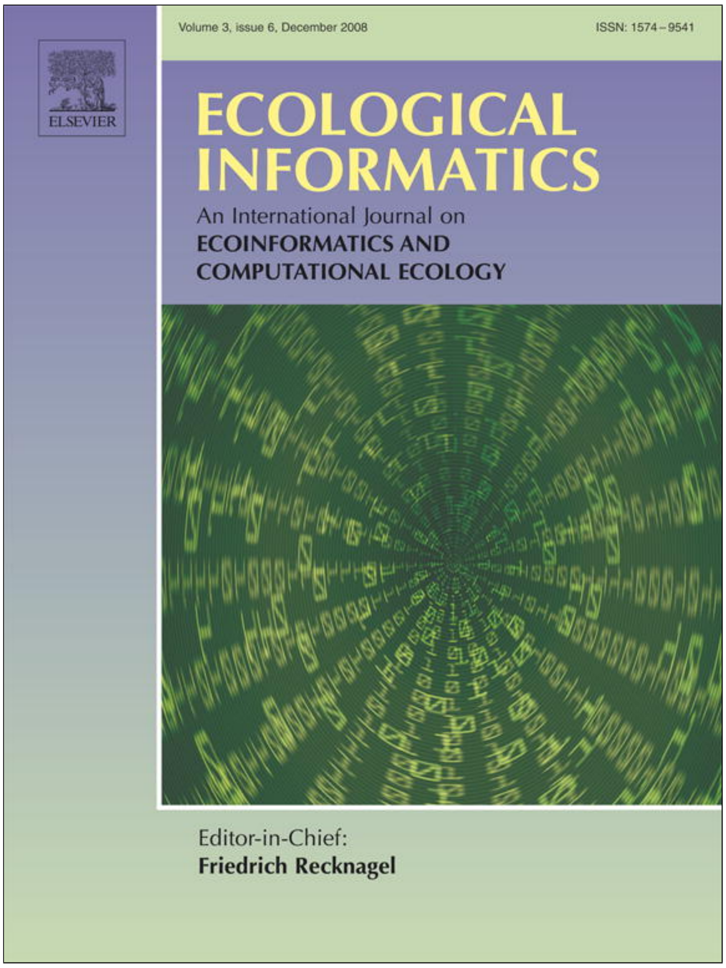
Volume 3, issue 6, December 2008
ISSN: 1574-9541
ELSEVIER
ECOLOGICAL INFORMATICS
An International Journal on
ECOINFORMATICS AND
COMPUTATIONAL ECOLOGY
Editor-in-Chief:
Friedrich Recknagel

ECOLOGICAL INFORMATICS 3 (2008) 375–386

available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/ecolinf

Predictive modeling and mapping sage grouse (*Centrocercus urophasianus*) nesting habitat using Maximum Entropy and a long-term dataset from Southern Oregon

Andrew C. Yost[a,*], *Steven L. Petersen*[b], *Michael Gregg*[c], *Richard Miller*[d]

[a] *Oregon Department of Forestry, 2600 State St., Salem OR 97310, United States*
[b] *Department of Plant and Wildlife Sciences, 275 WIDB, Brigham Young University, Provo, UT 84602, United States*
[c] *U.S. Fish and Wildlife Service, Mid-Columbia River NWR Complex, 64 Maple St., Burbank, WA 99323, United States*
[d] *Department of Rangeland Ecology and Management, 202 Strand Ag Hall, Oregon State University, Corvallis, OR 97331-2218, United States*

ARTICLE DATA

Article history:
Received 5 March 2008
Received in revised form
12 August 2008
Accepted 18 August 2008

Keywords:
Sage grouse
Predictive modeling and mapping
Maximum Entropy
Hart Mountain National Antelope Refuge

ABSTRACT

Predictive modeling and mapping based on the quantitative relationships between a species and the biophysical features (predictor variables) of the ecosystem in which it occurs can provide fundamental information for developing sustainable resource management policies for species and ecosystems. To create management strategies with the goal of sustaining a species such as sage grouse (*Centrocercus urophasianus*), whose distribution throughout North America has declined by approximately 50%, land management agencies need to know what attributes of the range they now inhabit will keep populations sustainable and which attributes attract disproportionate levels of use within a home range. The objectives of this study were to 1) quantify the relationships between sage grouse nest-site locations and a set of associated biophysical attributes using Maximum Entropy, 2) find the best subset of predictor variables that explain the data adequately, 3) create quantitative sage grouse distribution maps representing the relative likelihood of nest-site habitat based on those relationships, and 3) evaluate the implications of the results for future management of sage grouse. Nest-site location data from 1995 to 2003 were collected as part of a long-term research program on sage grouse reproductive ecology at Hart Mountain National Antelope Refuge. Two types of models were created: 1) with a set of predictor variables derived from digital elevation models, a field-validated vegetation classification, and UTM coordinates and 2) with the same predictors and UTM coordinates excluded. East UTM emerged as the most important predictor variable in the first type of model followed by the vegetation classification which was the most important predictor in the second type of model. The average training gain from ten modeling runs using all presence records and randomized background points was used to select the best subset of predictors. A predictive map of sage grouse nest-site habitat created from the application of the model to the study area showed strong overlap between model predictions and nest-site locations.

1. Introduction

Understanding the quantitative relationships between a species and the biophysical features of the ecosystem in which it occurs is fundamental when developing a sustainable resource management policy for that species and ecosystem. Predictive modeling and mapping that is based on these relationships forms an analytical foundation for informed

[*] *Corresponding author.* Tel.: +1 503 945 7410.
E-mail address: ayost@odf.state.or.us (A.C. Yost).

1574-9541/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecoinf.2008.08.004

conservation planning, mapping patterns of biodiversity, detecting distributional changes from monitoring data, and quantifying how variation in species performance relates to one or more controlling factors (Guisan and Hofer, 2003; McCune, 2006; Phillips et al., 2006). For example, to create management strategies with the goal of sustaining a species such as sage grouse (*Centrocercus urophasianus*) whose distribution throughout North America has declined by approximately 50% (Aldridge and Brigham, 2002), land management agencies need to know what portions of their former range they now inhabit and what attributes make these areas capable of population sustainability and attract disproportionate levels of use within a home range. This study tests the application of predictive modeling and mapping to sage grouse nesting habitat as a method for generating information valuable to a sustainable wildlife management policy.

Considerable work has been conducted evaluating habitat attributes at the site level for sage grouse nesting (Aldridge and Brigham, 2002; Gregg et al., 1994; Holloran et al., 2005; Sveum et al., 1998), however, few studies have evaluated the attributes at the landscape level. Gregg (2006) studied the nutritional ecology of sage grouse productivity and chick survival in Southern Oregon and northern Nevada. He found forb consumption and high insect availability in Spring were important for brood production and chick survival. The quality and availability of nutritional resources, however, are not distributed homogeneously across the landscape nor are the optimal locations that sage grouse select for nesting. Even if food resources were abundant and of high quality, selection of a nest site that increases the chances of exposure to predation or lethal climate conditions would have negative effects on grouse productivity. Moreover, how the spatial distribution of these and other attributes (i.e. topography) influence site selection and distribution during pre-nesting, nesting, rearing, or wintering at the landscape level is unknown. Gregg's (2006) study produced an extensive database of sage grouse nest-site locations over an eight-year time period sufficient for creating predictive models of the relationships between nest-site location and the biophysical attributes that might be important in sage grouse productivity. Therefore, the purpose of this study was to quantify the relative importance of the relationships between nest-site locations with the biophysical features that accompany those locations and then map the spatial distribution of sage grouse nesting habitat.

Advancements in computer technology, statistical modeling, and Geographic Information Systems (GIS) software allow the knowledge of animal/habitat relationships to be used for predicting the geographic distribution of individual populations of wildlife species. Predictive species mapping was defined by Franklin (1995) as predicting the distribution of a particular species across a landscape from mapped environmental variables. Predictive species mapping is founded in ecological niche theory and gradient analysis and rests on the premise that species distributions can be predicted from the spatial distributions of environmental variables that correlate with or control the occurrence of a plant or animal. Environmental conditions at occurrence localities constitute samples from a species' realized niche which is smaller than its fundamental niche (Hutchinson, 1957; Phillips et al., 2006). There are three major steps involved with predictive modeling and mapping: 1) collect species-level occurrence data and associated biophysical attributes of the landscape, preferably with a randomized sampling design, 2) build the models to determine the best subset of predictors and their parameter coefficients, and 3) application of the models to GIS data or new sites to forecast probability of occurrence for unsampled locations within the range of the study area.

Unlike vegetation monitoring datasets that typically contain some sampling sites with a particular species present and some where it was absent (Yost, 2008), wildlife sampling datasets often consist of "presence-only" data. General purpose statistical methods such as generalized linear models can be used for presence/absence datasets but there are a limited number of options available for presence-only datasets. Recently, Phillips et al. (2004, 2006) introduced the use of the Maximum Entropy (Maxent) method for modeling species geographic distributions with presence-only data. Maxent is a general purpose machine learning method for making predictions or inferences from incomplete information. The method estimates a target probability distribution across a study area by finding the probability distribution that is closest to uniform, or spread-out, subject to a set of constraints that represent our incomplete information about the target distribution. The information available about the target distribution presents itself as a set of real-valued variables, or "features" and the constraints are that the expected value of each feature should match its empirical average (average value for a set of sample points taken from the target distribution). When applied to presence-only distribution modeling, the pixels of the study area make up the space on which the probability distribution is defined, pixels with known species occurrence records constitute the sample points, and the features are the predictor variables that have digital geographic representation.

In addition to creating quantitative probability maps, the shape of response function and strength of predictability for each predictor variable can be graphically and quantitatively evaluated. This provides the capability to discover which gradients are most influential in predicting the likely occurrence of a particular species given they can be represented in a geographic database. Knowledge of the strength and functional response of species occurrences with each predictor provides valuable information for identifying which landscape features should be the focus of management for habitat sustainability.

The specific objectives of this study were to 1) quantify the relationships between sage grouse nest-site locations and a set of associated biophysical attributes with Maxent, 2) find the best subset of predictor variables that explain the data adequately, 3) create quantitative sage grouse distribution maps representing the relative likelihood of nest-site habitat based on those relationships, and 4) evaluate the implications of the results for future management of sage grouse.

2. Materials and methods

2.1. Data and study area

Locations of sage grouse nest sites from 1995 to 2003 were collected as part of a long-term research program on sage grouse reproductive ecology at Hart Mountain National Antelope Refuge (HMNAR) (Byrne, 2002; Coggins, 1998; Gregg, 2006). Nest sites were located by monitoring radiomarked females and confirmed by visually observing hens on nests. Universal Transverse Mercator (UTM) coordinates of nest sites were obtained using Garmin hand-held GPS units. The nest-site locations used for model building ($n=240$) were limited to those that fell within the circular boundary (313 km^2) of the vegetation classification GIS layer.

The Antelope Refuge is located in the High Desert Ecological province in southeast Oregon (Fig. 1). Climate, soils, and vegetation across the study areas are characteristic of those found in the High Desert, Klamath, and Humboldt ecological provinces. Mean annual precipitations across most of the two study areas range from 300 mm at lower elevations increasing to >400 mm at higher elevations. Topographic characteristic of this area includes mountains dissected by deep canyons, rocky tablelands, and rolling plains ranging in elevation from 4000 to 8000 ft. Primary plant alliances across the two refuges consist of subspecies of sagebrush. Wyoming big sagebrush (*A. tridentata wyomingensis*) (ARTRW) typically forms extensive stands across the warm, dry lower elevations of the refuge. Low sagebrush (*A. arbuscula*), (ARAR) forms large patches across low and mid-elevation sites. Mountain big sagebrush (*Artemesia tridentata vaseyana*) (ARTRVA) occupies the higher, cooler, wetter portions of the refuge, forming a complex matrix of patches with ARAR. Potential native understories are composed of deep and shallow rooted perennial tussock grasses and forbs as well as annual forbs. Associated shrubs include bitterbrush (*Purshia tridentata*) (PUTR), curl-leaf mountain mahogany (*Cercocarpus ledifolius*) (CELE), snowberry (*Symphoricarpos oreophilus*), horsebrush (*Tetradymia* spp.), and green (*Chrysothamnus viscidiflorus*) and grey rabbitbrush (*Ericameria nauseosa*).

2.2. Predictive modeling with Maxent

The estimated Maxent probability distribution is exponential in a weighted sum of environmental features divided by a scaling constant (Eq. (1)) to ensure that the probability values

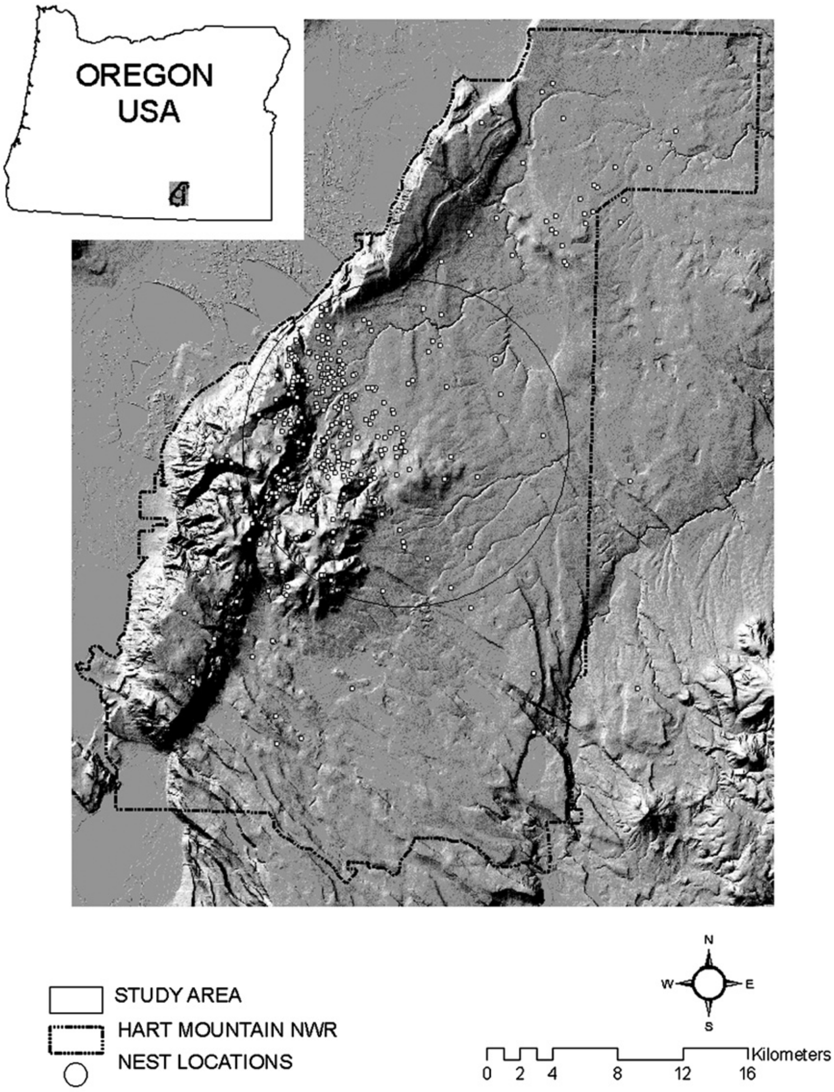

Fig. 1 – Hart Mountain National Wildlife Refuge, sage grouse nest-site locations, and circular boundary of the vegetation class GIS layer.

range from 0 to 1 and sum to 1. The Maxent probability distribution takes the form

$$q_\lambda(x)=\frac{e^{\lambda\cdot f(x)}}{Z_\lambda} \tag{1}$$

where λ is a vector of n real-valued coefficients or feature weights, f denotes the vector of all n features, and Z_λ is a normalizing constant that ensures that q_λ sums to 1. Maxent is a maximum-likelihood method that generates the probability distribution over the pixels in a grid of the modeling area. The program starts with a uniform distribution, and performs a number of iterations, each of which increases the probability of the sample locations for the species. The probability is displayed in terms of "gain", which is the log of the number of grid cells minus the log loss (average of the negative log probabilities of the sample locations). The gain starts at zero (the gain of the uniform distribution), and increases as the program increases the probabilities of the sample locations. The gain increases iteration by iteration, until the change from one iteration to the next falls below the convergence threshold, or until maximum iterations have been performed. The gain is a measure of the likelihood of the samples. For example, if the gain is 2, it means that the average sample likelihood is exp(2)≈7.4 times higher than that of a random background pixel. The uniform distribution has gain 0, so the gain can be interpreted as representing how much better the distribution fits the sample points than the uniform distribution does. The gain is closely related to "deviance", as used in generalized linear models (Phillips et al., 2006). The sequential-update algorithm is guaranteed to converge to the optimum probability distribution and because the algorithm does not use randomness, the outputs are deterministic.

To control over fitting, Maxent constrains the estimated distribution so that the average value for a given predictor is close to the empirical average (within empirical error bounds) rather than equal to it. This smoothing procedure is called regularization and users can alter the parameters to potentially compensate for small sample sizes.

The Maxent distribution is calculated over the set of pixels representing the study area that have data for all environmental variables. However, if the number of pixels is very large, processing time increases without a significant improvement in modeling performance. For that reason, when the number of pixels with data is larger than 10,000 a random sample of 10,000 "background" pixels is used to represent the variety of environmental conditions present in the data. The Maxent distribution is then computed over the union of the "background" pixels and the samples for the species being modeled. Maxent's predictions for each analysis cell can be represented as cumulative values representing as a percentage the probability value for the current analysis cell and all other cells with equal or lower probability. The cell with a value of 100 is the most suitable, while cells close to 0 are the least suitable within the study area (Hernandez et al., 2006). The formulaic description of the Maxent modeling procedure applied to species occurrence data and a description of the Maxent program (version 2.0) used to perform the modeling in this study is given by Phillips et al. (2006).

2.3. *Predictor variables*

The set of seven predictor variables included both east and north Universal Transverse Mercator (UTM) location coordinates for each nest-site location as recorded from a GPS unit. Models with and without location variables can be compared to assess the amount of spatial autocorrelation in the dataset and evaluate the adequacy of the remaining ecological and geophysical gradients at fitting the dataset. Prior to 1995 GPS was not used to identify location coordinates. The error associated with obtaining coordinates from topographic maps would be too high for producing reliable models, therefore, the nest sites located prior to 1995 were excluded from the modeling dataset.

Elevation of sampled plots ranged from approximately 5233 to 7057 masl and was obtained from a 10 m digital elevation model (DEM). Elevation for the area of analysis ranged from 4584 to 7602 masl. Slope values ranged from 0° to 55° and were also derived from the 10 m DEM using the surface analysis slope function in ArcGIS® Spatial Analyst. Similarly, aspect values for each nest site were obtained from the DEM and categorized as flat, north (316° to 45°), east (46° to 135°), south (136° to 225°), and west (226° to 315°).

A modified version of Iverson et al.'s (1997) Integrated Moisture Index (IMI) was used as a relative rating to moisture availability. The IMI was based on three topographic factors (hillshade, flow accumulation, curvature) derived from the 10 m DEM of the study area. The hillshade GIS layer contributed 50% to the IMI and was created with the "hillshade" command in Arc/Info Grid (Environ, Sys. Res. Inst., 2002). Landscape curvature, representing convexity and concavity across the landscape, contributed 15% to the IMI and was created with the "curvature" program in Arc/Info. Flow accumulation contributed 35% to the IMI and was created with the "Flow accumulation" program.

The vegetation cover type associated with each nest site was treated as a categorical predictor variable (Fig. 2). In a separate analysis, cover types were delineated using remote

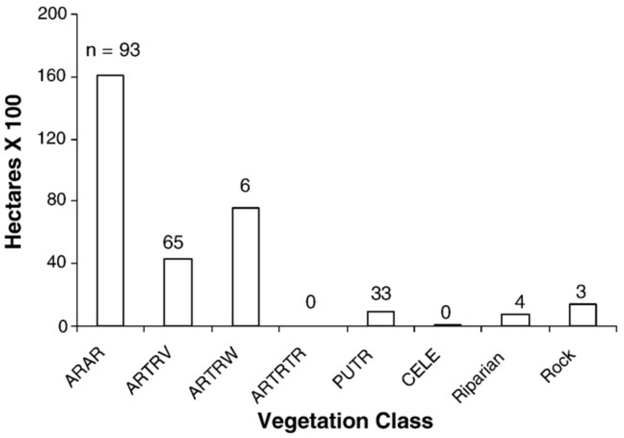

Fig. 2 – Amount of area, in thousands of hectares, within each modeling subcategory for the vegetation classification GIS layer. The number of nest sites from 1995 to 2003 for each subcategory is listed above each bar.

sensing and GIS. Plant community cover types were digitized on-screen from high-resolution (1:24,000) aerial photographs obtained from the United States Geological Survey (1 m resolution Digital Orthophoto Quadrangles) and the United States Department of Agriculture (USDA) Farm Service Agency's National Agriculture Imagery Program (NAIP, 2005) images. Differences in community cover types were identified using 1) distinct visual pattern recognized from the photographs, 2) plant association data obtained from field sample plots, and 3) the topographic position, aspect, slope, and elevation of individual cover types using digital elevation data. Polygons of each cover type were digitized within a 10 km radius (314 km^2) centered at the site of highest courtship activity (Fig. 1). An accuracy assessment was conducted comparing the actual vegetation type to the predicted type from 1500 randomly selected GIS locations that were visited during Summer 2006. The Kappa statistic (0.82) was better than substantial (Landis and Koch 1977) and the *p*-value for the z-statistic (72.23) was less than 0.001 indicating that the estimate was not due to chance alone.

2.4. *Model building*

The objective was to build a model with adequate performance with the best subset of predictors. This objective was accomplished by first building a GIS layer representing the spatial distribution of likely nesting habitat and secondly by identifying which variables, other than spatial coordinates, were most important in predicting that habitat. Maxent's jackknife test of variable importance can be used to evaluate the relative strengths of each predictor variable. The training gain is calculated for each variable alone and the drop in training gain when the variable is omitted from the full model. Therefore, to accomplish the first goal the modeling process started with a full model that contained all seven predictor variables. Then, the variable with the lowest decrease in the average training gain when omitted was removed and the remaining variables were used to build the model. For the second goal the north and east UTM predictor variables were omitted from the model building process to evaluate the predictive capabilities of the remaining five predictors.

Model performance can be evaluated by setting aside a subset of the presence records for training and use the remaining records to test the resulting model. Performance can vary depending upon the particular set of data withheld from building the model for testing, therefore, 10 random partitions of the presence records were made to assess the average behavior of Maxent, following Phillips et al. (2006). Each partition was created by randomly selecting 75% of the total 204 presence records (n=153) and 10,000 random background pixels treated as negative instances as training data. The remaining 25% of presence records (n=51) were used for testing the model. The full set of presence records were used to build the final reduced model to obtain the best estimate of the species distribution and for creating a GIS probability distribution map.

Linear, quadratic, product, and hinge functions of the predictor variables were selected for inclusion in the model. Model settings that allow the algorithm to get close to convergence are the maximum number of iterations, set to 1000, and the convergence threshold, set to 10^{-5}. The regularization multiplier was set to the default value of one.

The Maxent models were also evaluated with the binomial test to determine whether a model predicted the test localities significantly better than random. The binomial test requires that thresholds be used in order to convert continuous predictions into suitable and unsuitable areas for sage grouse nesting. After applying a threshold, model performance can be investigated using the extrinsic omission rate, which is the fraction of test localities that fall into pixels that are predicted as not suitable for sage grouse, and the proportional predicted area, which is the fraction of all the pixels that are predicted as suitable. The *p*-values associated with a cumulative threshold of one, five and ten are reported to show trend as the threshold varied.

The receiver operating characteristic (ROC) analysis was also used to evaluate how well the Maxent model compared to random prediction. The area under the ROC function (AUC) is an index of performance because it provides a single measure of overall accuracy that is independent of any particular threshold (Deleo, 1993). The ROC analysis assigns a threshold to the modeled probability values by which sampling units are classified as positive or negative for species presence. The sensitivity for a particular threshold is the fraction of all positive instances that are classified as present and specificity is the fraction of all negative instances that are classified as not present. A ROC plot is obtained by plotting all sensitivity values (true positive fraction) on the *y* axis against their equivalent (1 — specificity) values (false positive fraction) for all available thresholds on the *x* axis. In other words, a point (*x*, *y*) in the plot indicates that for some threshold, the classifier classifies a fraction *x* of negative examples as positive and a fraction *y* of positive examples as positive. Maxent treats the randomly selected background pixels as negative instances and the pixels in which the presence data fall as positive instances. The value of the AUC is typically between 0.5 and 1.0. A value of 0.8 indicates that, for 80% of the time, a random selection from the positive group will have a score greater than a random selection

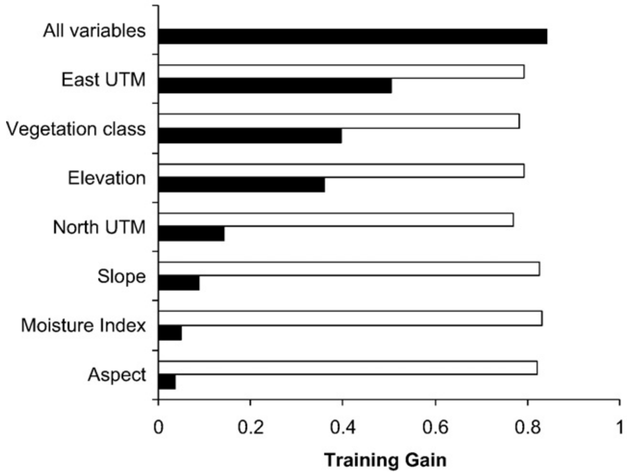

Fig. 3 – Training gain for each predictor variable alone (black bars) and the drop in training gain when the variable is removed from the full model (white bars).

from the negative class and a value of 0.5 indicates the model is no better than a random prediction. However, when ROC analysis is used on presence-only data, the maximum AUC is less than one (Wiley et al., 2003), and is smaller for wider-ranging species. The maximum achievable AUC can be shown to be equal to $1-a/2$, where a is the fraction of pixels covered by the species' distribution. A one-tailed Mann–Whitney-*U* statistic was used to test the null hypothesis that the AUC for the modeled predictions performed for the test data was not significantly ($\alpha=0.05$) different than random (Phillips et al., 2006). The sample for this test was comprised of $n=10$ sensitivity values (1 — test omission) at each 0.1 interval of the fractional predicted area from the Maxent omission output.

The success of the model can also be evaluated by visually inspecting how well the probability values in the output grid fit with the points of the presence records. Output grids are generated from application of the Maxent model to the set of GIS grids that represent each predictor variable. A good model will produce regions of high probability that cover the majority of presence records and areas of low probability should contain few to no presence points.

3. Results

3.1. *The spatial model*

The regularized training gain for the full model built with all presence records was 0.842. From the jackknife test of variable importance (Fig. 3) the single most important predictor variable, in terms of the gain produced by a one-variable model, was east UTM followed by vegetation class. North UTM decreased the gain the most when it was omitted from the full model, which suggests it contained the most information not present in the other variables. These results indicate there was a measurable spatial constraint to the nest location data within the modeling extent. Based on the amount of decrease in model gain when a variable was omitted, the order of variable removal for the spatial model was moisture index, slope, aspect, elevation, vegetation, and north UTM.

From the binomial test the *p*-values from all partitions and threshold categories were less than 0.005 indicating that Maxent produced predictions that were significantly better than random for all models regardless of the number of predictor variables or cumulative threshold value (Table 1). Binomial test *p*-values decreased considerably when the threshold changed from one to ten indicating a higher probability of rejecting the null hypothesis as threshold increased to 10. When a binary prediction is desired, the value of threshold to choose for establishing a boundary between the range of probability values classified as suitable habitat from those that are not is a critical issue and an area of research that remains to be done. A good rule needs to be developed to set a threshold operationally using intrinsic data (Phillips, 2006; Hirzel et al., 2006).

From the Mann Whitney Test the average AUC values (Fig. 4), and the individual AUC values ($n=10$) from all partitions, for all of the models, were statistically significant (p-value<0.05) indicating better-than-random prediction (Table 1). The average AUC values were relatively the same as model size decreased but dropped slightly with the one-variable model containing east UTM (Fig. 4).

The average training gain declined consistently as variables were removed (Fig. 4). There was a non-monotonic decline in the standard deviation (0.075 to 0.021) of the training gain values from the seven-variable to the one-variable partition and variability was higher in the behavior of the average test gain as model size decreased. The standard deviations of the test gain values from each partition were higher than those for the training gain ranging from 0.075 to 0.262.

Given the higher sensitivity of the average training gain relative to the average AUC value, the former metric was used to decide which of the best performing models should be used for mapping. Therefore, the logical choice of best model was the one that had the fewest predictors with an average

Table 1 – *p*-Values from the binomial test

Number of variables	Binomial test *p*-value for threshold			Average AUC *p*-value	Predictor variable removed
	1	5	10		
Spatial model					
7	1.87E−04	1.27E−07	4.98E−12	0.0079	
6	2.35E−04	1.27E−07	4.18E−12	0.0064	Moisture index
5	3.81E−04	3.55E−08	1.63E−12	0.0067	Slope
4	2.72E−04	3.28E−08	1.53E−11	0.0036	Aspect
3	2.42E−04	1.75E−08	1.80E−11	0.0041	Elevation
2	3.76E−02	1.77E−07	8.86E−10	0.0047	Vegetation
1	8.44E−04	2.58E−07	3.65E−09	0.0204	North UTM
Without east and north UTM					
5	2.23E−03	1.37E−05	1.65E−07	0.0132	
4	1.57E−03	1.06E−06	2.64E−08	0.0158	Slope
3	1.04E−03	7.65E−06	1.01E−05	0.0229	Moisture index
2	9.9E−04	4.03E−06	7.05E−08	0.0189	Aspect
1	1.33E−04	9.37E−05	2.53E−08	0.1684	Elevation

The average AUC *p*-values are from the Mann–Whitney Test. The column titled predictor variable removed lists the name of each variable as it was removed from the modeling set.

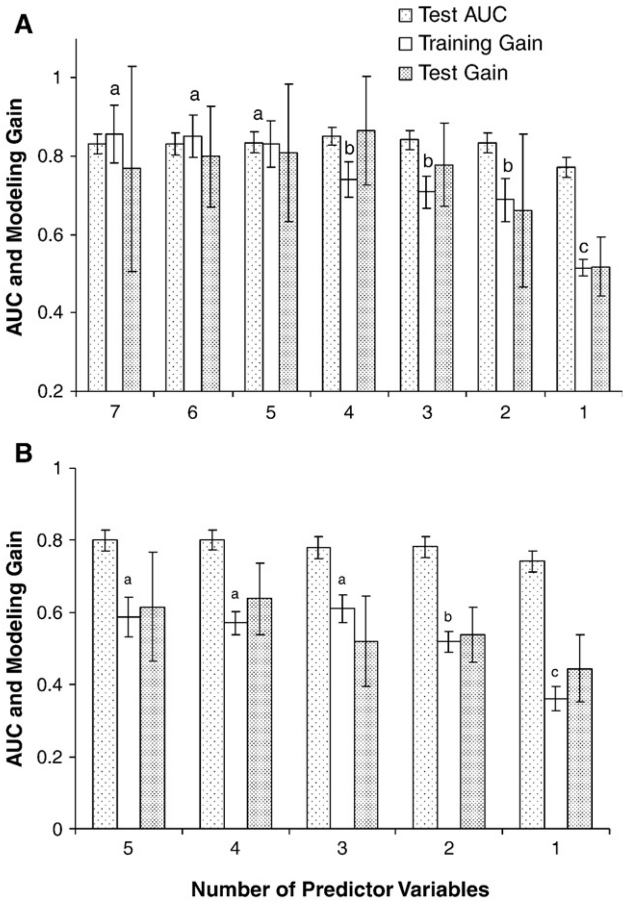

Fig. 4 – Values for the test AUC, training gain, and test gain averaged across the 10 random partitions of the presence records. The *x* axis represents the number of predictor variables in each model. Models that included east and north UTM as predictors are shown in the top graph (A) and those that omitted the spatial coordinates are shown in the bottom graph (B). Models with same lowercase letter above the bar for average training gain were not significantly different.

training gain not significantly different than the full model or the model with highest training gain. Using the overlap between 95% confidence intervals for training gain averages as the criteria for significance the five-variable model containing the spatial coordinates, vegetation class, elevation and aspect was not significantly different than the two larger models but was different than the remaining smaller models (Fig. 4). Therefore the five-variable model was used to create the distribution of potential nesting habitat for the Hart Mountain study area.

3.2. *Models without east and north UTM*

The regularized training gain for the five-variable model using all presence records but without the spatial coordinates was 0.604. The relative importance of the predictor variables, in terms of the gain was nearly the same as when the spatial coordinates were included in the model except that vegetation decreased the gain the most when omitted suggesting it contained the most information not contained in the other variables. The order of variable removal from the full model was slope, moisture index, aspect, and elevation leaving vegetation as the one-variable model (Table 1).

The results of the binomial test (Table 1) indicated that, on average, all of the models performed significantly better than random. An interesting feature of this analysis was the average AUC values (Fig. 4) were only slightly lower than those for the models with spatial coordinates and the decrease in value as variables were removed was small. The *p*-values from the Mann Whitney Test indicated that, on average, the AUC estimates for each model, except the one-variable model containing vegetation class, were significantly better-than-random prediction. However, the average *p*-values were generally larger than the model with spatial coordinates (Table 1).

Unlike the AUC values the average training and test gain were more sensitive to the removal of the spatial coordinates. There was a slight increase in average training gain for the three-variable model followed by a steep drop for the smaller models. The standard deviation of the five training gain averages ranged between 0.029 and 0.054 and the range for the test gain averages had consistently higher values from 0.076 to 0.152.

Using the overlap between 95% confidence intervals for training gain averages as the criteria for significance it appears that the Maxent model containing vegetation class, elevation, and aspect was not statistically different from the four or five-variable model. The average training gain for the two-variable model, containing vegetation and aspect, was significantly different from the three larger models and the one-variable model with vegetation only (Fig. 4).

This analysis shows that the spatial coordinates were powerful predictor variables for this dataset and the geographical extent of the study region. In fact, the average training gain for the model containing just the UTM coordinates (0.689 ± 0.034) alone was significantly higher than the best training gain among models with the coordinates omitted (0.610 ± 0.024). Nonetheless, vegetation class, elevation and aspect emerged as important landscape features for the nest-site locations of sage grouse at HMNAR.

3.3. *Predictor variables*

Both east and north UTM produced unimodal response profiles (Fig. 5). East UTM was skewed on the right and north UTM considerably flatter across the range of values. The exponent for east UTM dropped below zero at about 75% of its range of values which accurately reflects the lower density of nest locations in the easternmost portion of the study area.

The *P. tridentata* (PUTR) subcategory within the vegetation class predictor variable was associated with the highest increase in the exponent (Fig. 5). These results are consistent with what could logically be expected given PUTR had the highest ratio of presence records to area. Indeed, the value of this ratio, the number of presence records to area, was consistent with the pattern of values for the exponent across subcategories when the model contained only the vegetation class variable. It appears that the exponent value will be close to zero when a subcategory contains no presence records as was the case for ARTRTR and CELE and when the ratio is quite

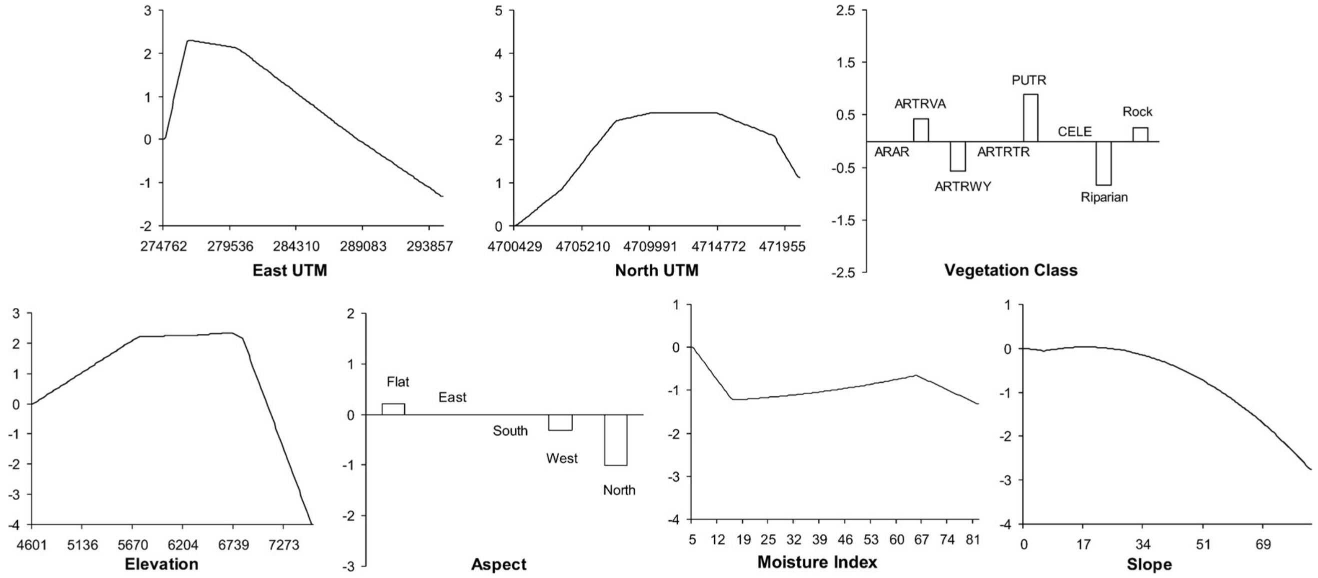

Fig. 5 – These curves show how each environmental variable affects the Maxent prediction when all presence records were used to build a full model. The raw Maxent model has the exponential form described by Eq. (1) and the curves show how the exponent changes as each environmental variable is varied, while all other variables are held constant at their average sample value. The shapes of these curves can be different than what is shown depending upon which predictor variables are included in the model.

small such as 0.0054 for the riparian subcategory. Negative exponent values were associated with subcategories that had even smaller presence records to area ratios such as 0.0008 for ARTRW) and 0.0021 for riparian. Aldridge and Boyce (2007) reported riparian areas were identified as risky habitats.

The value of the exponent started at zero for the lowest values of elevation, then increased to a little over two until 5760 ft asl, was relatively flat until 6760 ft after which it rapidly decreased to negative values for the highest elevations. Aspect was a difficult landscape feature to create a reliable predictor variable from because northern-most areas include the highest and lowest values and flat areas (32% of the study area) have no aspect. Thus, sine and cosine transformations of aspect are problematic. The alternative chosen for this study was to create a categorical variable with flat areas as one subcategory and all other areas with slope greater than zero as the other four categories. Aspect was a relatively weak predictor variable but the pattern of values for the exponent across the five subcategories may indicate an ecological signal. The pattern suggests that sage grouse show a slight avoidance of north and west facing slopes for nest sites. The thermal benefit of early morning exposure to solar radiation to nest success might have some driving force in the natural selection of the most successful nest-site-selection behaviors. This hypothesis then might be a reasonable explanation for the response in exponent across the five categories (Fig. 5). The value of the exponent was slightly positive for slopes of 12° to 24° and then became negative in a curvilinear trend as slope increased (Fig. 5). This pattern of response reflected the relationship that should be expected between nest-site selection and slope, that is, sage grouse avoid steep slopes for nest sites. The Integrated Moisture Index predictor variable performed poorly in the models and the response of the exponent across the range of values was negative. This result is consistent with the low increase in the Maxent exponent for the riparian subcategory in the vegetation class predictor and suggests that the digitally-derived hydrologic features of the landscape did not carry a meaningful signal for this data.

3.4. *Predictive mapping*

A five-variable model built from the full set of 204 nest-site locations was used to create the distribution map of sage grouse potential nest-site habitat for the study area. The predictor variables in this model included the spatial coordinates, vegetation, elevation, and aspect (Fig. 6). Visual inspection shows strong agreement between nest location points and the continuous probability distribution. The regions of highest nest-site locations were accurately associated with regions of high probability predicted by the model. However, even though Maxent predicted a relatively compact area of high nesting potential there were still a few nests placed in

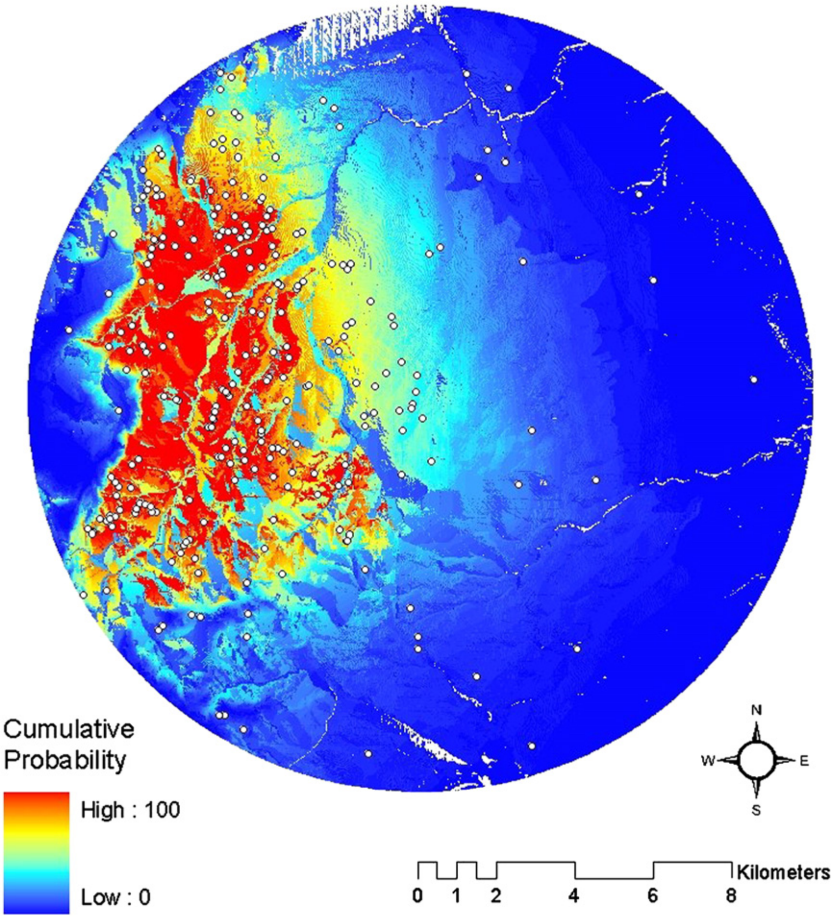

Fig. 6 – Predicted potential geographic distribution of nest-site habitat for sage grouse within the circular boundary of the vegetation class GIS layer. The color red represents areas with higher estimates for the probability of nest-site habitat.

areas quantified as low nesting potential over the 8 years of location information. This distribution map perhaps provides a foundation for further research to explore the nature of the relationship of nest success and recruitment to the probability distribution. This would be important for understanding the extent to which the Maxent models and probability maps predict source and sink habitats (Aldridge and Boyce, 2007).

The potential nesting habit distribution shown in Fig. 6 is, in large part, dependent upon the size and spatial extent of the vegetation classification GIS layer. Having a vegetation classification layer, along with layers for the other predictor variables, with a geographical boundary of the complete boundary of HMNAR would produce a slightly different nesting potential distribution than the one in Fig. 6.

4. Discussion

This study tests the application of predictive modeling and mapping to sage grouse nesting habitat with Maximum Entropy (Phillips, 2006) as a method for generating distribution information that is fundamental to a sustainable wildlife management policy. The Maximum Entropy modeling approach, when applied to Gregg's (2006) sage grouse dataset, was successful in achieving the objectives of this study. The relationships between sage grouse nest-site locations and a set of associated biophysical attributes were quantified using the Maxent software and probability distribution maps were created that locate the relative likelihood of nest-site habitat. The model adequately passed the fundamental test of predicting what could reasonably be expected, that is, there was good overlap between the spatial distribution of probability values (Fig. 6) and the highest densities of nest-site locations. The Maxent predictions are continuous thus allowing further distinction in areas suitable for nest-site location between marginally strong prediction versus those with increasingly stronger prediction. These distributions, therefore, provide an improvement over shaded outline maps of species distributions found in standard field guides (Phillips et al., 2006). Discrete distributions can be created from the continuous Maxent distribution by applying threshold values to filter output cells into categories of habitat suitability (Hirzel, 2006; Valverde and Lobo, 2007). Therefore, the continuous Maxent probability distribution is the more preferable modeling and mapping system to those that produce only discrete distribution output. More importantly, the comparisons of a suite of modeling methodologies made by Elith et al. (2006) found Maxent to be among the top performers.

Maxent was capable of combining linear, quadratic, product, and hinge features to capture the complex responses for the continuous variables and the two categorical predictors in the model. The response curves (Fig. 5) illustrate how the choices sage grouse made for nest-site selection were constrained by the chosen landscape features represented in the predictor variables. However, the shapes of these curves are not fixed and can change depending on the set of predictor variables and modeling features chosen for the model. The use of hinge features, in combination with Linear, Quadratic, and Product features, in the modeling process was critical to the flexible fitting of the response curves across the range of each predictor variable and resulted in measurable improvement in model predictability than when hinge features were not selected.

This study presented criteria for selecting a Maxent model with the best subset of predictor variables for the purposes of distribution mapping. The objective of the method was to identify a model with the fewest predictor variables that explained the data adequately (Burnham and Anderson, 2002). This objective was based on the principle of parsimony and the philosophy that models are only estimates of reality and that no single model is ever "true" or likely to perform well in all applications (Hilborn and Mangel, 1997). The process included creating a full model containing all the predictor variables, identifying the least informative predictor, creating a reduced model without that predictor, and repeating this process until only one variable remained. The model with the fewest predictor variables and an average training gain not significantly different than the model with highest training gain was selected as the best alternative. The overlap between 95% confidence intervals for training gain averages was used for the test of significance. Conversely it might be unwise to ignore small improvements in training gain with additional variables especially if the modeling objective is to find the most powerful set of predictor variables.

The Maxent models, expressed through the differences in training and test gain between models containing the UTM coordinates and those without them, suggest that the full set of environmental factors controlling the selection of nest sites within the study area were not sufficiently represented with the five other predictor variables. This interpretation assumes that a perfect model would contain a set of environmental variables that sufficiently described all the parameters of the species' fundamental niche relevant to its distribution at the grain of the modeling task (Phillips et al., 2006). Hence, the assumption is that in a perfect model the UTM coordinates would contribute very little to training gain values above what is produced by the other predictor variables. The results of Holloran et al. (2005) suggest that other attributes such as the amount of shrub and other plant species cover per unit area or the height of shrubs and other species might be important predictor variables. Indeed, in the eastern half of the study area the number of different classes and polygon density is much higher than the western half. Furthermore, Hernandez et al. (2006) confirmed the results of other researchers that the ecological characteristics of species affects model accuracy potential, where species widespread in both geographic and environmental space are generally more difficult to model than species with compact spatial distributions, as is the case with the sage grouse nest data. They also confirmed that the ability to model species effectively is strongly influenced by species ecological characteristics independent of sample size. Without the UTM coordinates the Maxent models produced higher probability values in areas where nest sites had not been recorded thus, potentially overestimating the distribution of nesting habitat.

Even though in a long-term, multi-partner study, from which the nest-site data emerged, it is difficult to keep protocols standardized, especially when technology changes and improves during the study, and data may not be of ideal quality, analysis of the data must be conducted regardless. Therefore it is worth considering the results of this study in context of the list of serious pitfalls that could affect the accuracy of predictive modeling and mapping with occurrence data described by Phillips et al. (2006). For example, all occurrence localities have some level of precision or error, can be biased by access conduits, sampling barriers, and variation in sampling effort over space and time. The choice of variables to use for building models directly affects the degree to which a model can be generalized to other areas and time periods. The set of modeling variables might be insufficient to describe all the parameters of a species fundamental niche relevant to its distribution at the grain of the modeling task. Large errors within the predictor variables will directly affect model accuracy. The results of this study would have naturally been different if the occurrence data with poor location information had been used or the vegetation layer encompassed an area of different size or shape. For example, there were 86 potential nest-site locations that were not used for model building because they fell outside of the boundary for the GIS layer for vegetation cover class. Nonetheless, these points could be used for further validating and refining the current model if the vegetation cover information with the same resolution were made available.

The Maxent modeling methodology provides a powerful analytical tool that is capable of predicting the potential distribution of ecological phenomena such as sage grouse nest-site habitat based on occurrence information collected from historical, georeferenced events. Predictions assume that the same or similar events will occur in the future within the same geographical extent as they have in the past. While these assumptions are not likely to be valid all the time for all species and locations given the inherent variability in biological phenomena and climatic factors, the Maxent models and distributions provide a means of modeling and mapping ecological events that provide important information for the future management and conservation of natural resources. Model predictions will typically be larger than a species' realized distribution because few species occupy all areas that satisfy their niche requirements (Phillips et al., 2006). Other drawbacks include: 1) it is not as mature a statistical method as Generalized Linear or Additive Modeling (GLM, GAM), and 2) the Maxent software presented by Phillips et al. (2006) would benefit from the implementation of the capability to calculate confidence intervals for individual probability estimates. The applicability of the Maximum Entropy principle to species distributions is supported by thermodynamic theories of ecological processes. The second law of thermodynamics specifies that in systems without outside influences, processes move in a direction that maximizes entropy (Schneider and Kay, 1994). Thus, in the absence of influences other than those considered in the model, the geographic distribution of a species will indeed tend toward the distribution of Maximum Entropy (Phillips et al., 2006).

4.1. *Management implications*

The type of vegetation was the most important predictor variable in the Maxent models of sage grouse nest-site locations within the extent of the analysis area. The PUTR subcategory had the highest exponent value followed by the ARTRVA and rock subcategories (Fig. 5). The result of the riparian subcategory having the most negative exponent value for nest-site location agrees with other research (Aldridge and Boyce, 2007) that reported riparian habitats as risky for chick survival. They found that more than half of the habitats identified as attractive nesting habitat were considered risky and considered them an "ecological trap." Therefore, caution must be applied before interpreting all vegetation subcategories that produced positive exponents as "source" habitats. The ARTRW subcategory also had a negative exponent value indicating possible avoidance of this vegetation class for nesting. These results, however, do not refute the conclusion reached by Holloran et al. (2005) that dense sagebrush stands with adequate herbaceous vegetation represent desirable sage grouse nesting habitat, but rather add to them the strong association of PUTR for the Hart Mountain population. A sustainable management plan should limit actions (i.e. prescribed fire, herbicides, overstocking of ungulates) that would reduce the ecological features with the highest Maxent exponent values in this study and promote the maintenance or restoration of these and other ecosystem features such as forbs and grasses recognized as important for egg production and chick survival (Gregg, 2006).

Acknowledgements

George Buckner and the Bureau of Land Management, Michael W. Byrne, Kreg A. Coggins, John A. Crawford, David W. Wrobleski, Travis Miller, Kevin Goldie, support staff at Hart Mountain National Antelope Refuge, and numerous field technicians.

REFERENCES

Aldridge, C.L., Brigham, R.M., 2002. Sage-grouse nesting and brood habitat use in southern Canada. Journal of Wildlife Management 66, 433–444.

Aldridge, C.L., Boyce, M.S., 2007. Linking occurrence and fitness to persistence: habitat-based approach for endangered greater sage-grouse. Ecological Applications 17 (2), 508–526.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodel Inference: A Practical–Theoretic Approach, 2nd ed. Springer-Verlag.

Byrne, M.W. 2002. Habitat use by female greater sage grouse in relation to fire at Hart Mountain National Antelope Refuge, Oregon. M.S. Thesis, Oregon State University, Corvallis, OR, USA.

Coggins, K.A. 1998. Relationship between habitat changes and productivity of sage grouse at Hart Mountain National Antelope Refuge, Oregon. M.S. Thesis, Oregon State University, Corvallis, OR, USA.

Deleo, J.M., 1993. Receiver operating characteristic laboratory (ROCLAB): software for developing decision strategies that account for uncertainty. Proceedings of the Second International Symposium on Uncertainty Modelling and analysis. IEEE, Computer Society Press, College Park, MD, pp. 318–325.

Elith, J., Graham, C.H., and the NCEAS Species Distribution Modelling Group, 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29 (2), 129–151.

Environmental Science Research Institute. 2002. ArcMap 8.2. Redlands, California, USA. ESRI. 1999–2002. ArcView GIS 3.3. Redlands, California, USA.

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

Gregg, M.A. 2006. Greater sage-grouse reproductive ecology: linkages among habitat resources, maternal nutrition, and chick survival. PhD Dissertation, Oregon State University. Corvallis, OR 97333, 217 pp.

Gregg, M.A., Crawford, J.A., Drut, M.S., DeLong, A.K., 1994. Vegetational cover and predation of sage grouse nests in Oregon. Journal of Wildlife Management 58, 162–166.

Guisan, A., Hofer, U., 2003. Predicting reptile distributions at the mesoscale: relation to climate and topography. Journal of Biogeography vol. 30 (8), 1233–1243.

Hernandez, P.A., Graham, C.H., Master, L.L., Albert, D.L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29, 773–785.

Hilborn, R., Mangel, M., 1997. The Ecological Detective: Confronting Models with Data. Princeton University Press, Princeton, New Jersey, USA.

Hirzel, A.H., Gwenaelle, L.L., Helfer, V., Randin, C., Guisan, A., 2006. Ecological Modelling 199, 142–152.

Holloran, M.M., Heath, B.J., Lyon, A.G., Slater, S.J., Kuipers, J.L., Anderson, S.H., 2005. Greater sage-grouse nesting habitat selection and success in Wyoming. Journal of Wildlife Management 69 (2), 638–649.

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

Iverson, L.R., Dale, M.E., Scott, C.T., Prasad, A., 1997. A GIS-derived integrated moisture index to predict forest composition and productivity of Ohio forests (U.S.A.). Landscape Ecology 12, 331–348.

Landis, R.J., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

McCune, B., 2006. Non-parametric habitat models with automatic interactions. Journal of Vegetation Science 17, 819–830.

Phillips, S.J., et al., 2004. A maximum entropy approach to species distribution modeling. In: Brodley, C.E. (Ed.), Machine Learning. Proc. of the Twenty-first Century International Conference on Machine Learning. ACM Press, Canada, p. 83.

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecollogical Modelling 190, 231–259.

Schneider, E., Kay, J., 1994. Life as a manifestation of the second law of thermodynamics. Mathematical and Computer Modelling 19 (6–8), 25–48.

Sveum, C.M., Edge, W.D., Crawford, J.A., 1998. Nesting habitat selection by sage grouse in south-central Washington. Journal of Range Management 51, 265–269.

Valverde, A.J., Lobo, J.M., 2007. Threshold criteria for conversion of probability of species presence to either–or presence–absence. Acta Oecologica 31 (3), 361–369.

Wiley, E.O., McNyset, K.M., Peterson, A.T., Robins, C.R., Stewart, A.M., 2003. Niche modeling and geographic range predictions in the marine environment using a machine-learning algorithm. Oceanography 16, 120–127.

Yost, A.C. 2008. Probabilistic modeling and mapping of plant indicator species in a Northeast Oregon industrial forest, USA. Ecological Indicators 8, 46–56.