Distribution to Environmental Management, August 2001, Riederalp, Switzerland.

* Corresponding author. Tel.: +1-619-594-8034; fax: +1-619-594-4938

E-mail address: jmiller@rohan.sdsu.edu (J. Miller).

0304-3800/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0304-3800(02)00196-5

the most frequently used methods, predictive vegetation modeling, can be defined as predicting the distribution of vegetation across a landscape based on the relationship between the spatial distribution of vegetation and certain environmental variables (Franklin, 1995; Guisan and Zimmermann, 2000). It requires digital maps of the environmental variables, as well as spatial information on the vegetation attribute of interest (e.g. species, type, abundance), usually from a sample of locations. The environment–vegetation relationship can be based on observed correlation or on the theoretical or experimental physiological limitations of different plant species. The statistical methods used to quantify this relationship have become increasingly flexible in order to describe what are now generally accepted to be non-Gaussian species response curves (Austin and Smith, 1989). The result is a vegetation map that is stored in a geographic information system (GIS), which allows for collection, analysis and display of spatial data.

The models used to generate the predictive vegetation maps can be separated methodologically into two types. “Static” or “equilibrium” models make the simplifying assumptions that vegetation distribution is in temporary (or pseudo-) equilibrium with the environment (Guisan and Theurillat, 2000), and that the relationship between vegetation distribution and environmental variables detected in a sample of observed distributions extends throughout the study area (Franklin, 1995). Static models do not directly consider dynamic ecological processes such as competition, predation and disturbance, all of which can affect the spatial arrangement of vegetation. Dynamic vegetation models attempt to simulate these processes to produce more realistic process-based maps, but can be challenging to develop because they require a large number of parameters. For this reason they are beyond the scope of this study.

Static predictive models are often developed without considering the spatial pattern that exists in biogeographical data. The traditional statistical methods used to analyze the environmental–vegetation relationships are commonly based on the implicit assumption that the distribution of vegetation is random and, therefore, each observation is independent. This assumption violates one of the basic tenets of geography, the direct relationship between distance and likeness (Tobler, 1979), as well as of ecological theory, that elements of an ecosystem close to one another are more likely to be influenced by the same generating process and will therefore be similar (Legendre and Fortin, 1989). Ignoring spatial dependence in biogeographical data can lead to poorly specified models in general and inflated significance estimates for explanatory variables in particular (Legendre, 1993). Some of the spatial correlation can be explained by the independent variables used in the model. Environmental variables such as precipitation, temperature and elevation exhibit spatial dependence, some of which is responsible for the spatial patterning in vegetation distribution, but remaining spatial dependence can result from either unmeasured environmental variables or biotic processes that cause spatial clustering.

Some studies have attempted to eliminate spatial dependence by manipulating the sampling strategy to avoid autocorrelated observations (Sokal and Oden, 1978; Legendre and Fortin, 1989; Davis and Goetz, 1990; Smith, 1994), while Borcard et al. (1992) were able to separate the spatial component that was related to vegetation pattern from the environmental component using correspondence analysis. One problem caused by spatially dependent data is that each observation contributes less information and the degrees of freedom used in analyses are exaggerated. Thomson et al. (1996) used a method to modify the degrees of freedom based on spatial dependence in order to proceed with analysis. Anselin (1993) used a maximum likelihood regression method to deal specifically with spatially dependent continuous data.

Two conceptually different models were used in this study to predict vegetation presence at unsampled locations in the study area. Both were used also to assess the explicit inclusion of spatial dependence as a predictor variable. One method, generalized linear models (GLM), is basically model-driven; i.e. a pre-specified model form is fit to the data. In order to provide insight into suitable transformations of the predictor variables, a more flexible extension of GLM, generalized additive models (GAM), were used. The second method, classification tree (CT) analysis, is data-driven and allows for the development of a model whose form is directly a function of that particular data set. Both of these methods can be used for vegetation mapping because they each can be manipulated to produce a probability surface, sometimes referred to as suitability (Carver, 1991), of vegetation presence. The specific objectives of this study were to: (1) develop models that describe the presence of four vegetation alliances, two shrubland and two woodland (see the National Vegetation Classification System (NVCS)—Grossman et al., 1998), based on environmental variables using GLM and CT; (2) use indicator kriging based on observed presence/absence data to represent spatial dependence and add this variable to GLM and CT models of each alliance; (3) generate binary maps of predicted presence/absence for each alliance from each of the four models; and (4) compare predicted maps of each alliance developed from GLM and CT models, with and without spatial dependence, in terms of prediction accuracy.

2. Study area

The Mojave Desert, the smallest North American desert, covers 74,000 km^2. Its location, between the Great Basin Desert to the north and the Sonoran Desert to the south, has resulted in its characterization as an ecotone, with both Great Basin and Sonoran vegetation, as well as its own endemic species (Rowlands et al., 1982). The study area is a portion of the Mojave Desert Ecoregion within California, referred to as the Eastern California Subsection (Fig. 1). The Mojave Vegetation Mapping Project (MVMP), sponsored by the Department of Defense (DoD) and carried out by the US Geological Survey (USGS), provided data and support for this project (www.mojavedata.gov).

The physiography of the Mojave Desert region is mainly one of basins and ranges. The basins generally range from 600 to 1200 m and can have dry lakes or playas (Norris and Webb, 1990). However, the lowest point in the Mojave Desert (as well as the Western Hemisphere) is at −86 m in the Death Valley basin. The highest point in the Mojave is Telescope Peak at 3368 m. Several other ranges greater than 2000 m are also found there (Clark, Kingston, New York, and Providence Mountains).

The Mojave Desert climate is characterized by low, unevenly distributed precipitation, temperature extremes, windy conditions and high light intensity (Schoenherr, 1992). Temperatures throughout all of the Mojave Desert range from a mean minimum January temperature of −2.4 °C at Beatty, Nevada to a mean July maximum temperature of 47 °C at Death Valley (Rowlands et al., 1982). A typical daily temperature range is 28 °C (Schoenherr, 1992). Due to its position on the leeward side of the Sierra Nevada and Transverse Ranges, the Mojave Desert gets very little precipitation, and the amount varies greatly yearly as well as locationally. Winter precipitation accounts for most of the average annual precipitation. Mean annual precipitation for Death Valley was 41.4 mm, and for Victorville in the south-central region of the Mojave Desert, it was 135.7 mm (Rowlands et al., 1982).

One result of the combination of low precipitation and high evaporation rate is the presence of alkaline soils, although Mojave Desert soils vary widely in their properties. Many soils also have a high proportion of sand and coarse fragments with low organic material, while others are made up of silt and clay with high organic content (Rowlands et al., 1982). The most common land forms in this section of the Mojave Desert are alluvial fans, bajadas and alluvial plains (42%), rocky highlands (45%), washes (5%), playas (2.5%) and sand sheets and dunes (3.5%) (www.mojavedata.gov).

3. Data and methods

3.1. CTs and GLM

Decision tree-based analysis (Breiman et al., 1984) has been used in ecological studies (discussed by De'ath and Fabricius, 2000) including vegetation modeling (Moore et al., 1991a; Lees

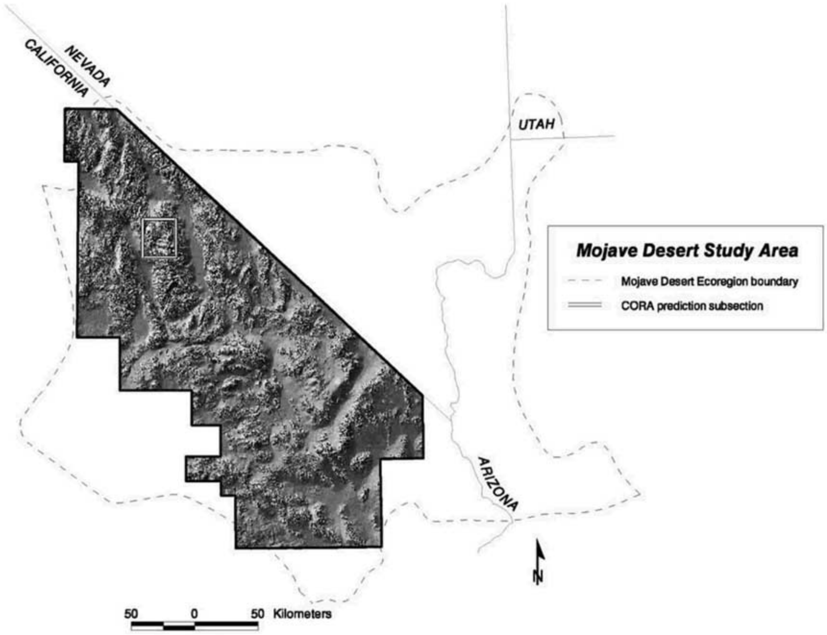

Fig. 1. The Mojave Desert Study Area (box shows mapped subsection for CORA predictions used in Figs. 5 and 6).

and Ritman, 1991; Lenihan and Neilson, 1993; Michaelsen et al., 1994; Franklin, 1998; Franklin et al., 2000; Guisan et al., 1998; Vayssières et al., 2000; Meentemeyer et al., 2001—reviewed by Franklin, 1995; Guisan and Zimmermann, 2000). Many of these studies emphasize the intuitive ecological sense of the models developed by decision tree analysis, including their ability to express complex relationships among the predictor variables that are non-linear, non-additive and hierarchical. Rather than estimating a mean value for a range of environmental variables associated with the vegetation types (as with most parametric techniques), decision trees identify specific thresholds of environmental conditions above or below which a species or vegetation type can be found (Moore et al., 1991a).

Decision trees can be used with continuous response variables such as species abundance (regression trees) or with categorical response variables such as species type (CTs). This study focuses on the prediction of a dichotomous categorical response variable (vegetation alliance presence/absence) using CTs. CT models can estimate the "probability" of class membership based on the proportion of observations of each class (presence or absence in this study) at any terminal node of the tree. These proportions can then be used to produce a discrete suitability surface analogous to probability of occurrence.

GLMs have been used extensively in vegetation modeling research (Nicholls, 1989; Le Duc et al., 1992; Austin et al., 1994; Brown, 1994; Augustin et al., 1996; Franklin, 1998; Guisan et al., 1998, 1999; Guisan and Theurillat, 2000; Vayssières et al., 2000—reviewed in Franklin, 1995; Guisan and Zimmermann, 2000). GLMs are a suite of parametric methods (see McCullagh and Nelder, 1989) that allow more flexible relationships to be specified, in the form of a number of link functions, between the response and predictor variables than linear regression models. When response data are binary, the appropriate GLM is a logistic model, which uses a logit link to describe the relationship between the response and the linear sum of the predictor variables (see Hosmer and Lemeshow, 1989). The product of logistic regression analysis can also be used to describe the probability of membership in response classes.

An even more flexible extension of GLMs is GAM, discussed thoroughly by Hastie and Tibshirani (1990). A suitable link based on the distribution family of the response data is again selected, but the relationship between the response and predictor variables is described by a number of smoothing functions rather than a coefficient, typically resulting in non-parametric shapes that are more descriptive of the data. Although GAMs have been used to develop vegetation models (Yee and Mitchell, 1991; Bio et al., 1998; Leathwick, 1998; Franklin, 1998; Frescino et al., 2001—reviewed in Franklin, 1995; Guisan and Zimmermann, 2000), here they were used to suggest an appropriate shape for GLM specification, as in Brown (1994) and Franklin (1998).

3.2. *Environmental predictor variables*

The vegetation models were developed and predictions subsequently generated using digital layers of climatic and topographic variables (Table 1) in a GIS. The relationship between climate and vegetation distribution is based largely on organisms' physiological tolerances and has been used historically to map vegetation (see Austin et al., 1994; Franklin, 1995 for review). Simple topographic variables such as elevation, slope and aspect are often empirically important, but as they represent indirect gradients related to vegetation distribution (*sensu* Austin and Smith, 1989), their predictive power is less than that of complex topographic variables (e.g. solar radiation, topographic moisture) that are more directly related to vegetation distribution (Franklin et al., 2000).

The climate variables consisted of precipitation and temperature, both of which are important in the altitudinal and latitudinal "zoning" of plants described by Hunt (1966) in Death Valley. Minimum temperature and available water have been significant in explaining the distribution of Mojave Desert shrubs (Beatley, 1975; Parker, 1991). These variables were interpolated to a resolution of 1 km^2 and include mean minimum and maximum monthly temperature for each month, and annual and quarterly mean precipitation (see www.mojavedata.gov and methods described in Franklin et al., 2001).

Terrain variables have been correlated with vegetation distribution at a finer scale than climate variables (Franklin, 1995) and those used here include both simple and complex. A United States Geological Survey (USGS) 7.5' digital elevation model (DEM) was used to provide elevation

Table 1
Environmental variables used in this study

Variable name	Variable	Range of values
Sumprecip	Average summer precipitation	11–146 mm
Winprecip	Average winter precipitation	45–579 mm
Jantemp	Minimum January temperature	−11.3–4.8 °C
Jultemp	Maximum July temperature	16.6–44.4 °C
Elevation	Elevation; from USGS 7.5' DEM	−85–3390 m
Slope	Slope	0–78°
Swness	Cosine(aspect−225°) (Franklin et al., 2000)	−1–1
Lpos4	Landscape position; Average difference between cell and neighbors; positive in valleys, neutral in mid-slope position, and negative on ridges (Fels, 1994)	−1732–2311
Solrad	Solar radiation (Dubayah, 1994)	0–383 W/m^2
TMI	Topographic moisture index; number of cells draining into a cell divided by the tangent of slope (Beven and Kirkby, 1979)	0–22.6
Landform	Geomorphic landform (Dokka et al., 1999)	29 nominal classes
Landcomp	Surface composition	6 aggregated nominal classes: CalcCarb, Evap, IgnPlut, IgnVolc, Meta, Sed

Climate variables are 1 km resolution; all others are 30 m resolution.

values; from this slope and aspect were derived. Parker (1991) found that slope was an important determinant in Sonoran vegetation distribution, while elevation was important in determining the range of several *Yucca* species in the Mojave Desert (Yeaton et al., 1985). Aspect was scaled to an index of “southwestness” using a cosine transform (cos(aspect − 225°)), where higher values indicated more xeric exposures, in order to distinguish between pole-facing (moist), neutral, and equator-facing (dry) slope aspects.

Elevation, slope, and aspect were subsequently used to derive the three more complex topographic variables: potential solar radiation, landscape position and topographic moisture index. Topographic moisture is related to the water availability of a site (Moore et al., 1991b) and hillslope position is related to soil depth, texture and potential soil moisture (reviewed in Franklin, 1995). Hillslope position and slope are also important proxy measures of soil texture (Fels, 1994), which was a significant factor in Mojave Desert (Beatley, 1975; McAuliffe, 1994) and Sonoran Desert (Parker, 1991) vegetation patterns. Aspect is strongly associated with potential solar radiation (Dubayah, 1994), but both southwestness (indirect) and potential solar radiation were tested as explanatory variables because previous studies have given conflicting evidence as to which is more strongly related to vegetation patterns (Franklin, 1998; Franklin et al., 2000; Franklin, 2002).

Valverde et al. (1996) found that landform was the most important of several topography-related variables in determining vegetation distributions. They suggest that it measures an indirect gradient along which temperature, exposure and geology vary. Two categorical geology/geomorphology variables were used here (see Dokka et al., 1999). One has land surface composition aggregated into six classes and the second has 29 landform classes.

3.3. Vegetation response variables

The vegetation variable predicted was the vegetation assemblage or type at the alliance level of NVCS. An alliance is defined as “a physiognomically uniform group of plant associations sharing one or more dominant or diagnostic species, which as a rule are found in the upper-most stratum of the vegetation” (Grossman et al., 1998). However, it should be noted that it is actually the plant *species* defining the alliance classification whose responses to the environmental variables affect the spatial patterning of the alliances (Franklin, 1995). Modeling plant species may be more rigorous, but appropriate data are sometimes less available, so for mapping and data practicality, vegetation type or alliance (as in Fischer, 1990; Lenihan and Neilson, 1993; Zimmermann and Kienast, 1999) is the response variable used in this study.

Alliance data were collected using three different sampling strategies. A modified gradient-directed sampling strategy (from Austin and Heyligers, 1989) was used to select the field plot locations where vegetation would be sampled for the Mojave Vegetation Mapping Project (MVMP) (Franklin et al., 2001). Stratification was aimed at maximizing the observed floristic variation by sampling across a broad spectrum of environmental conditions. The purpose of the survey was to define vegetation alliances by quantitative analysis of plot species composition, as well as to develop predictive models under the auspices of MVMP. Based on gradient-directed sampling 1133 observations were collected in 1998–1999. Each plot was of uniform dimension (1 km radius), and classification of the alliances was based on detailed observations of cover by species. To this, MVMP added 676 plots from five “retrospective” datasets collected between the 1970s and 1990s. Although plot size and geographic sampling intensity varied, species cover data were available so that each plot could be assigned an alliance label by vegetation experts through ordination/classification analysis (T. Keeler-Wolf, pers. comm.).

Finally, in spring 2000, MVMP had an additional 2353 locations surveyed using a modified sampling protocol (K. A. Thomas, pers. comm.). Field observers traveled along all major roads in the study area and recorded the alliance as well as several dominant species for every mile or whenever vegetation appeared to change for an area consistent with the MVMP minimum mapping area (5 ha). While these data were sufficient for modeling the spatial distribution of alliances, the lack of comprehensive species cover data rendered them insufficient for defining the alliances themselves.

After observations that extended beyond any of the 12 digital environmental layers were removed, the total sample available for modeling was 3819 plots. The main purpose of our research was to develop predictive models so despite inconsistency in the three sampling strategies, the data were combined. Limitations in the use of these models for explanation rather than, or in addition to, prediction, should be noted. Four alliances were selected for modeling (Table 2) and a dataset was developed for each, coded to represent presence/absence of that alliance. Each of these four datasets was then divided randomly into a 75% training portion (used to build the models) and a 25% test portion (used to assess the models).

3.4. Spatial dependence variable

In addition to the 12 environmental predictor variables described above, a variable used to represent spatial dependence was calculated. When it results from unspecified biotic processes or unmeasured environmental variables, spatial dependence, as evidenced by clustering in alliances, can be an important addition to predictive models. A logistic model that includes a spatial dependence variable, usually the sum of neighborhood presence values, is formally called an autologistic model (see Besag, 1972, 1974), and has been used to model plant presence/absence (Wu and Huffer, 1997). In CT models, spatial dependence has been specified indirectly with the use of geographic coordinates as variables (e.g. Franklin, 1998).

The calculation of a spatial dependence variable is straightforward when presence/absence information is known for all locations, but when only a sample has been observed the specification is more complex. From the sample data, spatial information has to be generated, usually either by simulation or interpolation. Generally, studies that have used simulation to this end have been based on small, regularly shaped study areas with either extensive or complete response data available (Besag, 1972, 1974; Augustin et al., 1996, 1998; Gumpertz et al., 1997; Wu and Huffer, 1997; Hoeting et al., 2000). While simulation is better at preserving any “roughness” characteristic of the sample dataset, when data are sparse, interpolation methods with more “smooth” effects may be more robust.

Kriging, one of the most widely used interpolation methods, attempts to optimize interpolation by dividing spatial variation into three components: deterministic variation, spatial autocorrelation (defined by a variogram), and noise (Burrough and McDonnell, 1998). The non-linear form of kriging used with binary response data (e.g. presence/absence or continuous data discre-

Table 2
Vegetation alliances modeled

Label	Alliance name	*n* test	*n* train	Dominant and indicator species	Habitat
ATCA	*Atriplex canascens*—Shrubland alliance	7	16	*A. canascens*, *Bromus madritensis*	Margins of playas
CORA	*Coleogyne ramosissima*—Shrubland alliance	21	110	*C. ramosissima*, *Atriplex confertifolia*, *Ephedra nevadensis*, *Ephedra viridis*, *Eriogonum fasciculatum*, *Salizaria mexicana*	Widespread: shallow rocky soils on upper bajadas, pediments and hill slopes
PIMO	*Pinus monophylla*—Woodland alliance	12	38	*P. monophylla*, *Artemisia tridentata*, *Quercus cornelius-mulleri*, *Nama californica*	Upper elevations: cool, moist mountain areas
YUBR	*Yucca brevifolia*—Wooded shrubland alliance	87	265	*Y. brevifolia*, *Artemisia tridentata*, *Artemisia confertifolia*, *C. ramosissima*, *Opuntia acanthocarpa*	Narrow zone, base of mountains

The data set of 3819 observations was divided randomly into a 75% train and 25% test subsets. *n* test gives the number of observations present in the $n = 960$ test dataset; *n* train gives the number of observations present in the $n = 2859$ training dataset.

tized based on a threshold value) is called indicator kriging and while the methods used are roughly the same, the output is different. Kriging produces a surface of estimated values based on specified assumptions about the three components in the model, whereas indicator kriging produces a surface with the probability that the condition coded “1” would occur (Burrough and McDonnell, 1998). When continuous data that have been thresholded are the response, it is the probability that the threshold will be exceeded that is mapped; when binary presence/absence data are used, it is the probability of presence that is mapped.

For each of the four alliances, a lattice of probability values was calculated by indicator kriging in GS+ software using four sets of the 3819 sample data points recoded “0” for absent and “1” for present for each alliance. The lattice values were interpolated to a continuous grid in ARCVIEW GIS software at a resolution of 30 m to be consistent with the environmental variables. This resulted in four maps with values that represented the probability that a specific alliance would be present in each 30 m grid cell, based on the presence/absence of that alliance in the 3819 points. To represent the neighborhood around each cell, the kriged values for the eight surrounding grid cells for each observation were summed using ARC/Info and added to the modeling datasets as the kriged dependence variable (K_x):

$$(\mathrm{K_x}) = \sum_{i=1}^{8} \mathrm{IK}_i. \qquad (1)$$

The kriged value for each cell (IK_i) can range from 0 to 1, therefore the kriged dependence term representing the neighborhood sum, K_x, can range from 0, indicating no observations of presence nearby, to 8, indicating a cluster of observations of presence (Besag, 1974; Augustin et al., 1996).

3.5. Vegetation models

Two classification tree models were developed for each of the four alliances: one using up to 12 environmental predictor variables, and a second to which the kriged spatial variable was added. The trees were pruned (based on cross-validation) to sizes that ranged from 8 to 25 terminal nodes. Exploratory GAMs were developed for each alliance with all 13 predictor variables using a stepwise (forward/backward elimination) procedure. Plots of the additive contribution of each variable to each of the four response functions were examined in order to estimate the appropriate shape for continuous variables (e.g. linear, second order polynomial, or piecewise linear) to be used in GLMs. Pairwise interaction terms that were indicated by CT or that were suggested by biophysical principles (e.g. temperature/precipitation; aspect/temperature) were also tested for significance. From this, two GLMs were developed for each of the four alliances: one that used predictor variables selected in CTs or retained in GAMs; and a second to which the kriged dependence variable was added and from which any resulting non-significant variables were removed. The GLMs were developed based on a combination of stepwise and subjective, iterative variable addition and subtraction methods with a goal of minimizing the C_p (Atkinson, 1981) statistic.

To summarize, a total of four models for each of the four alliances were developed using SPLUS statistical software: (1) a CT model with the 12 environmental variables; (2) the same CT model to which the kriged variable was added (these models will be referred to as “K_CT”); (3) a GLM with a subset of the twelve environmental variables; (4) the same GLM to which the kriged variable was added (referred to as “K_GLM” models). The four models developed for each alliance (CT, K_CT, GLM, K_GLM) were assessed using two different measures: adjusted D^2 and area under the curve (AUC) of the receiver operating characteristic (ROC) plot. The adjusted D^2 is suitable for comparison of similar *conceptual* models with different combinations of variables and interaction terms (Guisan and Zimmermann, 2000) and, as with the R^2 in linear models, a higher value indicates that the model explains more deviance. Comparisons can be made with the adjusted D^2 between GLMs with and without the kriged variable, but due to different assumptions about the error function involved in CTs, the adjusted D^2

is not appropriate for comparisons between CTs and GLMs (Austin et al., 1994; Franklin, 1998).

Another comparison among models involved assessing the classification accuracy of the resulting predictions, typically with an error matrix that shows omission (false negative) and commission (false positive) errors (see Fielding and Bell, 1997 for review). Sensitivity (fraction of observed present correctly predicted) and specificity (fraction of observed absent correctly predicted) can also be calculated from the error matrix components. The intermediate step between GLM and CT model predictions and error matrix calculation involves discretizing the vector of probability values into model predictions of presence versus absence. This "threshold" value of the probability of presence predicted by the model, above which an alliance is predicted present and below which it is predicted absent, can be optimized based on, among other things, relative importance of omission versus commission errors, and whether an alliance is rare or common in the sample.

CT model predictions are based on the proportion of observations of presence and absence at a terminal node. A probability threshold of 0.5 is often used, but when an alliance is rare, as with ATCA in this study, a lower threshold can significantly reduce the omission errors (Fielding and Bell, 1997; Franklin, 1998).

An increasingly used measure of binary classification accuracy that is threshold-independent is the ROC plot. The ROC technique has been used in medical and engineering research and can gauge how well a "receiver" (in this case a model) assigns cases into dichotomous categories (Fielding and Bell, 1997). A ROC plot is obtained by plotting sensitivity values on the y-axis against 1—specificity values for a range of probability thresholds on the x-axis. The AUC provides a measure of overall accuracy based on several different probability thresholds, and can be translated as the probability that the model will correctly distinguish between two cases (DeLeo, 1993).

Both the GLM and CT models can be used to generate predictions of alliance presence for unsampled locations based on their integration with the digital maps of the predictor variables. The implementation of the logistic models for predictions is straightforward—each predictor variable is multiplied by its model coefficient then summed to provide the linear predictor (LP) for the alliance. Second order polynomials are treated as two separate terms (one that is squared) with two different coefficients, categorical variables are treated so that only one class is used in the equation at a time, and piecewise linear variables are broken into two parts: one whose coefficient has a linear effect and the other whose coefficient has a constant effect.

In order to obtain probability values between 0 and 1, a logistic transformation of LP is used, e.g.

$$\text{Prob(Alliance)} = \frac{e^{\text{LP}}}{(1 + e^{\text{LP}})} \qquad (2)$$

Each CT model produces a set of hierarchical decision rules (see Fig. 2) based on threshold values for the continuous predictor variables and specific classes for the categorical predictor variables. The variables and rules are selected by maximizing the homogeneity of training data observations at each terminal node, which represents a set of specific environmental conditions. A tree that classifies perfectly would have only observations of a uniform class at each terminal node. The proportion of observations correctly classified at each terminal node (in practice, this value is often less than 1) can be used to represent the likely proportion of similarly classified observations of unsampled data at the environmental conditions defined by that terminal node. Therefore this proportion of presence can be used to estimate a suitability or "probability" that is analogous to the probability of presence produced by the logistic models.

A binary presence/absence map (30 m grid cells) was then developed for each alliance from each of these four models based on an optimum probability threshold. Because the entire study area is very large (ca. 56 million 30 m grid cells), a subsection of the area, containing a sufficient number of occurrences for each alliance, was used for presenting the mapped predictions. The quantity and spatial distribution of grid cells predicted to be present from the four different models were compared for each alliance.

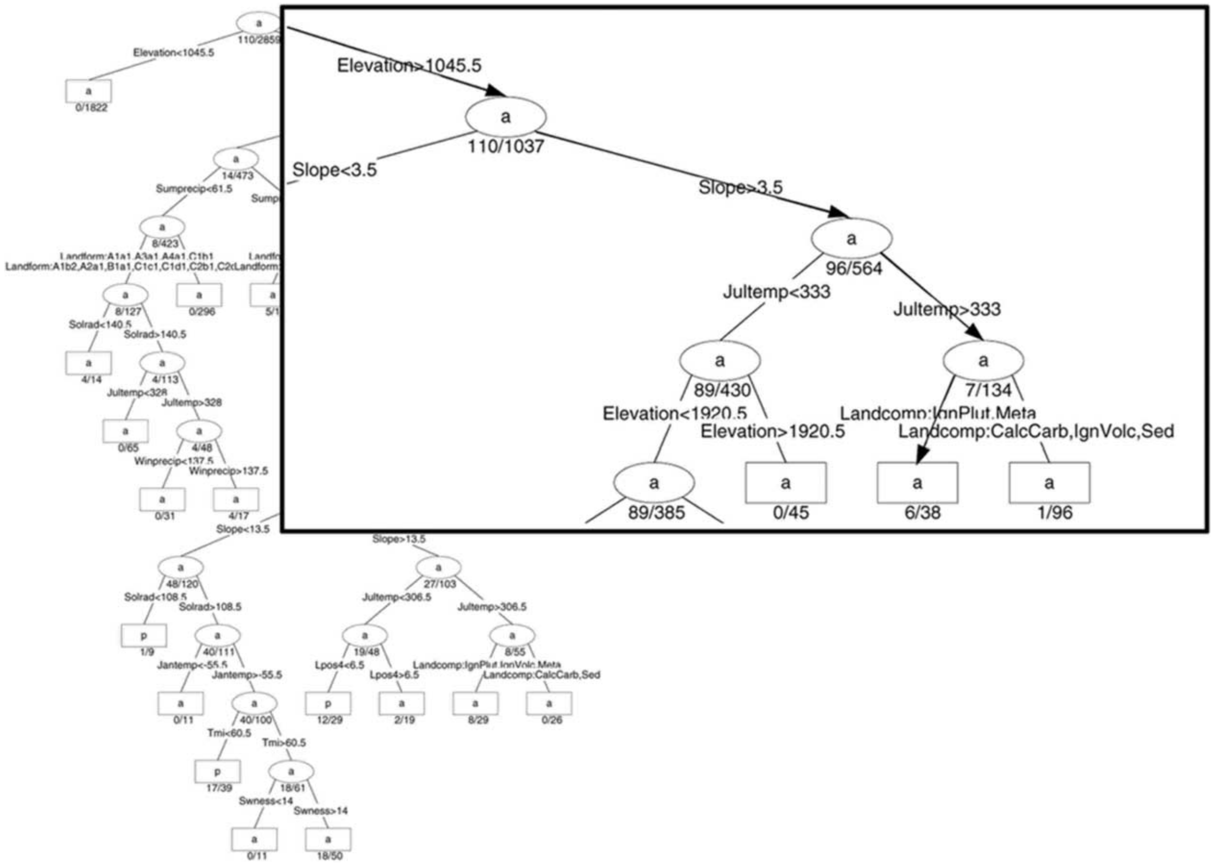

Fig. 2. Classification tree for CORA.

4. Results

4.1. Overall model assessment

Model fits based on the adjusted D^2 values were higher when the kriged dependence term was included (Table 3). The alliance PIMO has the strongest environmental correlations, specifically with elevation, temperature and precipitation and its predictive models were therefore improved the least by including the kriged dependence term. The models developed for the ATCA alliance have the lowest D^2, which is likely a result of the scarcity of ATCA observations in the dataset (23 out of 3819 total).

Using prediction errors (sensitivity and specificity), comparisons were made between CT models and GLM in terms of classification accuracy

Table 3
Adjusted D^2 for all models, and change when kriged variable was added

Models	ATCA	ΔD^2	CORA	ΔD^2	PIMO	ΔD^2	YUBR	ΔD^2
CT	0.698	–	0.572	–	0.905	–	0.675	–
K_CT	0.886	0.188	0.779	0.207	0.924	0.019	0.846	0.171
GLM	0.179	–	0.179	–	0.817	–	0.521	–
K_GLM	0.596	0.417	0.596	0.417	0.858	0.041	0.730	0.209

Table 4
Optimum probability threshold and sensitivity/specificity for all models based on test data ($n = 960$)

Models	ATCA			CORA			PIMO			YUBR		
	Optimum probability	Sensitivity	Specificity	Optimum probability	Sensitivity	Specificity	Optimum probability	Sensitivity	Specificity	Optimum probability	Sensitivity	Specificity
CT	0.2	28.6	99.5	0.1	85.7	89.9	0.2	41.2	99.7	0.2	65.5	94.6
K_CT	0.3	42.9	99.2	0.2	85.7	95.6	0.2	41.7	99.4	0.2	83.9	95.7
GLM	0.1	0.0	100	0.2	85.7	75.6	0.3	100	86.9	0.3	85.1	69.3
K_GLM	0.2	62.5	99.7	0.2	100	94.4	0.3	91.7	93.3	0.4	90.8	94.9

(Table 4). An optimum probability threshold was selected for each alliance based on plots of sensitivity and specificity at a range of probability thresholds 0, 0.1, 0.2,...−0.9 (c.f. Fielding and Bell, 1997 Fig. 2 and Franklin, 1998 Fig. 6). As is usually the case in vegetation mapping, the percent present correctly classified was considered more important than percent absent correctly classified and the thresholds were chosen accordingly. The GLM for the ATCA alliance was the poorest model: none of the probability of presence values exceeded 0.1. In addition to being the most rare of the four alliances, ATCA occurs mainly along playa edges—an environmental correlation that should be captured by one of the two categorical landform/landcomp variables. While landform was the most important variable in the CT model for ATCA, it was not selected at all in GLM. The CORA K_GLM model has the best combination of sensitivity and specificity with the test data.

Table 5 shows AUC for all four models for each alliance using both training and test data. In general, the accuracies of all models for all alliances using training data were high, ranging from an AUC of 0.89 for GLM for ATCA to 1.0 for both K_CT and K_GLM for PIMO. As expected, the accuracy of models using test data is lower than that of models using training data, with the exception of K_GLM for CORA. The difference in accuracy between training data and test data is most notable for CTs, indicating that these models are less robust and less useful for prediction. CTs are particularly sensitive to outliers (Breiman et al., 1984) and their performance could be attributed to this.

4.2. CT results

Fig. 2 shows the pruned CT model developed for the CORA alliance. The highlighted path can be translated as the following set of decision rules: “Where elevation is greater than 1045.5 m and slope is greater than 3.5° and Jultemp is greater than 33.3° and landcomp is Ignplut or Meta, 38 observations occurred in the training dataset and six were CORA.” For predictive purposes, where these same environmental conditions exist throughout the unsampled study area, there is a 0.16 (6/38) probability or suitability of CORA presence. Variables in CT models are selected to create splits that maximize the resulting node homogeneity, therefore the variables used in early splits can be considered to be more important (and in fact the amount of deviance explained at each split is calculated). When the kriged dependence term was added as a predictor variable (K_CT model), it was used in the first three splits, replacing elevation and slope and resulting in a model that had higher prediction accuracy with the test data, but at the expense of more satisfying and generalizable ecological relationships. Although slope remained an important variable even after the kriged dependence term was added, elevation was not used in a split until much later and solrad is used much earlier in the K_CT model than in the CT model (Table 6). This is probably a function of the arbitrariness with which variables with similar effects are selected in the CT models. Different variables could explain very similar amounts of deviance but sort the data quite differently. The kriged dependence term was the most important in three of the four models in

Table 5
Area under the curve (AUC) from ROC plots depicting model accuracy

Models	ATCA		CORA		PIMO		YUBR	
	Train	Test	Train	Test	Train	Test	Train	Test
CT	0.993	0.610	0.975	0.895	0.999	0.786	0.978	0.768
K_CT	0.999	0.709	0.994	0.911	1.000	0.786	0.993	0.956
GLM	0.890	0.689	0.923	0.890	0.998	0.981	0.953	0.851
K_GLM	0.998	0.655	0.981	0.983	0.998	0.985	0.987	0.981

Table 6
Variables used in classification tree models Atca2, Cora2, Pimo2, and Yubr2 are the kriged dependence varialbes

Model	Number of variables	Number of terminal nodes	Variables used in order of importance in tree construction
ATCA	8	11	Landform Swness Sumprecip Jultemp Elevation Winprecip Jantemp Slope
K_ATCA	6	8	Atca2 Landform Solrad TMI Winprecip Elevation
CORA	10	25	Elevation Slope Sumprecip Jultemp Landform Landcomp Solrad Winprecip Lpos4 TMI
K_CORA	10	24	Cora2 Slope Solrad Elevation Landcomp Sumprecip Swness Jultemp Jantemp Winprecip
PIMO	5	8	Elevation Jultemp Slope Landcomp Sumprecip
K_PIMO	6	8	Elevation Pimo2 Slope Sumprecip Winprecip Lpos4
YUBR	8	25	Sumprecip Jultemp Landcomp Landform Elevation Jantemp Winprecip Slope
K_YUBR	12	25	Yubr2 Elevation Lpos4 Sumprecip Slope Solrad Swness Winprecip Landform Landcomp TMI Jultemp

which it was used. As discussed earlier, the distribution of the PIMO alliance is highly correlated with elevation, which remained the most important variable even after the kriged dependence term is added.

4.3. GLM results

The GLMs were specified by using variables that were significant in GAMs or were used in the CT models and are summarized in Table 7. Response shapes tested were suggested by GAMs and included linear, piecewise linear and second order polynomial. In general, the climate variables explained the most deviance in the models, particularly Jantemp and Jultemp. Slope, swness, and elevation were also retained often in the models. In contrast to the CT models, the variable that explained the most deviance for PIMO was sumprecip rather than elevation. The correlation between these variables is high ($r = 0.786$) but because sumprecip is specified as a linear term and elevation is a second order polynomial (and uses two degrees of freedom), sumprecip makes a more parsimonious model.

4.4. Spatial predictions

The four binary presence/absence maps for each alliance were derived using the optimum probability threshold and compared in terms of spatial pattern of presence as well as number of grid cells predicted present. Fig. 3 shows the percentage of the mapped area predicted present by a union of all four models, divided by the total number of cells predicted to be present by each of the models individually. In other words, for PIMO, a union of all four models predicted 22% of cells in the mapped area to have PIMO present. Of these cells predicted present, there was less disparity among the four models for PIMO than for any of the other alliances. In general, the model that predicted the highest proportion of cells present was CT, but generalizations regarding the kriged variable are more difficult to make. With CORA and PIMO, both models with the kriged variable predicted a smaller proportion of cells present, but with YUBR, the kriged variable increased the proportion of cells predicted present. The ATCA model has a much higher proportion of cells predicted present by the classification tree model (and resulting higher commission errors) but no comparisons can be made between the GLM and K_GLM models because none of the GLM predictions exceeded the optimum probability threshold (all were < 0.1).

Fig. 4 shows the amount of spatial coincidence of pairs of model predictions. YUBR showed the most uniformity in model predictions spatially. The addition of the kriged variable had the least effect on the PIMO models based on the model evaluations, but there was a surprising lack of

Table 7
Results from GLMs

Alliance, variables	Response function		Deviance explained	Prob(χ^2)
		Null deviance = 197.85		
ATCA (GLM)				
Slope	Linear		20.4	< 0.00001
Jantemp	Linear		0.1221	0.7268
Swness	Piecewise linear (swness < 52, swness ≥ 52)		10.9	0.00098
Sumprecip:Jultemp	Interaction term		4.6	0.033
		Model $C_p = 169.4$		
ATCA (K_GLM)				
Slope	Linear		20.4	< 0.00001
Jantemp	Linear		0.1221	0.7268
Swness	Piecewise linear (swness < 52, swness ≥ 52)		10.9	0.00098
Atca2	Linear		86.8	< 0.00001
Sumprecip:Jultemp	Interaction term		0.002	0.964
		Model $C_p = 84.13$		
		Null deviance = 932.4		
CORA (GLM)				
Slope	2nd order polynomial		87.8	< 0.00001
Jultemp	Linear		111.8	< 0.00001
Jantemp	2nd order polynomial		85.6	< 0.00001
Elevation	2nd order polynomial		19.2	0.00007
Sumprecip	2nd order polynomial		15.9	0.00034
Winprecip	Linear		4.9	0.026
Slope:Lpos4	Interaction term		2.16	0.142
		Model $C_p = 613.2$		
CORA (K_GLM)				
Slope	2nd order polynomial		87.8	< 0.00001
Jultemp	Linear		111.8	< 0.00001
Jantemp	2nd order polynomial		85.6	< 0.00001
Elevation	2nd order polynomial		19.2	0.00007
Sumprecip	2nd order polynomial		15.9	0.00034
Winprecip	Linear		4.9	0.026
Cora2	Linear		225.6	< 0.00001
Slope:Lpos4	Interaction term		0.123	0.73
		Model $C_p = 390.3$		
		Null deviance = 403.9		
PIMO (GLM)				
Sumprecip	Linear		157.6	< 0.00001
Landcomp	Categorical variable		23.2	0.0003
Slope	Linear		29.3	< 0.00001
Elevation	2nd order polynomial		117.4	< 0.00001
Elevation:Solrad	Interaction term		2.8	0.09
		Model $C_p = 75.4$		
PIMO (K_GLM)				
Sumprecip	Linear		157.6	< 0.00001
Landcomp	Categorical variable		23.2	0.0003
Slope	Linear		29.3	< 0.00001
Elevation	2nd order polynomial		117.4	< 0.00001
Pimo2	Linear		15.7	0.00007
Elevation:Solrad	Interaction term		3.9	0.049

Table 7 (*Continued*)

Alliance, variables	Response function	Deviance explained	Prob(χ^2)
	Model $C_p = 57.6$		
	Null deviance = 1765.1		
YUBR (GLM)			
Sumprecip	2nd order polynomial	604	< 0.00001
Slope	Linear	74	< 0.00001
Jultemp		89.9	< 0.00001
Jantemp	2nd order polynomial	56.3	< 0.00001
Elevation	2nd order polynomial	34.6	< 0.00001
Sumprecip:Jultemp	Interaction term	5.6	0.018
Jantemp:Winprecip	Interaction term	58.3	< 0.00001
	Model $C_p = 870$		
YUBR (K_GLM)			
Sumprecip	2nd order polynomial	604	< 0.00001
Slope	Linear	74	< 0.00001
Jultemp	2nd order polynomial	89.9	< 0.00001
Jantemp	2nd order polynomial	56.3	< 0.00001
Elevation	2nd order polynomial	34.6	< 0.00001
Yubr2	Linear	430.1	< 0.00001
	Model $C_p = 501$		

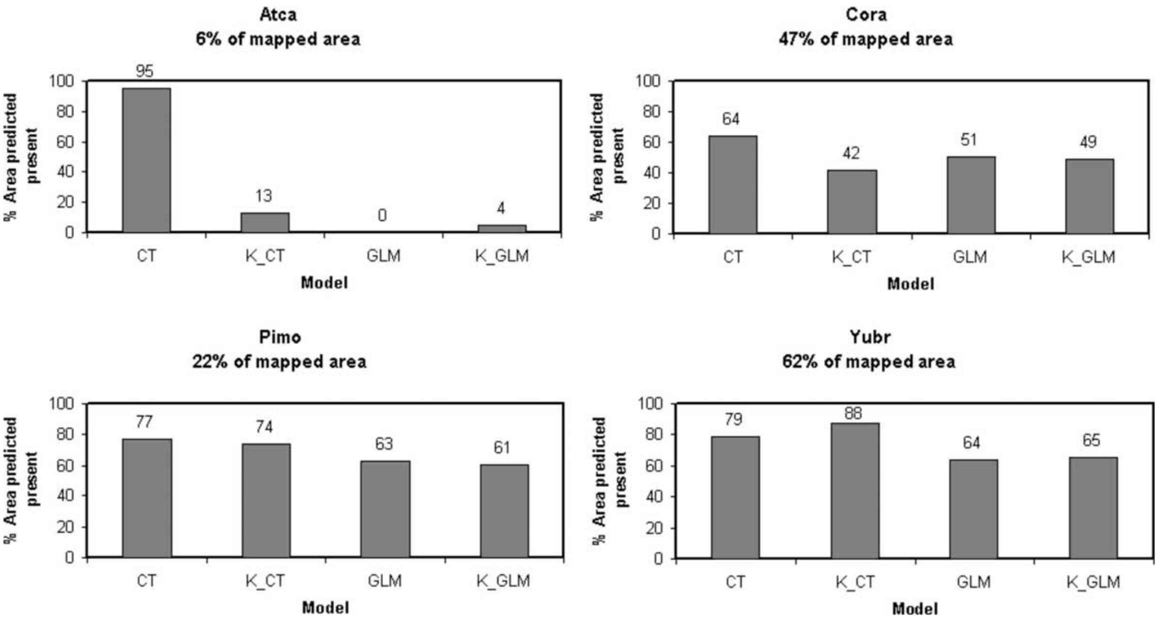

Fig. 3. Percentage of total area of each alliance predicted present by all four models combined that was predicted present by each of the models individually. Numbers under alliance names show the proportion of grid cells predicted present by any of the four models in the mapped subsection.

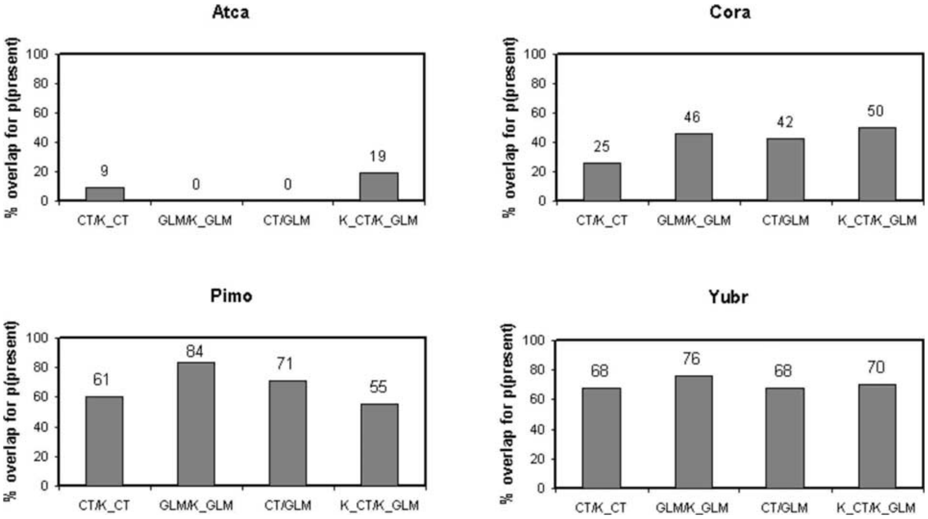

Fig. 4. Percentage of spatial overlap between model predictions. "CT/K_CT" refers to a comparison between classification tree model and classification model with kriged dependence term; "GLM/K_GLM" between GLM and GLM with kriged dependence term; "CT/GLM" between classification tree and GLM; and "K_CT/K_GLM" between classification tree with kriged dependence term and GLM with kriged dependence term.

consistency in the overlap between the CT and K_CT models (61%). According to the AUC measure (see Table 5), the CT and K_CT models would differ in discriminating between presence and absence in 2% of the cases, yet the two models showed only 25% spatial overlap in their predictions. Although, the change in model performance after the kriged variable was added was more drastic with the GLM and K_GLM models (see Tables 3 and 5), the spatial agreement in the predictions generated by these two models was still consistently better than that of CT and K_CT. Fig. 5 shows the subsection of the study area used to generate predictions for the CORA alliance with the kriged dependence term that ranges in value from 0 (indicating that no observations of CORA were nearby) to 6.7 (indicating probable clustering of CORA). The predicted probability maps resulting from the four models for CORA (before using the probability threshold to make binary maps) are shown in Fig. 6. The probability map generated by the classification tree with the kriged dependence term (Fig. 6B) is notably smooth and ecologically unrealistic, due to the use of the kriged dependence variable in the first three splits, although there seem to be few commission errors compared to the three other maps. Both GLMs (with and without the kriged dependence term, Fig. 6C,D) look similar, with a potentially high number of commission errors, but the map produced by the model with the kriged dependence term should have fewer omission errors and higher sensitivity (as was shown for the test data, Table 4).

In summary, the CT model predicted CORA to occur between 1046 and 1920 m, on gentle slopes receiving low solar radiation, and on moderate slopes between midslopes. The GLM also predicted CORA to occur at low July and January temperatures on moderate slopes. Both GLMs and CTs predicted PIMO to occur at elevations greater than 1924 m, except where the summer precipitation is below 51.5 mm. If the summer precipitation is greater than 51.5 mm but less than 63.5 mm and landcomp is either CalcCarb or IgnVolc, then it is predicted to be absent. YUBR was predicted by CT to occur generally where summer precipitation is between 38 and 68 mm, but will occur below 38 mm when July temperature is less than 30.6 °C and landcomp is CalcCarb, IgnPlut or IgnVolc. The variable representing the interaction between January temperature and winter precipitation was

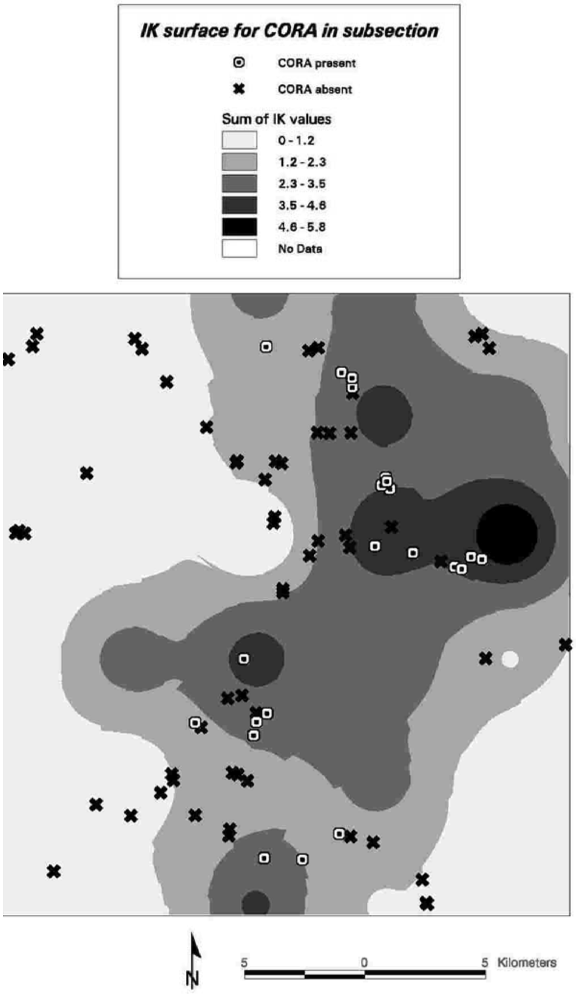

Fig. 5. Kriged dependence term for CORA alliance in mapped subsection. Each pixel contains the sum of the kriged probability of presence for its eight neighbors.

highly significant in GLM and the two variables were also often nested in the CT model decision rules (e.g. where winter precipitation is less than 179 mm, YUBR is predicted to be absent unless January temperature exceeds −3.1 °C). As discussed earlier, landform was the most important variable in the ATCA CT model. The initial split resulted in predictions of ATCA to be absent on any landform other than alluvial fan, older alluvial deposits, fluvial terrace, and active alluvial plain.

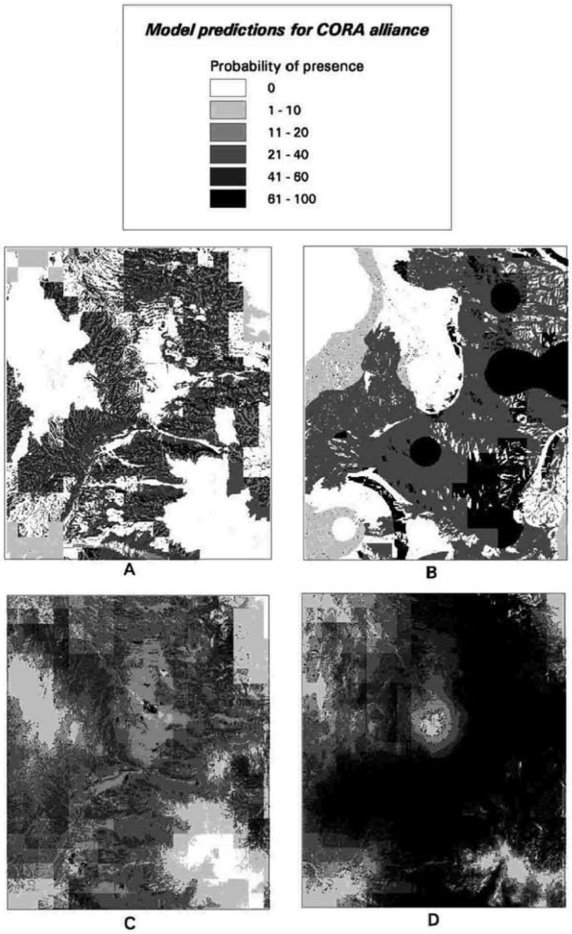

Fig. 6. Predictions generated for test area with (A) CORA classification tree ($P = 0.1$); (B) CORA classification tree with kriged dependence term ($P = 0.2$); (C) CORA GLM ($P = 0.2$); (D) Cora GLM with kriged dependence term ($P = 0.2$). Optimum probability thresholds are given in parentheses.

5. Conclusions

The CTs suggested several pair-wise interaction terms that were also significant in the GLMs. Although a measure to compare the performance of both models in terms of deviance explained could not be fully implemented, comparisons could be made between the models based on classification accuracy using threshold-dependent and threshold-independent measures. These measures indicated that the CT models had higher classification accuracy on the training data than

GLMs, but accuracy of the CT models degraded more drastically when they were assessed using the test data.

In general, the kriged dependence term improved the model for all alliances, with PIMO showing the least dramatic improvement because its distribution is more easily delimited by environmental variables (elevation, summer precipitation, slope). The PIMO alliance consistently had better performing models by all measures, but this was expected due to its clear correlation with high elevation and high precipitation/low temperature. The ATCA alliance had the poorest performing models and this could be attributed to the fact that it was the most rare of the four alliances, and further its best-known ecological correlation is with a landform type, which was not easily specified in GLMs.

In terms of model specification, there is a great deal of subjectivity and expertise (Austin and Meyers, 1996) involved in variable selection for GLMs. With 12 explanatory variables, it would have been very time-consuming to test all possible interactions and response shapes, so the models that resulted could probably be improved upon. Although the choice between two variables with very similar effects in CT models is also somewhat arbitrary and the variables can be difficult to interpret in terms of response functions (Austin et al., 1994), the resulting trees are easy to test for ecological realism. However, CT models are adversely affected by outliers, which can cause very different tree results when they are included. Also, CT models partition the data based on one predictor variable at a time, therefore the resulting predicted maps are more likely to adhere to existing spatial patterns in the input data. Possibly due to the arbitrariness with which both models select equally good predictor variables, even CT and GLM models with very similar classification accuracy can generate very different spatial predictions.

While the addition of the kriged dependence term always improved the model performance by all measures, it is, in effect, relying too heavily on the sample data used to construct the models and will therefore produce less generalizable models for prediction. In all models with the exception of the PIMO alliance, the kriged dependence term was very important and may have been replacing more suitable predictive environmental variables. When spatial dependence does exist but is not included explicitly in the model, the importance of some predictor variables may be overstated, as it is their spatial autocorrelation that is being seized upon as being important to the model. However, a variable representing true spatial dependence requires complete information on alliance presence as well as absence, and, lacking this, methods to interpolate it from sample data (as with kriging) tend to overestimate the true condition. Additionally, the predicted maps generated by the models with the kriged dependence term appeared less ecologically realistic in some cases than the models with the environmental variables. The kriged dependence variable maps show unrealistic smooth circles around alliances—an obvious oversimplification of the true spatial pattern even when extreme spatial dependence occurs. The spatial dependence that the kriged dependence term was intended to estimate can only be as accurate or complete as the sample data on which the models are built and rare alliances could result in distorted kriged dependence terms. Future work will focus on using spatial dependence in a more restricted way in vegetation models so that environmental variables are not replaced by an over-specified dependence variable (see Gotway and Stroup, 1997; Pebesma et al., 2000; Bishop and McBratney, 2001).

Acknowledgements

This research was supported in part by United States Geological Survey (USGS) and Department of Defense (DoD) grants (to J. Franklin) and the Department of Geography, San Diego State University (support for J. Miller). We thank the following people for help and advice including comments on the manuscript: K. A. Thomas, T. Keeler-Wolf, P. Stine, D. Shaari, J. Michaelsen, S.J. Rey, M. F. Goodchild, J. Rogan, A. Hope, and D. Apalatea. We thank A. Guisan and A. Lehman for providing the opportunity to participate in the Advances in GLM/GAM workshop and the motivation to write this paper, and T. Edwards and the anonymous reviewers for improving the manuscript with their comments.

References

Anselin, L., 1993. Discrete space autoregressive models. In: Goodchild, M., Parks, B., Steyaert, L. (Eds.), Environmental Modeling with GIS. Oxford University Press, Oxford, pp. 454–469.

Atkinson, A., 1981. Likelihood ratios, posterior odds and information criteria. J. Econometrics 16, 15–20.

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

Augustin, N., Mugglestone, M., Buckland, S., 1998. The role of simulation in modelling spatially correlated data. Environmetrics 9, 175–196.

Austin, M., Heyligers, P., 1989. Vegetation survey design for conservation: gradsect sampling of forests in north-eastern New South Wales. Biol. Conservation 50, 13–32.

Austin, M., Meyers, J., 1996. Current approaches to modelling the environmental niche of eucalypts: implication for management of forest biodiversity. Forest Ecol. Manage. 85, 95–106.

Austin, M., Smith, T., 1989. A new model for the continuum concept. Vegetation 83, 35–47.

Austin, M., Meyers, J., Belbin, L., Doherty, M., 1994. Modelling of landscape patterns and processes using biological data, Sub Project 5: Simulated Data Case Study, CSIRO Division of Wildlife and Ecology, Canberra (Reprint No. 2703).

Beatley, J., 1975. Climates and vegetation pattern across the Mojave/Great Basin Desert transition of southern Nevada. Am. Midland Nat. 931, 53–70.

Besag, J., 1972. Nearest-neighbour systems and the autologistic model for binary data. J. Roy. Stat. Soc. B 34, 75–83.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. Roy. Stat. Soc. B 36, 192–236.

Beven, K., Kirkby, M., 1979. A physically based variable contributing area model of basin hydrology. Hydrol. Sci. Bull. 24, 43–69.

Bio, A., Alkemade, R., Barendregt, A., 1998. Determining alternative models for vegetation response analysis: a nonparametric approach. J. Vegetation Sci. 9, 5–16.

Bishop, T., McBratney, A., 2001. A comparison of prediction methods for the creation of field-extent soil property maps. Geoderma 103, 149–160.

Borcard, D., Legendre, P., Drapeau, P., 1992. Partialling out the spatial component of ecological variation. Ecology 73, 1045–1055.

Breiman, L., Freedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth, Belmont, CA.

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

Burrough, P., McDonnell, R., 1998. Principles of Geographical Information Systems. Oxford University Press, New York.

Carver, S., 1991. Integrating multi-criteria evaluation with geographic information systems. Int. J. Geographical Inf. Syst. 5, 321–339.

Davis, F., Goetz, S., 1990. Modeling vegetation pattern using digital terrain data. Landscape Ecol. 41, 69–80.

De'ath, G., Fabricius, K., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81, 3178–3192.

DeLeo, J., 1993. Receiver operating characteristic laboratory ROCLAB: software for developing decision strategies that account for uncertainty, Proceedings of the 2nd International Symposium on Uncertainty Modelling and Analysis, IEEE Computer Society Press, College Park, MD, pp. 318–325.

Dokka, R.K., Christenson, C., and Watts, J., 1999. Geomorphic Landform and Surface Composition GIS of the Mojave Desert Ecosystem in California. Available from: www.mojavedata.gov/mdep/geomorphic/glmetadata.html

Dubayah, R., 1994. Modeling a solar radiation topoclimatology for the Rio Grande River Basin. J. Vegetation Sci. 5, 627–640.

Fels, J., 1994. Modeling and mapping potential vegetation using digital terrain data, Ph.D. Dissertation, North Carolina State University, pp. 287.

Fielding, A., Bell, J., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conservation 241, 38–49.

Fischer, H., 1990. Simulating the distribution of plant communities in an alpine landscape. Coenoses 51, 37–43.

Franklin, J., 1995. Predictive vegetation mapping: geographic modeling of biospatial patterns in relation to environmental gradients. Prog. Phys. Geography 19, 474–499.

Franklin, J., 1998. Predicting the distributions of shrub species in California chaparral and coastal sage communities from climate and terrain-derived variables. J. Vegetation Sci. 95, 733–748.

Franklin, J., McCullough, P., Gray, C., 2000. Terrain variables used for predictive mapping of vegetation communities in southern California. In: Wilson, J., Gallant, J. (Eds.), Terrain Analysis: Principles and Applications. Wiley & Sons, New York, pp. 331–353.

Franklin, J., Keeler-Wolf, T., Thomas, K., Shaari, D., Stine, P., Michaelsen, J., Miller, J., 2001. Stratified sampling for field survey of environmental gradients to define vegetation alliances in the Mojave Desert. In: Millington, A., Walsh, S., Osborne, P. (Eds.), GIS and Remote Sensing Applications in Biogeography and Remote Sensing. Kluwer Academic Publishers, Netherlands, pp. 229–251.

Franklin, J., 2002. Enhancing a regional vegetation map with predictive models of dominant plant species in chaparral. Appl. Vegetation Sci. 5, 135–146.

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

Gotway, C., Stroup, W., 1997. A generalized linear model approach to spatial data analysis and prediction. J. Agric. Biol. Environ. Stat. 2, 151–178.

Grossman, D., Faber-Langendeon, D., Weakley, A., Anderson, M., Bourgeron, P., Crawford, R., Goodin, K., Landaal, S., Metzler, K., Patterson, K., Pyne, M., Reid, M., Sneddon, L., 1998. International Classification of Ecological Communities: Terrestrial Vegetation of the United States The National Vegetation Classification System, Development, Status and Applications, vol. I. The Nature Conservancy, Washington, DC.

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

Guisan, A., Zimmermann, N., 2000. On the use of static distribution models in ecology. Ecol. Model. 135, 147–186.

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

Guisan, A., Weiss, S., Weiss, A., 1999. GLM versus CCA spatial modeling of plant species distributions. Plant Ecol. 1431, 107–122.

Gumpertz, M., Graham, J., Ristaino, J., 1997. Autologistic model of spatial pattern of phytophthora epidemic in bell pepper: effects of soil variation on disease presence. J. Agric. Biol. Environ. Stat. 2, 131–156.

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

Hoeting, J., Leecaster, M., Bowden, D., 2000. An improved model for spatially correlated binary responses. J. Agric. Biol. Environ. Stat. 51, 102–114.

Hosmer, D., Lemeshow, S., 1989. Applied Logistic Regression. Wiley & Sons, New York.

Hunt, C., 1966. Plant Ecology of Death Valley, U.S. Geological Survey Professional Paper 509, USGS, Washington, DC, pp. 1–68.

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

Le Duc, M., Hill, M., Sparks, T., 1992. A method for predicting the probability of species occurrence using data from systematic surveys. Watsonia 19, 97–105.

Lees, B., Ritman, K., 1991. Decision-tree and rule-induction approach to integration of remotely sensed and GIS data in mapping vegetation in disturbed or hilly environments. Environ. Manage. 156, 823–831.

Legendre, P., 1993. Spatial autocorrelation: problem or new paradigm? Ecology 74, 1659–1673.

Legendre, P., Fortin, M.J., 1989. Spatial pattern and ecological analysis. Vegetation 80, 107–138.

Lenihan, J., Neilson, R., 1993. A rule-based vegetation formation model for Canada. J. Biogeography 20, 615–628.

McAuliffe, J., 1994. Landscape evolution, soil formation, and ecological processes in Sonoran Desert bajadas. Ecol. Monogr. 642, 111–148.

McCullagh, P., Nelder, J., 1989. Generalized Linear Models. Chapman & Hall, New York.

Meentemeyer, R., Moody, A., Franklin, J., 2001. Landscape-scale patterns of shrub-species abundance in California chaparral: the role of topographically mediated resource gradients. Plant Ecol. 156, 19–41.

Michaelsen, J., Schimel, D., Friedl, M., Davis, F., Dubayah, R., 1994. Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys. J. Vegetation Sci. 5, 673–686.

Moore, D., Lees, B., Davey, S., 1991. A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. Environ. Manage. 151, 59–71.

Moore, I., Grayson, R., Ladson, A., 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrol. Process. 5, 3–30.

Nicholls, A., 1989. How to make biological surveys go further with generalised linear models. Biol. Conservation 50, 51–75.

Norris, R., Webb, R., 1990. Geology of California. Wiley & Sons, New York.

Parker, K., 1991. Topography, substrate, and vegetation patterns in the northern Sonoran Desert. J. Biogeography 18, 151–163.

Pebesma, E., Duin, R., Bio, A., 2000. Spatial interpolation of sea bird densities on the Dutch part of the North Sea, ICG report Department of Physical Geography, Utrecht University, The Netherlands, pp. 130.

Rowlands, P., Johnson, H., Ritter, E., Endo, A., 1982. The Mojave Desert. In: Bender, G. (Ed.), Reference Handbook on the Deserts of North America. Greenwood Press, Westport, CT, pp. 103–162.

Schoenherr, A., 1992. A Natural History of California. University of California Press, Berkeley, CA.

Smith, P., 1994. Autocorrelation in logistic regression modeling of species' distributions. Global Ecol. Biogeography Lett. 4, 47–61.

Sokal, R., Oden, N., 1978. Spatial autocorrelation in biology 1. Methodology. Biol. J. Linnean Soc. 10, 199–228.

Thomson, J., Weiblen, G., Thomson, B., Alfaro, S., Legendre, P., 1996. Untangling multiple factors in spatial distributions: lilies, gophers and rocks. Ecology 776, 1698–1715.

Tobler, W., 1979. Cellular geography. In: Gale, S., Olsson, G. (Eds.), Philosophy in Geography. Reidel, Dordrecht, pp. 379–386.

Valverde, P., Zavala-Hurtado, A., Montana, C., Ezcurra, E., 1996. Numerical analysis of vegetation based on environmental relationships in the Southern Chihuahuan Desert. Southwestern Nat. 414, 424–433.

Vayssières, M., Plant, R., Allen-Diaz, B., 2000. Classification trees: An alternative nonparametric approach for predicting species distributions. J. Vegetation Sci. 11, 679–694.

Wu, H., Huffer, F., 1997. Modelling the distribution of plant species using the autologistic regression model. Environ. Ecol. Stat. 4, 49–64.

Yeaton, R.I., Yeaton, R.W., Waggoner, J., Horenstein, J., 1985. The ecology of Yucca (Agavaceae) over an environmental gradient in the Mohave Desert: distribution and interspecific interactions. J. Arid Environ. 8, 33–44.

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

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