A Comparison of Methods for Incorporating Spatial Dependence in Predictive Vegetation Models: A Mojave Desert Case Study

Jennifer Miller

[1]Department of Geology and Geography, West Virginia University,
Morgantown, WV 26506-6300
Tel: +1 304-293-5603 x4341
FAX: +1 304-293-6522
Email: jennifer.miller@mail.wvu.edu

Abstract

In this study, presence/absence models of eleven vegetation alliances in a portion of the Mojave Desert (California, USA) were developed using generalized linear models (GLM) and classification tree (CT) models, and two different methods for explicitly incorporating spatial dependence. In the first method, spatial dependence was included as a model term, along with environmental variables, and predictions were generated. In the second method, the residuals from the models using environmental variables were added to the model predictions. Accuracy was assessed with relative-operator characteristic (ROC) plots, using a portion of the sample data not used for model development. Spatial predictions of alliance presence/absence were compared among all twelve of the models.

In general, incorporating spatial dependence improved the classification accuracy of most of the models. Models for more common alliances showed a greater increase in accuracy, while models for rare alliances were less consistent. This was most likely related to the relative of data on 'presence' observations for rare alliances. The residual interpolation method had more consistently positive results in terms of increased accuracy than the method of including spatial dependence as a predictor variable. One problem associated with the latter method is that the resulting models may become less generalizable, as they rely too heavily on presence data rather than environmental correlates. Additionally, the spatial patterning of the dependence variables, particularly the overly smooth kriged surface, can result in ecologically unrealistic predictions.

1. Introduction

Predictive vegetation modeling (PVM) quantifies the relationship between vegetation distribution and environmental gradients and applies the resulting model to unsampled areas. The result is a map that shows the geographic distribution of some vegetation attribute (e.g., probability of presence or mean abundance) as a function of digital maps of the environmental variables. There is an increasingly wide variety of statistical methods from which to choose, ranging from more traditional generalized regression to artificial neural networks and genetic algorithms (for review, see Franklin, 1995; Guisan and Zimmermann 2000). Method selection is based upon, among other things, data characteristics (known vs. unknown parameters, distribution, measurement level), model use (prediction vs. inference), and intended final product (categorical map, abundance map).

These models are typically static and probabilistic in nature, and rely on the assumption that

vegetation is in equilibrium with its environment. Many frequently used methods such as generalized regression further assume that the distribution of vegetation is random, and therefore each observation is independent, an assumption which casual observation proves unrealistic. This lack of independence and randomness in natural distributions led Waldo Tobler (1970) to coin the phrase 'first law of geography,' to describe the regularity that near things are more related than distant things (also see Sui 2004). Spatial dependence in biogeography, for example, results from the fact that plants that are close together are more likely to be influenced by the same generating process and will therefore be similar (Legendre and Fortin 1989).

Failing to account for 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 structure can be explained by the predictor variables used in the model. Environmental variables such as precipitation, temperature and elevation exhibit spatial dependence, some of which is responsible for spatial clustering in vegetation distribution, while extant spatial dependence can result from either unmeasured environmental variables or biotic processes that cause spatial clustering.

Spatial dependence has been identified as an important area of future research in habitat distribution models in general (Franklin 1995; Guisan and Zimmermann 2000). Although traditionally ignored, many vegetation modeling studies that do acknowledge spatial dependence attempt to eliminate it by manipulating the sampling strategy to avoid autocorrelated observations (Legendre and Fortin 1989; Davis and Goetz 1990; Borcard et al. 1992; Smith 1994). Several studies have indicated the importance of including spatial dependence in models as a way of clarifying the influence of environmental predictor variables (Wu and Huffer 1997; Hubbell et al. 2001; Keitt et al. 2002). Borcard et al. (1992) and Legendre and Legendre (1998) used partial regression to separate the explanatory ability (of vegetation distribution) of environment from spatial factors (see also Lobo et al. 2002; Lobo et al. 2004; Graae et al. 2004; Nogués-Bravo and Martinez-Rica 2004 for recent examples). However, the potential *predictive* ability of spatial dependence in PVM has only recently been explored (see Miller et al. submitted, for review).

The aim of this study was to compare the predictive accuracy of vegetation models in which spatial dependence has been explicitly incorporated. Two types of models are utilized, generalized regression models (GLM) and classification trees (CT), and two methods for describing spatial dependence are compared. The first method of incorporating spatial dependence uses geostatistical techniques to interpolate a surface from sample data to obtain an additional variable of neighborhood (3 x 3 grid cells) presence/absence. The second method of incorporating spatial dependence considers the spatial dependence in the model residuals to most likely represent spatial dependence in the vegetation distribution unexplained by the model variables. Residuals from nonspatial GLM and CT models are interpolated and added to the model results to form new predictions.

The models are used to predict presence/absence of eleven vegetation alliances, ranging from rare to common, in a section of the Mojave Desert, CA. Twelve environmental variables were used as predictors in the models. These variables have been used successfully in previous studies, or have an ecological basis for being associated with vegetation distribution. A total of 3819 observations of alliance presence/absence were collected and compiled and divided into 75:25 train:test modeling dataset. Classification accuracy is compared for models using two different methods of incorporating spatial dependence (model term and model residuals), two different statistical methods (GLM and CT), and several different types of vegetation distributions (rare to common, generalist to specialist).

2. Data

2.1 Study area

The study area for this research is a portion of the Mojave Desert Ecoregion within California, referred to as the Eastern California Subsection (figure 1). This region is characterized by basin and range physiography, much of it at elevations between 600 and 1200 m, with some mountain ranges exceeding 1600 m. The Mojave Desert climate is characterized by low, unevenly distributed precipitation, temperature extremes, windy conditions and high light intensity (Schoenherr 1992). It serves as a transition zone between the Great Basin Desert to the north and the Sonoran Desert to the south, and contains, in addition to both Great Basin and Sonoran vegetation, its own endemic species as well (Rowlands et al. 1982). 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). 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, although most of it occurs between October and April. One result of the combination of low precipitation and high evaporation rate is the presence of alkaline soils with low moisture retention capabilities. 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.mojave data.gov).

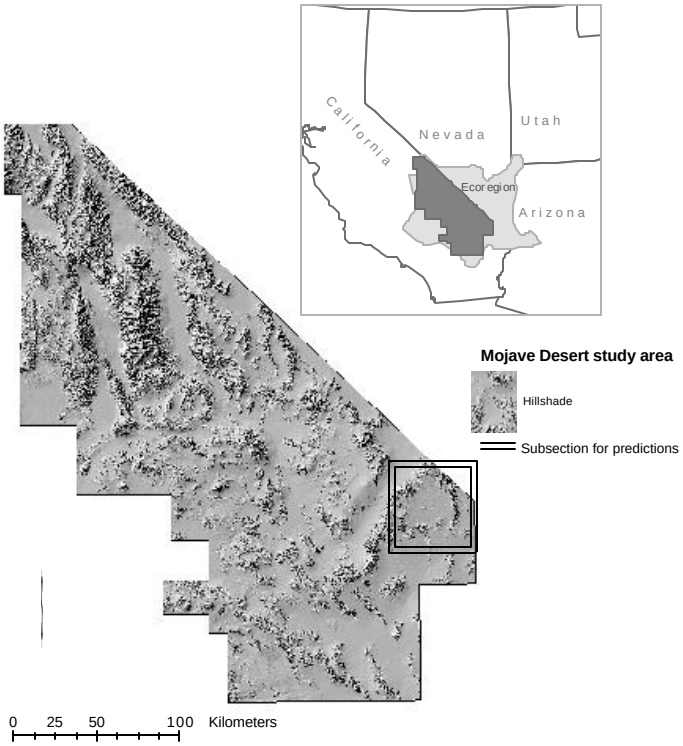

Figure 1: Mojave Desert study area. Outlines section is used for predictions in figure 4.

2.2 Environmental variables

The explanatory variables used here included climate, topography, and landform (table 1).

The relationship between climate and vegetation distribution is based largely on the plants' physiological tolerances and has been used historically to map vegetation (see Austin et al. 1994; Franklin 1995 for review). The climate variables used here 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 one square kilometer 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).

Table 1: Environmental variables used in this study. Climate variables are 1-km resolution; all others are 30-m resolution.

Variable name	Description	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	From USGS 7.5' DEM	-85 – 3390 m
Slope	Derived from DEM	0 – 78
Swness	Cosine (aspect - 225°) (Franklin et al., 2000)	-1 – 1
Lpos4	Landscape position; Average difference between cell and 4 neighbors (positive in valleys, neutral in mid-slope position, negative on ridges) (Fels 1994)	-1732 – 2311
Solrad	Potential 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 M.Kirkby 1979)	0 – 22.6
Landform	Geomorphic landform (Dokka 1999)	29 nominal classes
Landcomp	Surface composition (Dokka 1999)	6 aggregated nominal classes

Topographic 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 (Wilson and Gallant 1998; see Florinsky 1998 for review of relationships between topographic variables and landscape characteristics). A United States Geological Survey

(USGS) 7.5' digital elevation model (DEM) was used to provide elevation values (30m resolution); 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 explaining 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)), a modification of an original formula proposed by Beers et al (1966). Higher values indicate more xeric exposures, and pole-facing (moist), neutral, and equator-facing (dry) slope aspects can be more easily distinguished.

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).

Elevation, slope, and aspect were subsequently used to derive three more complex topographic variables: landscape position, potential solar radiation, and topographic moisture index. Landscape position describes the position of a cell relative to surrounding cells (upslope or downslope). Potential solar radiation is related to the water availability of a site (I. Moore et. al. 1991) and topographic moisture is related to soil depth, texture and potential soil moisture (reviewed in Franklin 1995). Landscape 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.

Vegetation in the desert has a particularly close relationship to landform, as it relates to both nutrient and moisture availability. 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 for more detail). Landcomp aggregates land surface composition into six classes and landform has 29 landform classes.

Many of the environmental variables used can have values that are correlated (e.g., climate and elevation). The problem of multicollinearity can affect coefficient and significance estimates in GLMs. The problem is less severe in CTs—when predictor variables have similar effects, one is chosen arbitrarily. However, these variables were selected for this study based on their ecological significance, and the emphasis was on prediction rather than inference.

2.2 Vegetation variables

The vegetation response variable predicted was at the alliance level of the National Vegetation Classification System (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, p. 23). Eleven vegetation alliances (table 2) were selected for modeling here, with a goal of achieving a representative variety of distribution types in the Mojave Desert (e.g., rare, common; specific and general environmental relationships).

A dataset with 3819 observations of presence/absence for all eleven alliances was compiled from three different sampling strategies described briefly here (see Miller 2005 for more details). About 30% of the dataset was collected in 1998-1999 from a gradient-directed sample; 10% of the dataset came from five 'retrospective' datasets collected between 1970-1990; and 60% of the dataset resulted from a modified roadside sample. The primary goal of this study was to develop predictive models so, despite inconsistency in the three sampling

strategies, and lack of unbiased sampling in the latter two, all of the data were combined. Therefore, limitations in the use of these models for explanation rather than, or in addition to, prediction, should be noted. The data were partitioned into 75:25 train:test portions following a heuristic suggested by Fielding and Bell (1997) for presence/absence data with more than ten predictor variables.

Table 2: Vegetation alliances modeled(source: Thomas et al., 2004), and the proportion of the full (test and train, n = 3819) dataset in which they are present. Species abbreviations comprise the first two letters of genus and specific epithet of indicator species. Number of observations of present (P) are given for test and train data (which consisted of 960 and 2859 observations respectively).

Label, (Proportion)	**Alliance name**	**P test**	**P train**	**Habitat**
ATCA (0.006)	*Atriplex canascens* Shrubland Alliance	7	16	Soil of old beach, lake deposits; dissected alluvial fans, rolling hills. Wetland habitats such as washes, playa lakebeds and shores.
ATCO (0.028)	*Atriplex confertifolia* Shrubland Alliance	34	73	Bajadas, flats, edges of playas, lower slopes, rocky hills, valleys, and minor rills and washes. Soils variable; Wetland habitats such as ashes, and playa lakebeds.
CORA (0.034)	*Coleogyne ramosissima* Shrubland Alliance	21	110	Widespread; shallow rocky soils on upper bajadas, pediments and hill slopes, above 1000 m.
EPNE (0.006)	*Ephedra nevadensis* Shrubland Alliance	5	17	Dry, open slopes; ridges; breaks with southern exposures; canyons; floodplains, arroyos; and washes. Well-drained soils, with gravel or rock, may be alkaline or saline.
GALL (0.011)	*Pleuraphis rigida* Herbaceous Alliance	9	34	Flat ridges, lower bajadas, slopes, dune aprons and stabilized dunes.
LATR (0.158)	*Larrea tridentata* Shrubland Alliance	145	460	Alluvial fans; bajadas; upland slopes; minor intermittent wash channels; Soils well drained, also desert pavement surface.
LATR-AMDU (0.427)	*Larrea tridentata - Ambrosia dumosa* Shrubland Alliance	417	1214	Alluvial fans; bajadas; upland slopes; minor washes and rills. Soils well-drained, colluvial, sandy, and/or alluvial, often underlain by a caliche hardpan; may be calcareous and/or have pavement surface.
MESP (0.007)	*Menodora spinescens* Dwarf-shrubland Alliance	10	17	Ridges, slopes, upper alluvial fans and bajadas. Soils bedrock or alluvium derived.
PIMO (0.013)	*Pinus monophylla* Woodland Alliance	12	38	Upper elevations; cool, moist mountain areas
YUBR (0.092)	*Yucca brevifolia* Wooded Shrubland Alliance	87	265	Narrow zone, base of mountains. Gentle alluvial fans; ridges, gentle to moderate slopes. Soils colluvial, alluvial derived: coarse sand, very fine silt, gravel, or sandy loam. Often bimodal soils with both coarse sands and fine silts.
YUSC (0.047)	*Yucca schidigera* Shrubland Alliance	49	132	Rocky slopes, upper bajadas, and alluvial fans. Soils well drained, derived from various substrates including granitic, limestone, volcanic, metamorphic.

The data used here were collected under the auspices of the Mojave Vegetation Mapping Project (MVMP), the purpose of which was to use PVM to develop accurate vegetation maps efficiently. The methods for incorporating spatial dependence explored here were focused primarily on increasing prediction accuracy, using the available data and methods that were not prohibitively computationally intensive or complex, rather than on generating inferences about the nature of spatial dependence observed in this data. It should be mentioned that, while accuracy assessment is an important component of PVM research, when the emphasis is on prediction, all available data should be used to develop the ultimate models (Fielding and Bell, 1997). Accuracy may still be assessed based on the test data, as that would represent a more pessimistic and realistic measure of model success, then training and test data should be combined for ultimate model development.

3. Methods

3.1 Statistical models

Two conceptually different modeling methods were used here: generalized linear models (GLMs) and classification tree (CT) models. 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 for CT models, of vegetation presence.

3.1.1 Generalized linear models

GLMs are one of the most commonly used methods to relate vegetation to its environment (see Guisan et al. 2002). GLMs extend more traditional linear regression models to allow for non-normal error distributions and accommodate many different response variable distributions (e.g., Logistic, Poisson: McCullagh and Nelder 1989). Model specification with GLMs is fairly subjective, and as a result they are less data-driven and exploratory as nonparametric models such as CTs . One limitation associated with GLMs is the stepwise variable selection procedure often utilized. Although this process can be automated (resulting in new problems related to variable inclusion based solely on significance estimate), the procedure used here is an iterative and subjective process, requiring expert information on variable inclusion and exclusion, as well as appropriate transformations.
Logistic regression uses a logit link to describe the relationship between the response and the linear sum of the predictor variables. was used here, and was used here, as the response data were binary. The GLMs were developed based on a combination of stepwise and subjective, iterative, variable addition and subtraction methods with a goal of minimizing the AIC statistic (Akaike 1973; Hastie et al. 2001). Pairwise interaction terms based on biophysical principles (e.g., elevation/aspect) or position in the CT structure (described below) were also tested for significance. Generalized additive models were used to identify higher order relationships (polynomial, piecewise linear) between the environment gradients and response variable (see Brown 1994; Franklin 1998; Miller and Franklin 2002 for similar methods) which were then specified as such in the GLM. Once a subset of variables was selected for the nonspatial model, the spatial dependence term was added (always as a linear term) and any subsequently non-significant variables were removed.

3.1.2 Classification trees

CT models are rule-based and nonparametric. The rules are developed by partitioning data into subsets that are increasingly homogeneous with respect to the response variable (Breiman et al. 1984). All splits in the predictor variables are examined and a split is selected to maximize homogeneity in the resulting two branches. The splitting continues until either the resulting branches are homogeneous or a minimum number of observations remains in the

subset. The terminal node is the end of the branch and is defined by the hierarchical rules that precede it. Associated with each terminal node is the number of points in the training data that were observed in locations that met the environmental criteria, as well as the number that are correctly classified. This proportion can be interpreted as the 'suitability' (Pontius and Schneider 2001) for a class to occur, analogous to the probability that results from GLMs. CTs are particularly appropriate when the form of the relationship between response and predictor variables is unknown, as they are considered to be 'data-driven', and therefore more exploratory than parametric methods such as GLMs. Classification trees produce a set of decision rules that identify not only multiple conditions that are associated with alliance presence, but also conditions that are associated with its absence. Therefore, CT models may describe more adequately common or generalist alliances that are associated with more than one set of environmental conditions, especially when "indirect gradients" are used. Classification trees are particularly well suited for environmental modeling as they are able to express complex relationships among the predictor variables that are nonlinear, 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), classification trees identify specific thresholds of environmental conditions above or below which a species or vegetation type can be found, and these can be used to formulate conditional statements from which predicted surfaces can be generated.

Each CT model was given all predictor variables (twelve environmental variables for nonspatial models, one additional spatial variable for each of the three spatial models), then was pruned (based on cross-validation, see Breiman at al. 1984) to sizes that ranged from 6 to 31 terminal nodes for the non-spatial CT models, and 3 to 27 nodes for the spatial CT models.

3.2 Spatial dependence methods

3.2.1 Spatial dependence variable as a model term

Two geostatistical interpolation methods (kriging and simulation) were used to calculate spatial dependence terms based on the distribution of presence/absence in the training data, and these terms were included with other environmental variables in GLM and CT models in the methods described below. Both kriging and simulation are based on the spatial structure of the sample data, divided into three components: deterministic variation, spatial autocorrelation (defined by a variogram), and noise (Burrough and McDonnell 1998). Kriging methods result in one set of predicted values that are optimized based on the variogram and the spatial configuration of the data, but the result is overly smooth. Rather than one optimal prediction, simulation generates a series of equally probable predictions, maintaining some of the 'roughness' of the data (Burrough and McDonnell 1998). Geostatistical models explicitly include spatial dependence, where its importance is such that a predicted value is mainly determined by its distance to other values. But additional predictor variables are less easily and flexibly included in geostatistical models, and spatial dependence is usually not the only factor related to vegetation distribution. A spatial dependence term has been explicitly incorporated along with environmental variables in logistic models, formally called autologistic models, following work by Besag (1972) and subsequent modifications by Augustin et al. (1996). However, including a spatial dependence term requires complete information on the distribution of the response variable, which is rarely available. Gibbs sampler and Markov chains methods (Augustin et al., 1996; 1998) and Markov chain Monte Carlo methods (Gumpertz et al. 1997; Wu and Huffer 1997) have been used to "fill in the blanks" of the sample data, but these are all very computationally intensive. Here we use geostatistical interpolation methods to calculate

spatial dependence terms used in the models.

Indicator kriging is the non-linear form of kriging used with binary response data and its product is a surface with the probability that the condition coded '1' will occur (Burrough and MacDonnell 1998). When presence/absence data are used, it is the probability of presence that is predicted. Similarly, indicator simulation is used with binary sample data (Burrough and MacDonnell 1998). The result is a layer with values of '1' and '0' based on the variogram and the proportion of 1 and 0 in the sample data.

An indicator variogram was fit to the training data for each alliance. All variograms were fit using the common heuristic "by eye" approach (Gotway and Hartford 1996), and are therefore highly subjective. Three of the most commonly used variogram models were tested: spherical, exponential, and Gaussian. Only spherical, which describes a clear range and sill, and exponential, which describes a more gradual approach to the range (Burrough and MacDonnell 1998) were used. For comparison purposes, a variogram was fit to all alliances, even when spatial dependence was not apparent (the sill-to-noise ratio was low). Indicator kriging of the variograms and sample data was used to calculate a layer of probability values for each alliance. Similarly, indicator simulation was used to calculate a layer with values of 0 and 1 that mirrored the sample data proportions. The resolution of these spatial dependence layers (500 m) was chosen to approximate reasonable spatial dependence resolution, given the resolution of the training data and environmental variables, without resulting in unduly large processing time.

This resulted in eleven maps (one for each alliance) with values that represented the probability that a specific alliance would be present in each 500 m grid cell based on indicator kriging (referred to as "K"); eleven maps with values of 1 and 0 indicating whether an alliance is predicted to be present or absent based on indicator simulation ("1sim"); and eleven layers with values that represented the mean of ten simulations ("Msim"). Generally, the mean of 100 simulations should approximate the kriged result (Burrough and MacDonnell 1998)—the mean of ten simulations could retain some characteristic roughness of the data, with the flexibility of having non-binary values.

To represent the neighborhood around each cell, the values of each of the spatial dependence variables for the eight surrounding grid cells for each observation were summed using ARC/Info and added to the modeling datasets as the spatial dependence (SD) term (for K, Msim, or 1sim):

$$\text{Sd term} = \sum_{i=1}^{8} P(pres)_i \quad , \qquad (1)$$

The kriged/simulated value for each cell (P(pres)i) can range from 0 to 1, therefore the spatial dependence term representing the neighborhood sum, K_x/Msim_x, can range from 0, indicating no observations of presence nearby, to 8, indicating a cluster of observations of presence (Besag 1972; Augustin et al. 1996). Each of these spatial dependence variables was added to the GLM and CT models, along with the environmental variables. Once a subset of variables was selected for the nonspatial model, the spatial dependence term was added (always as a linear term) and any subsequently non-significant variables were removed.

3.2.2 Interpolated residuals

Model residuals, the difference between actual values and model predictions, can identify specific points that are either under- or over-predicted by the model. If a model fits the data well, there should be no systematic pattern in the residuals. Model residuals that exhibit spatial dependence can indicate a mis-specified model, either an important, and spatially dependent predictor variable has not been included correctly (i.e., incorrect model

specification) or at all (McMillen, 2003). Adding the model residuals to predictions can be an efficient way to increase model accuracy by applying to local adjustments to predictions based on 'global' models, when suitable variables are either unmeasurable or costly to obtain. Residuals were obtained for the training data for both GLM and CT models. Variograms were fit to the residual values and ordinary kriging was used to interpolate surfaces from the residuals. The result was a CT model residual surface and a GLM residual surface for each of the eleven alliances. Each of the residual surfaces was added to the model predictions and the resulting surfaces were named CT_Kres and GLM_Kres.
Gaussian conditional simulation assumes that the data used are normally distributed. The GLM and CT residuals were normalized to fit variograms, the variograms were used to simulate surfaces, and the surfaces were then transformed back from normalized to regular values and added to the model predictions. The result was a CT_Simres and GLM_Simres surface for each alliance.

3.3 Accuracy assessment

Classification accuracy was the focus of model assessment in this work, and ROC plots were used as the accuracy metric as they have several advantages over similar measures (e.g., Kappa). Most importantly they are threshold- and prevalence-independent (Fielding and Bell 1997; Fielding 1999). A ROC plot is obtained by plotting all sensitivity values (true positive fraction) on the y-axis against their equivalent (1 – specificity) (false positive fraction) values on the x-axis. Two or more models can be plotted together and their respective true and false positive fractions can be visually assessed. The area under the curve (AUC) of the resulting plot provides a measure of overall accuracy at all available thresholds. ROC curves are used only with binary data, therefore an AUC value greater than 0.5 indicates the model performs better than random sorting. AUC values greater than 0.75 are considered potentially useful in a species modeling context (Elith and Burgman, 2002; Pearce et al., 2002). Based on plots of sensitivity and specificity for a range of probability values, a threshold was selected to produce the binary present/absent maps.
To summarize, a total of twelve model prediction for each of the eleven alliances were compared:

1) a non-spatial CT model based on the environmental variables (referred to as NS);
2) the same CT model to which the kriged SD term was added (CT_K);
3) the same CT model to which the mean simulation SD term was added (CT_Msim);
4) the same CT model to which the simulation SD term was added (CT_1sim);
5) nonspatial CT model prediction to which kriged residuals were added (CT+Kres);
6) nonspatial CT model prediction to which simulated residuals were added;
7) a nonspatial GLM based on the environmental variables (NS);
8) the same GLM to which the kriged SD term was added (GLM_K);
9) the same GLM to which the mean simulation SD term was added (GLM_Msim);)
10) the same GLM to which the simulation SD term was added (GLM_1sim);
11) nonspatial GLM prediction to which kriged residuals were added (GLM+Kres); and
12) nonspatial CT prediction to which simulated residuals were added (GLM+Simres).

The classification accuracy of the models was compared using ROC plots based on predictions on the test data.

4. Results

4.1 Comparison of modeling success: methods and alliances

The AUC for all model predictions are shown in table 3. Figure 2 summarizes the AUC values (on test data) for each of the twelve models across all eleven alliances. Model

accuracy varied widely, with an AUC range of 0.3-0.4 between the least and most accurate model for all alliances. Summaries of model accuracy on test data for each alliance are shown in figure 3. The lowest model accuracies occurred with the rarest alliances, ATCA, GALL, and EPNE, although EPNE had very high accuracies with some models (GLM, GLM_1sim, and CT_Kres). However, when an alliance is very rare, higher model accuracy as assessed by ROC plots can occur based on a very small increase in the probability for one observation of presence. ATCA and GALL had low accuracies for all models. YUSC showed a slight decrease in accuracy when spatial dependence was incorporated for both GLM and CT models, however both of the non-spatial model accuracies were quite high. YUBR had consistently high accuracies for all models, indicating that it had specific environmental preferences that were described well by the variables used here. The two most common alliances (LATR and LATRAMDU) had both GLM and CT models that were improved by incorporating spatial dependence. Generalist alliances such as these are more likely to be influenced by local factors that were described by incorporating spatial dependence. It should be noted that spatial dependence in the data is both a function of the spatial structure of the population from which the samples were taken, as well as the ability of the samples to characterize spatial dependence.

Table 3: AUC from ROC plots for all models assessed with test data. The models shown are nonspatial (NS), including kriged SD term (K), including mean simulation term (Msim), including single simulation term (1sim), kriged residuals added to nonspatial model predictions (Kres), and simulated residuals added to model predictions (Simres). All values are significant at $p < 0.001$ unless otherwise noted. * $p < 0.05$; † $p > 0.05$.

	GLM						CT					
Alliance	NS	K	Msim	1sim	Kres	Simres	NS	K	Msim	1sim	Kres	Simres
ATCA	0.670† *(0.117)*	0.708* *(0.125)*	0.708* *(0.124)*	**0.711*** *(0.125)*	0.597† *(0.139)*	**0.855** *(0.055)*	0.610† *(0.125)*	0.638† *(0.125)*	0.709* *(0.124)*	0.639† *(0.125)*	0.627† *(0.134)*	0.716† *(0.118)*
ATCO	0.875 *(0.037)*	0.908 *(0.033)*	**0.932** *(0.025)*	**0.932** *(0.024)*	**0.955** *(0.011)*	0.822 *(0.050)*	0.822 *(0.050)*	0.787 *(0.051)*	0.857 *(0.044)*	0.826 *(0.053)*	0.895 *(0.035)*	0.895 *(0.030)*
CORA	0.901 *(0.022)*	0.965 *(0.007)*	**0.966** *(0.007)*	0.923 *(0.018)*	**0.960** *(0.009)*	0.904 *(0.026)*	0.895 *(0.039)*	0.855 *(0.056)*	0.741 *(0.070)*	0.923 *(0.018)*	0.908 *(0.035)*	0.775 *(0.068)*
EPNE	0.937 *(0.025)*	0.689† *(0.148)*	0.781* *(0.136)*	**0.952** *(0.019)*	0.697† *(0.149)*	0.598† *(0.146)*	0.791* *(0.136)*	0.598† *(0.146)*	0.597† *(0.146)*	0.507† *(0.128)*	**0.949** *(0.028)*	0.807* *(0.131)*
GALL	**0.706*** *(0.074)*	0.583† *(0.106)*	0.579† *(0.106)*	0.599† *(0.098)*	0.674† *(0.109)*	**0.746** *(0.044)*	0.635† *(0.114)*	0.510† *(0.095)*	0.534† *(0.102)*	0.525† *(0.092)*	0.616† *(0.112)*	0.509† *(0.121)*
LATR	0.735 *(0.023)*	0.814 *(0.020)*	**0.818** *(0.019)*	0.790 *(0.020)*	**0.817** *(0.021)*	0.770 *(0.023)*	0.716 *(0.024)*	0.813 *(0.021)*	0.782 *(0.023)*	0.760 *(0.024)*	0.734 *(0.025)*	0.748 *(0.023)*
LATR AMDU	0.839 *(0.013)*	**0.886** *(0.011)*	0.876 *(0.011)*	0.844 *(0.012)*	**0.898** *(0.010)*	0.817 *(0.014)*	0.825 *(0.013)*	0.869 *(0.012)*	0.862 *(0.012)*	0.833 *(0.013)*	0.884 *(0.011)*	0.827 *(0.013)*
MESP	**0.849** *(0.077)*	0.634† *(0.104)*	0.630† *(0.105)*	0.794 *(0.086)*	0.623† *(0.103)*	0.721* *(0.106)*	0.661† *(0.134)*	0.595† *(0.103)*	0.594† *(0.103)*	0.571† *(0.099)*	0.712* *(0.104)*	0.580† *(0.109)*
PIMO	**0.910** *(0.064)*	0.872 *(0.075)*	0.789 *(0.089)*	0.786 *(0.089)*	0.901 *(0.068)*	**0.914** *(0.062)*	0.786 *(0.089)*	0.832 (0.083)	0.831 *(0.083)*	0.831 *(0.84)*	0.840 *(0.079)*	0.757* *(0.097)*
YUBR	0.926 *(0.012)*	0.967 *(0.112)*	**0.969** *(0.009)*	0.956 *(0.011)*	**0.977** *(0.005)*	0.938 *(0.002)*	0.867 *(0.028)*	0.941 *(0.018)*	0.962 *(0.012)*	0.937 *(0.016)*	0.965 *(0.006)*	0.899 *(0.022)*
YUSC	0.903 *(0.014)*	0.876 *(0.032)*	0.896 *(0.026)*	0.902 *(0.014)*	0.909 *(0.022)*	0.828 *(0.038)*	**0.917** *(0.014)*	0.769 *(0.044)*	0.783 *(0.043)*	0.814 *(0.035)*	0.881 *(0.032)*	0.849 *(0.030)*

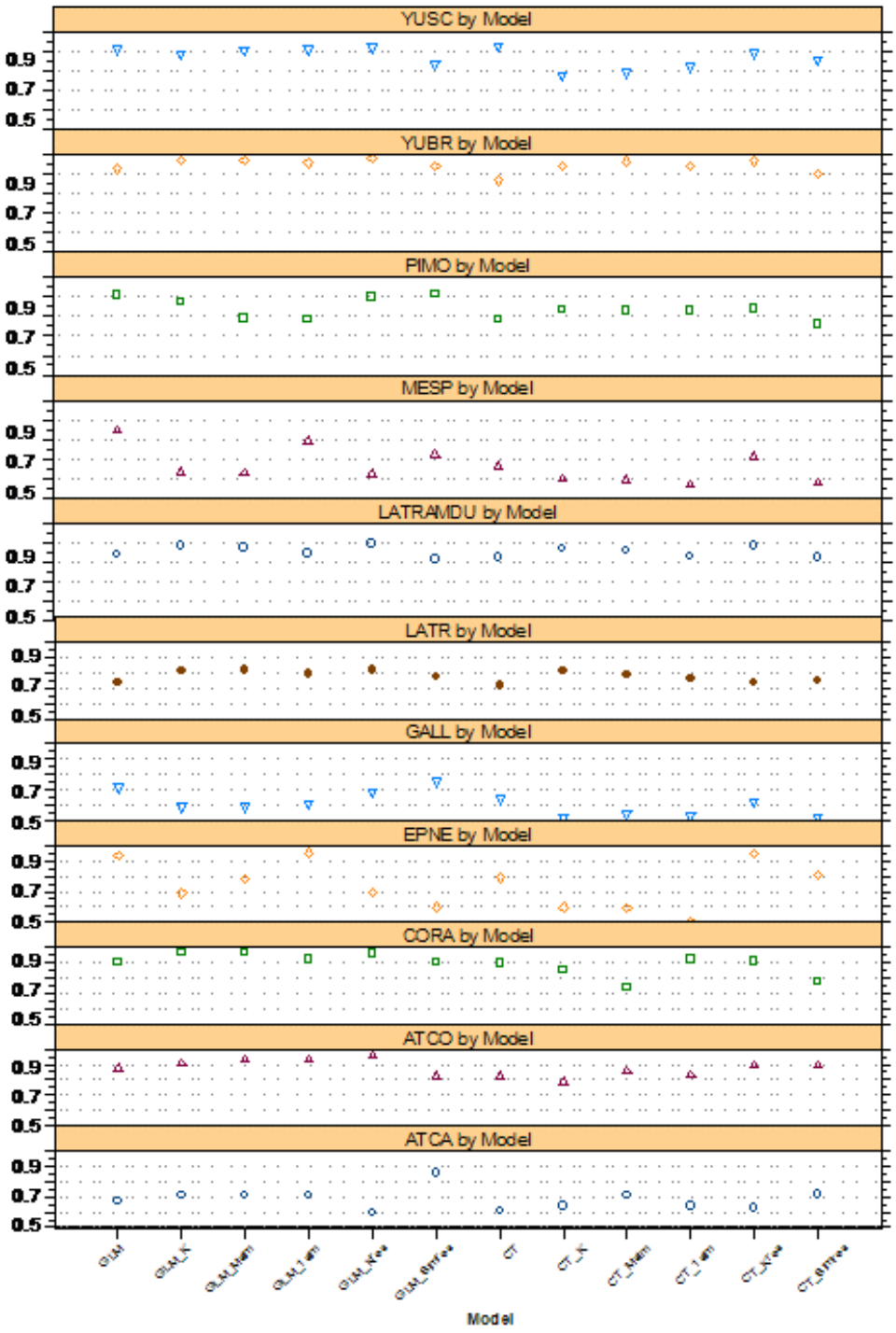

Figure 2: Summary of AUC on test data for each model

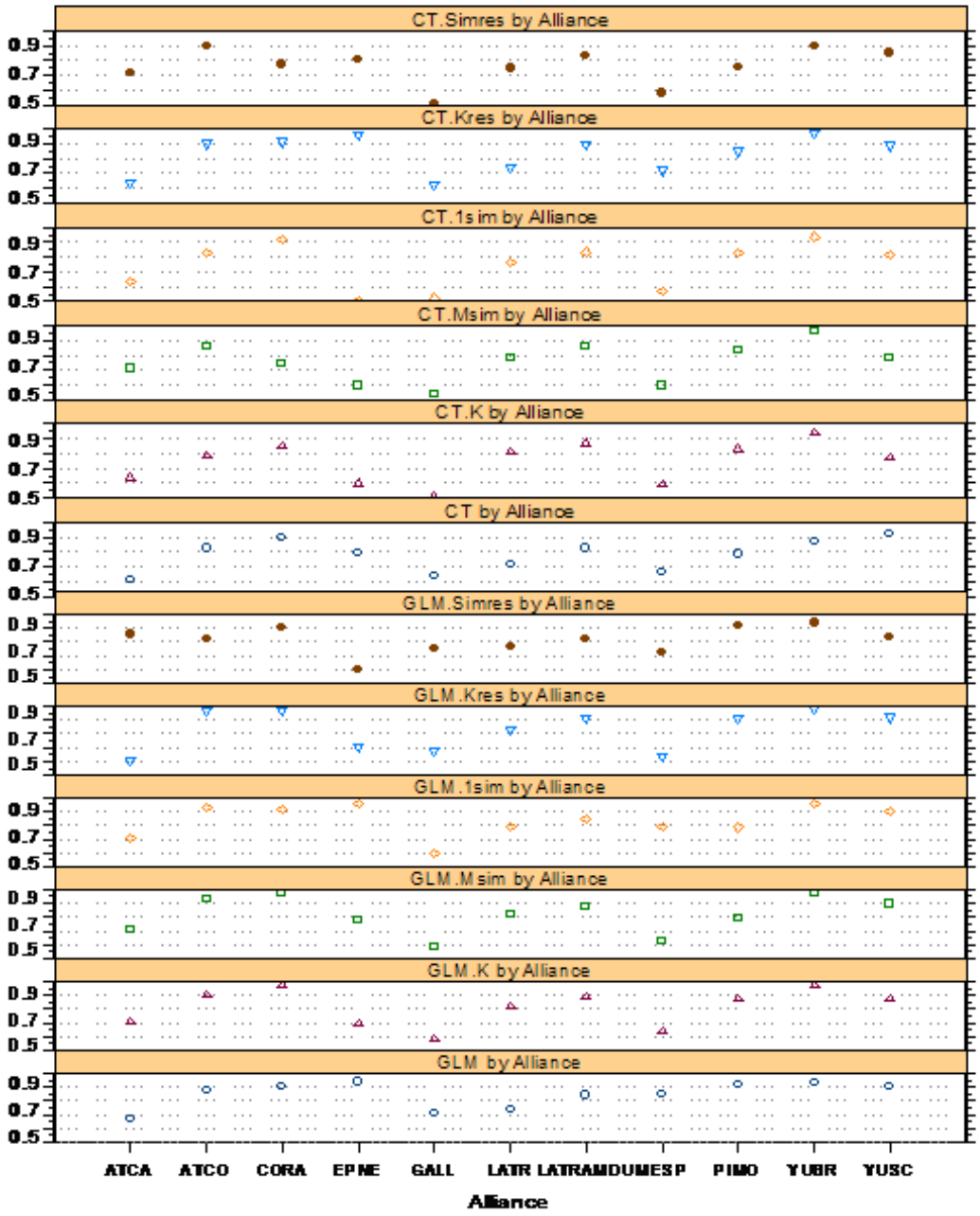

Figure 3: Summary of AUC on test data for each alliance

Figure 4 illustrates the potential advantages of incorporating spatial dependence, using predictions of YUBR presence in a section of the study area outlines in figure 1. A threshold was selected based on plots of sensitivity, specificity, and total accuracy (see Miller, 2005 for details) in order to create the presence/absence maps, with test data plotted on it. While this was not always the case, the nonspatial GLM here resulted in a number of commission errors (false positives) shown in the dashed circle. All of the spatial methods corrected this by predicting these observations to be absent. Conversely, the non-spatial CT model resulted in a number of omission errors (false negatives) that were subsequently corrected by the spatial models. In general, where spatial dependence does improve the model, adding the kriged residuals consistently produces among the best results for both GLM and CT models. When the nonspatial models had moderate to high accuracy, spatial dependence improves the accuracy at least slightly. If the nonspatial models had high accuracy, particularly the GLMs, spatial dependence tends to decrease the model accuracy. The environmental variables could be capturing most of the important spatial dependence, rendering the explicit spatial dependence, as a model term or residuals, redundant. In most cases, when the nonspatial model accuracy is low, spatial dependence does not improve it. Some alliances are not amenable to modeling in this context, and the spatial dependence used here may not be as important in describing their distribution.

5. Discussion

In general, GLMs produced more accurate models than CTs. This is somewhat different from results reported by other PVM studies that compared GLMs and CTs (Franklin, 1998; Vayssières et al., 2000; Cairns, 2001). I used GAMs and CTs with the same data to suggest variable transformations used in the GLMs, and this may be responsible for the generally higher accuracy in the GLMs. Although the GLM accuracy was higher in general, CT models had relatively higher accuracy with common alliances than with rare alliances; GLMs had somewhat higher accuracies with rare alliances than with common alliances.

When spatial dependence was incorporated explicitly as a model term, the resulting prediction accuracy was consistently improved for more common alliances. The variogram used to interpolate the spatial dependence term is only as good as the data to which it is fit. More common alliances provide more information on locations of present observations, and that generally results in more evidence of positive spatial autocorrelation. In a recent review of habitat distribution models that have incorporated spatial dependence, Miller et al., (submitted) concluded that limited availability of sample data at appropriate and varying spatial resolutions has been an important limiting factor in the ability of models to describe spatial dependence well enough to include it.

However, a more common lliance may have very high nonspatial model accuracy, in which case incorporating spatial dependence as a model term reduces the prediction accuracy (e.g., YUSC). When incorporated explicitly as a model term, spatial dependence tends to overwhelm the effects of the environmental variables, resulting in less robust models. Adding interpolated residuals resulted in the highest (or close to the highest) accuracies for nine alliances: ATCA, ATCO, CORA, EPNE, GALL, LATR, LATRAMDU, PIMO and YUBR. The only two alliances with highest model accuracy that did not involve interpolated residuals were MESP and YUSC, both of which had highest accuracy with non-spatial models (GLM and CT respectively). Incorporating spatial dependence as a model term resulted in model accuracies as high as the residual models with three alliances: CORA, EPNE, and LATR.

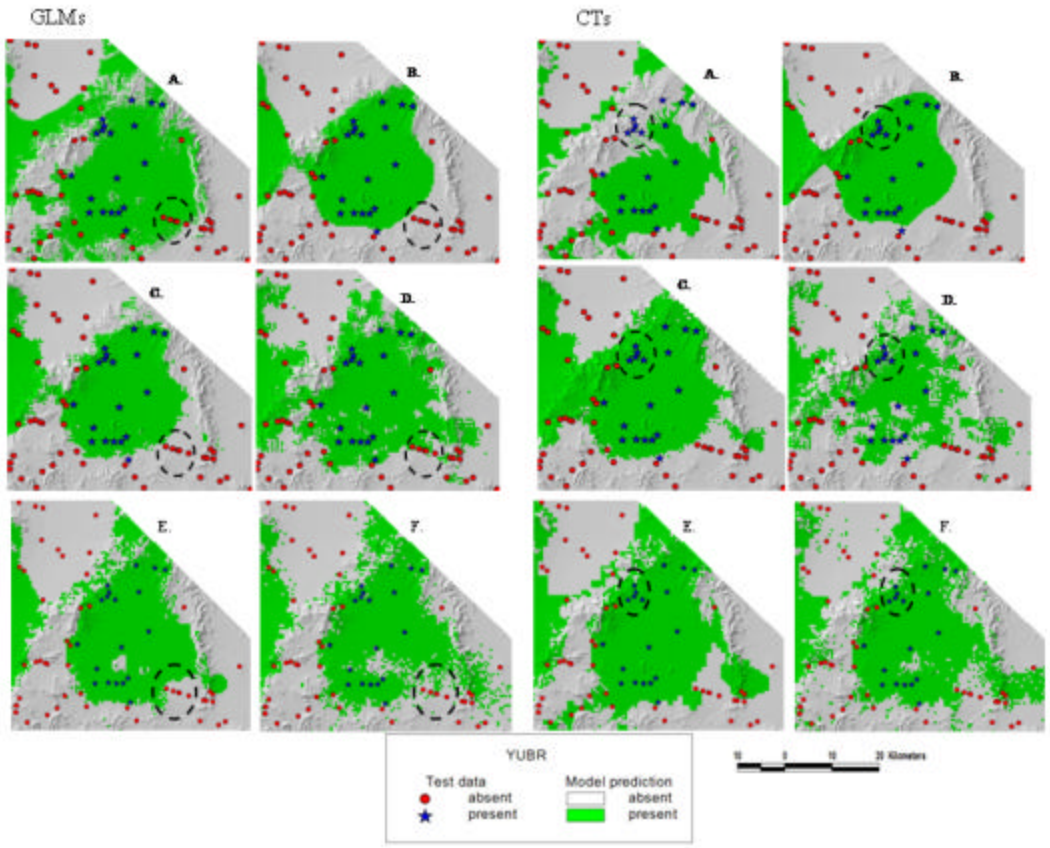

Figure 4: Maps of predicted presence of YUBR in the section shown in figure 1: (A) non-spatial model; (B) model with kriged SD term (C) model with mean simulation SD term; (D) model with simulation SD term; (E) prediction plus kriged residuals; (F) prediction plus simulated residuals. Dashed circles indicate FP errors that are corrected in the spatial GLMs, and FN errors that are corrected in the spatial CT models.

Although both methods of incorporating spatial dependence had similar effects for some alliances, in general I found that incorporating model residuals more consistently increased prediction accuracy. This is likely due to a combination of factors. Using model residuals allows for important environmental correlations to be maintained, rather than replaced, resulting in ultimately more generalizable results (as demonstrated with the test data here). Spatial dependence in the model residuals is based on more than just spatial dependence among observations, allowing for a more complex result to be included in the models.

6. Acknowledgements

The author wishes to thank Janet Franklin and the Geography Department at San Diego State University for support during this research.

7. References

Akaike H (1973) Information theory and an extension of the maximum likelihood. Pages 267-281 *in* 2nd International Symposium of Information Theory, Budapest

Augustin N, Mugglestone M, and Buckland S (1996) An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology* **33**:339-347

Austin MP, and Smith TM (1989) A new model for the continuum concept. *Vegetatio* **83**:35-47

Austin MP, Nicholls AO, Doherty MD, and Meyers JA (1994) Determining species response functions to an environmental gradient by means of a beta-function. *Journal of Vegetation Science* **5**:215-228

Beatley J (1975) Climates and vegetation pattern across the Mojave/Great Basin Desert transition of Southern Nevada. *The American Midland Naturalist* **93**:53-70

Beers T, Dress P, and Wensel L (1966) Aspect transformation in site productivity research. *Journal of Forestry* **64**:691-692

Besag J (1972) Nearest-neighbour systems and the autologistic model for binary data. *Journal of the Royal Statistical Society B* **34**:75-83

Beven K, and M.Kirkby (1979) A physically based variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin* **24**:43-69

Borcard D, Legendre P, and Drapeau P (1992) Partialling out the spatial component of ecological variation. *Ecology* **73**:1045-1055

Breiman L, Freedman J, Olshen R, and Stone C (1984) *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA

Brown D (1994) Predicting vegetation types at treeline using topography and biophysical disturbance variables. *Journal of Vegetation Science* **5**:641-656

Burrough P, and McDonnell R (1998) *Principles of Geographical Information Systems*. Oxford University Press, Oxford

Cairns DM (2001) A comparison of methods for predicting vegetation type. *Plant Ecology* **156**:3-18

Davis F, and Goetz S (1990) Modeling vegetation pattern using digital terrain data. *Landscape Ecology* **4**:69 - 80

Dokka R, C. Christenson, and J. Watts (1999) Geomorphic landform and surface composition GIS of the Mojave Desert Ecosystem in California. Pages 39-40 *in* IV International Conference on GeoComputation, Fredericksburg, VA

Dubayah R (1994) Modeling a solar radiation topoclimatology for the Rio Grande River Basin. *Journal of Vegetation Science* **5**:627-640

Elith J, and Burgman M (2002) Predictions and their validation: rare plants in the Central Highlands, Victoria, Australia.In: Scott J, Heglund P, Morrison M, Haufler J, Raphael M, Wall W, and Samson F, eds. *Predicting Species Occurrences; Issues of Accuracy and Scale*. Island Press, Washington, 303-313

Fels J. 1994. Modeling and Mapping Potential Vegetation using Digital Terrain Data. Ph.D. North Carolina State University, Raleigh.

Fielding AH, and Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* **24**:38-49

Fielding AH (1999) How should accuracy be measured?In: Fielding AH, editor. *Machine Learning Methods for Ecological Applications*. Kluwer Academic, Boston, MA, 209-223

Florinsky IV (1998) Combined analysis of digital terrain models and remotely sensed data in landscape investigations. *Progress in Physical Geography* **22**:33-60

Franklin J (1995) Predictive vegetation mapping: ge ographic modeling of biospatial patterns in relation to environmental gradients. *Progress in Physical 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. *Journal of Vegetation Science* **9**:733-748

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

Franklin J, Keeler-Wolf T, Thomas K, Shaari D, Stine P, Michaelsen J, and Miller J (2001) Stratified sampling for field survey of environmental gradients in the Mojave Desert Ecoregion.In: Mill ington A, Walsh S, and Osborne P, eds. *GIS and Remote Sensing Applications in Biogeography and Remote Sensing*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 229-251

Gotway CA, and Hartford AH (1996) Geostatistical methods for incorporating auxiliary information in the prediction of spatial variables. *Journal of Agricultural, Biological, and Environmental Statistics* **1**:17-39

Graae BJ, Økland RH, Petersen PM, and Fritzbøger B (2004) Influence of historical, geographical and environmental variables on understorey composition and richness in Danish forests. *Journal of Vegetation Science* **15**:465-474

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, and Sneddon L. 1998. International Classification of Ecological Communities: Terrestrial Vegetation of the United States. The National Vegetation Classification System, Development,Status and Applications. 1, The Nature Conservancy, Washington, D.C.

Guisan A, and Zimmermann N (2000) Predictive habitat distribution models in ecology. *Ecological Modelling* **135**:147-186

Guisan A, Edwards T, and Hastie T (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling* **157**:89-100

Gumpertz M, Graham J, and Ristaino J (1997) Autologistic model of spatial pattern of Phytophthora epidemic in bell pepper: Effects of soil variation on disease presence. *Journal of Agricultural, Biological and Environmental Statistics* **2**:131-156

Hastie T, Tibshirani R, and Friedman J (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, New York

Hubbell SP, Ahumada JA, Condit R, and Foster RB (2001) Local neighborhood effects on long-term survival of individual trees in a neotropical forest. *Ecological Research* **16**:859-875

Hunt CB (1966) Plant Ecology of Death Valley. *Geological Survey Professional Paper* **509**:1-68

Keitt TH, Bjornstad ON, Dixon PM, and Citron-Pousty S (2002) Accounting for spatial pattern when modeling organism-environment interactions. *Ecography* **25**:616-625

Legendre P, and Fortin MJ (1989) Spatial pattern and ecological analysis. *Vegetatio* **80**:107-138

Legendre P (1993) Spatial autocorrelation: problem or new paradigm? *Ecology* **74**:1659-1673

Legendre P, and Legendre L (1998) *Numerical Ecology*, 2nd English Edition edition. Elsevier, Amsterdam

Lobo JM, Lumaret J-P, and Jay-Robert P (2002) Modelling the species richness of French dung beetles (Coleoptera, Scarabaeidae) and delimiting the predictive capacity of different groups of explanatory variables. *Global Ecology & Biogeography* **11**:265-277

Lobo JM, Jay-Robert P, and Lumaret J-P (2004) Modelling the species richness distribution for French Aphodiidae (Coleoptera, Scarabaeoidea). *Ecography* **27**:145-156

McAuliffe J (1994) Landscape evolution, soil formation, and ecological processes in Sonoran desert bajadas. *Ecological Monographs* **64**:111-148

McCullagh P, and Nelder J (1989) *Generalized Linear Models*. Chapman & Hall, London

McMillen DP (2003) Spatial autocorrelation or model misspecification? *International Regional Science Review* **26**:208-217

Miller J, and Franklin J (2002) Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. *Ecological Modelling* **157**:227-247

Miller J (2005) Incorporating spatial dependence in predictive vegetation models: Residual interpolation methods. *The Professional Geographer* **57**:169-184

Miller J, Franklin J, and Aspinall RJ (submitted) Incorporating spatial dependence in predictive vegetation models.

Moore I, Grayson R, and Ladson A (1991) Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. *Hydrological Processes* **5**:3-30

Nogués-Bravo D, and Martinez-Rica JP (2004) Factors controlling the spatial species richness pattern of four groups of terrestrial vertebrates in an area between two different biogeographic regions in northern Spain. *Journal of Biogeography* **31**:629-640

Parker K (1991) Topography, substrate, and vegetation patterns in the northern Sonoran Desert. *Journal of Biogeography* **18**:151-163

Pearce JL, Venier LA, Ferrier S, and McKenney DW (2002) Measuring prediction uncertainty in models of species distribution.In: Scott J, Heglund P, Morrison M, Haufler J, Raphael M, Wall W, and Samson F, eds. *Predicting Species Occurrences; Issues of Accuracy and Scale*. Island Press, Washington, 383-390

Pontius RG, and Schneider LC (2001) Land-cover change model validation by an ROC method for the Ipswich watershed, Massachusetts, USA. *Agriculture, Ecosystems and Environment* **85**:239-248

Rowlands P, Johnson H, Ritter E, and Endo A (1982) The Mojave Desert.In: Bender G, editor. *Reference Handbook on the Deserts of North America*. Greenwood Press,

Westport, CT, 103-162

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

Smith PA (1994) Autocorrelation in logistic regression modeling of species' distributions. *Global Ecology and Biogeography Letters* **4**:47-61

Sui D (2004) Tobler's first law of geography: A big idea for a small world? *Annals of the Association of American Geographers* **94**:269-277

Thomas K, Franklin J, Keeler-Wolf T, and Stine P. 2004. Mojave Desert Ecosystem Program Central Mojave Vegetation Mapping Project. USGS-BRD, DoD, Flagstaff.

Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. *Economic Geography* **46**:234-240

Valverde P, Zavala-Hurtado A, Montana C, and Ezcurra E (1996) Numerical analysis of vegetation based on environmental relationships in the Southern Chihuahuan Desert. *The Southwestern Naturalist* **41**:424-433

Vayssières MP, Plant RE, and Allen-Diaz BH (2000) Classification trees: an alternative non-parametric approach for predicting species distributions. *Journal of Vegetation Science* **11**:679-694

Wilson JP, and Gallant JC (1998) Terrain-based approaches to environmental resource evaluation.In: Lane SN, Richards KS, and Chandler JH, eds. *Landform Monitoring, Modelling and Analysis*. John Wiley & Sons, Chichester, England, 219-240

Wu H, and Huffer FW (1997) Modelling the distribution of plant species using the autologistic regression model. *Environmental and Ecological Statistics* **4**:49-64

Yeaton RI, W. YR, Waggoner J, and Horenstein J (1985) The ecology of Yucca (Agavaceae) over an environmental gradient in the Mojave Desert: distribution and interspecific interactions. *Journal of Arid Environments* **8**:33-44