Journal of Wildlife Management 74(5):1112–1121; 2010; DOI: 10.2193/2009-002

Tools and Technology Article

Ranking Mahalanobis Distance Models for Predictions of Occupancy From Presence-Only Data

SUZANNE C. GRIFFIN,[1] *Wildlife Biology Program, College of Forestry and Conservation, University of Montana, Missoula, MT 59812, USA*
MARK L. TAPER, *Department of Ecology, Lewis Hall, Montana State University, Bozeman, MT 59717, USA*
ROGER HOFFMAN, *Olympic National Park, Division of Natural Resources, 600 Park Avenue, Port Angeles, WA 98362, USA*
L. SCOTT MILLS, *Wildlife Biology Program, College of Forestry and Conservation, University of Montana, Missoula, MT 59812, USA*

ABSTRACT The Mahalanobis distance statistic (D^2) has emerged as an effective tool to identify suitable habitat from presence data alone, but there has been no mechanism to select among potential habitat covariates. We propose that the best combination of explanatory variables for a D^2 model can be identified by ranking potential models based on the proportion of the entire study area that is classified as potentially suitable habitat given that a predetermined proportion of occupied locations are correctly classified. In effect, our approach seeks to minimize errors of commission, or maximize specificity, while holding the omission error rate constant. We used this approach to identify potentially suitable habitat for the Olympic marmot (*Marmota olympus*), a declining species endemic to Olympic National Park, Washington, USA. We compared models built with all combinations of 11 habitat variables. A 7-variable model identified 21,143 ha within the park as potentially suitable for marmots, correctly classifying 80% of occupied locations. Additional refinements to the 7-variable model (e.g., eliminating small patches) further reduced the predicted area to 18,579 ha with little reduction in predictive power. Although we sought a model that would allow field workers to find 80% of Olympic marmot locations, in fact, <3% of 376 occupied locations and <9% of abandoned locations were >100 m from habitat predicted by the final model, suggesting that >90% of occupied marmot habitat could be found by observant workers surveying predicted habitat. The model comparison procedure allowed us to identify the suite of covariates that maximized specificity of our model and, thus, limited the amount of less favorable habitat included in the final prediction area. We expect that by maximizing specificity of models built from presence-only data, our model comparison procedure will be useful to conservation practitioners planning reintroductions, searching for rare species, or identifying habitat for protection.

KEY WORDS habitat model, Mahalanobis distance, *Marmota olympus*, Olympic marmot, Olympic National Park, presence-only data.

Predictive modeling of distribution of a species or of habitat suitable for a species is a key component of many conservation programs and ecological studies (e.g., Guisan and Zimmermann 2000, Fortin et al. 2005, Guisan and Thuiller 2005, MacKenzie et al. 2006). Although predictive models are often, and perhaps best, built using techniques such as logistic regression that rely on both presence and absence data (e.g., Peeters and Gardeniers 1998, Manel et al. 1999, Mladenoff et al. 1999), in many applications absence data are unavailable, unreliable, or incomplete. At the time of a survey, a species may have undergone declines for reasons unrelated to habitat quality (van Manen et al. 2005, Thompson et al. 2006). For instance, wide-ranging animals may be absent from an area (Clark et al. 1993, Pearce and Boyce 2006), a species that exists in a metapopulation may be temporarily extinct at a suitable site (Hanski 1998, Ozgul et al. 2006), and even organisms that are present at the time of a survey may not be detected (McArdle 1990, MacKenzie et al. 2003). To accommodate such situations, several methods have been developed to predict distribution or rank potential habitat using only presence data (e.g., Busby 1991, Clark et al. 1993, Hirzel et al. 2002, Lele and Keim 2006). These methods extend the toolkit of the habitat modeler.

The Mahalanobis distance (D^2) statistic has been successfully used to identify suitable habitat from presence data in a variety of situations (Corsi et al. 1999, Boetsch et al. 2003, Browning et al. 2005, Thompson et al. 2006) and in a recent comparison of several presence-only methods, this modeling approach performed particularly well (Tsoar et al. 2007). In this method, every map cell is assigned a score based on how similar it is to the multivariate mean of the habitat characteristics of the occupied map cells. In addition to requiring only presence data, D^2-based models do not require multivariate normality in the habitat data and they specifically account for covariance among habitat variables (Knick and Dyer 1997). Mahalanobis distance-based habitat models most commonly have been developed as a practical aid to conservation efforts (e.g., to identify potential reintroduction sites [Thatcher et al. 2006, Thompson et al. 2006], to guide surveys for rare plants [van Manen et al. 2005]).

A limitation of the D^2 method is that there are no significance tests or other established methods to determine importance of the explanatory covariates (Johnson and Gillingham 2005). Whereas other habitat modeling approaches (e.g., logistic regression) allow stepwise inclusion and exclusion of covariates, or permit comparison of how well several competing models fit the data at hand, we are unaware of any metrics that evaluate effects of individual covariates on the specificity of the D^2 statistic. Including even one extraneous or redundant covariate greatly increases the number of parameters that must be estimated in the covariance matrix and so reduces precision of each estimate,

[1] *E-mail: olympicmarmots@aol.com*

lessening reliability of the final model. Furthermore, the D^2 statistic only identifies how dissimilar any given location is to the average occupied location with respect to all the habitat covariates. Thus, inclusion of covariates that do not differ in distribution between occupied locations and the rest of the study area will dilute distinctions between the 2 types of locations by inflating the within-group variance relative to the between-group variance. Thus, it is possible that when there are many candidate habitat covariates, inclusion of extraneous or redundant covariates may actually reduce specificity of the model.

Several recent papers have argued that it is possible to identify variables that are most constant and, therefore, useful predictors of habitat by partitioning the variance in the D^2 into the principle components of the correlation matrix (Dunn and Duncan 2000, Browning et al. 2005, Rotenberry et al. 2006). The variables that contribute most heavily to the eigenvectors with the smallest eigenvalues are considered the important predictors of habitat. There are some apparent weaknesses in this decomposition approach. At a fundamental level, these eigenvectors are more poorly estimated than the eigenvectors associated with the larger eigenvalues. Browning et al. (2005) attempted to account for this weakness by bootstrapping their data and eliminating eigenvectors that appeared to be unstable. Second, small variance components will arise when ≥ 2 variables are highly correlated, regardless of their importance to the organism (S. Cherry, Montana State University, unpublished data). Even if these statistical factors were unimportant in a given situation (i.e., a very large data set with little or no correlation in the explanatory variables), the use of only the smallest k principle components could result in the prediction of considerably more habitat than would be predicted by a full model (Rotenberry et al. 2006, figs. 1, 2). Ignoring these limitations, the decomposition approach, which Rotenberry et al. (2006) refer to as niche identification, may contribute to theoretical understanding of a species' biology and may be a useful tool in predicting a species' general response to changing ecological conditions. However, the primary objective of many modeling efforts is to make the model as specific as possible by reducing the amount of unsuitable habitat predicted to be suitable while simultaneously classifying a large portion of the suitable habitat as suitable.

If high model specificity is a desired outcome, we propose that the best combination of explanatory variables for a D^2 model can be identified by ranking potential models based on the proportion of the entire study area that is classified as potentially suitable habitat given that a predetermined proportion of occupied locations are correctly classified. Because performance of any given model may depend on the particular sample of presence points, we recommend averaging results from many bootstrap replicates from the presence data.

We developed a habitat model for the Olympic marmot (*Marmota olympus*) using this model-ranking approach. Olympic marmots are large, ground-dwelling squirrels found on the upper slopes of the Olympic Mountains in northwest Washington State, USA (Fig. 1). Their range is largely contained within the 3,700-km^2 Olympic National Park. Since 1999, the park's Resource Management Plan has called for determining present distribution of marmots within the park and developing a long-term monitoring program for the species. Olympic and other alpine-dwelling marmots inhabit high-elevation meadows, often interspersed with talus or rock outcrops, on moderately steep, southeast- to southwest-facing slopes (Barash 1989, Armitage 2000). Marmots dig extensive burrow systems and consequently require well-developed soil. Marmots are likely restricted to high elevations both by distribution of meadows and by these species' intolerance of high temperatures (Türk and Arnold 1988, Melcher et al. 1990). These habitat requirements are generalities; Olympic marmots occupy all aspects and a wide range of slopes, and they and their burrows are often found at forest edges.

This semifossorial rodent shares much with many species of conservation concern. Little was known about its distribution when we began our study; like other marmots (Ozgul et al. 2006), Olympic marmots are believed to persist as a metapopulation with periodic local extinctions leaving suitable habitat temporarily vacant, and Olympic marmots were declining and recently had disappeared from many apparently suitable habitat patches (Griffin et al. 2008). Additionally, our survey data more readily lent itself to a presence-only model than a presence–absence model. Our objective was to compare the ability of each of $2^n - 1$ possible Mahalanobis distance models, given n habitat covariates measured at locations where marmots were present, to identify accurately a predetermined proportion of suitable habitat while minimizing the amount of the total landscape predicted.

STUDY AREA

We initially restricted our study to Olympic National Park plus a 1.5-km buffer due to limited Geographic Information System (GIS) cover-type data. Practically, our study area included almost the entire range of the Olympic marmot. We imposed an elevation cut-off of 1,300 m (Fig. 1), which was lower than the lowest known marmot colony. Finally, we removed all map cells classified as 71–100% closed canopy in the GIS cover-type layer, because marmots were not found in closed canopy forest. Thus, our final study area encompassed 78,302 ha of open or lightly wooded, high-elevation terrain within or adjacent to the park.

METHODS

The Mahalanobis Distance Statistic

The Mahalanobis distance statistic (D^2) represents the standardized squared distance between the covariate values for a given sample and the mean vector of these covariates for the occupied locations used to build the model (hereafter, training data). In the context of habitat modeling, a D^2 value is computed for each map cell in the study area based on the value of the habitat covariates under consideration in that cell, relative to the average values of those covariates in the training data as follows:

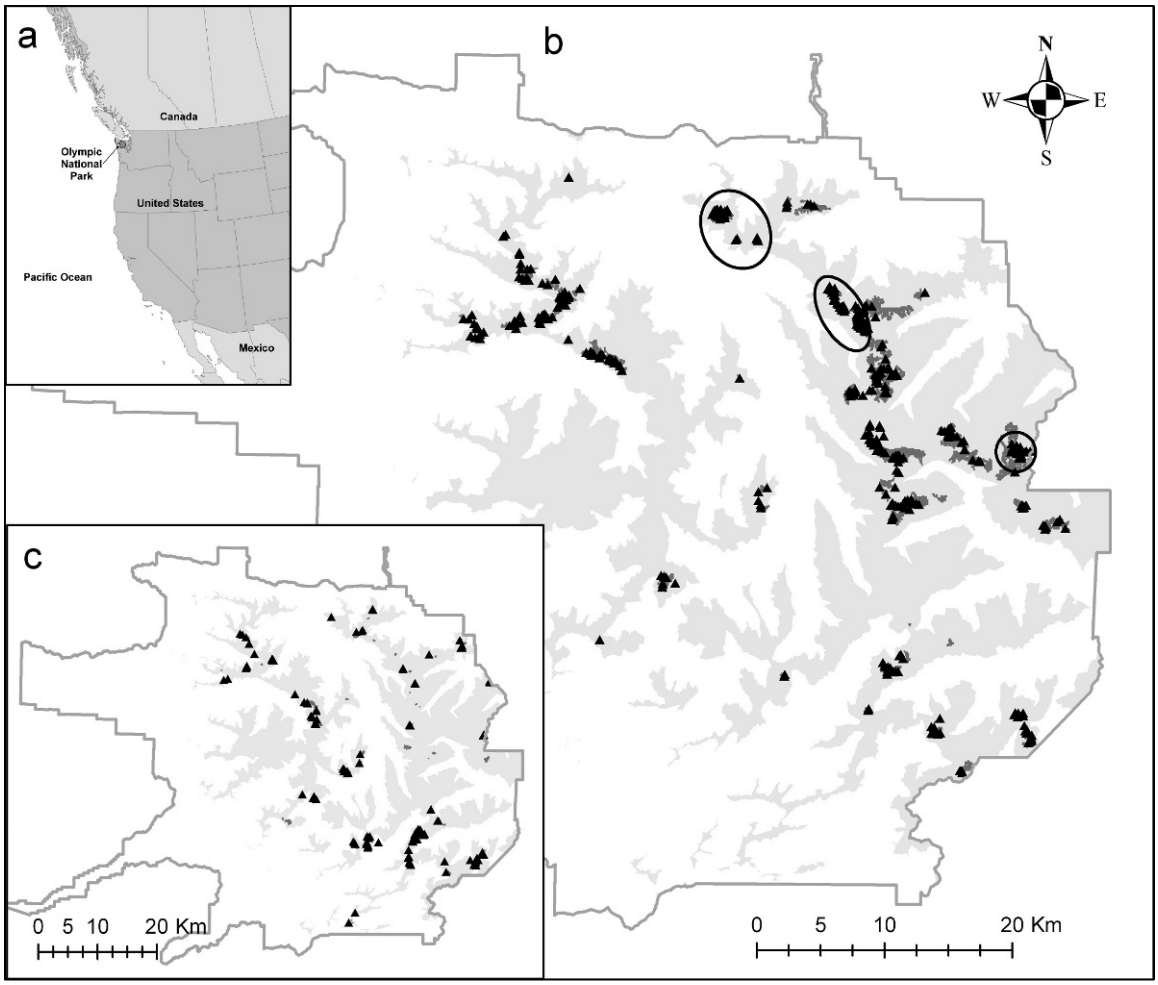

Figure 1. (a) Location of Olympic National Park, Washington, USA. (b) Polygons of habitat known to be occupied by Olympic marmots in 2002–2005 (dark gray shading) and 376 point locations used in development of habitat models for the species (black triangles); intensive study sites are circled. (c) Polygons known to be abandoned (dark gray shading) and the 114 abandoned point locations (black triangles) used to test the habitat model. Areas within the park >1,300 m elevation are shown with light gray shading in (b) and (c).

$$D^2 = (\hat{\underline{\mu}} - \underline{x})' \hat{\Sigma}^{-1} (\hat{\underline{\mu}} - \underline{x}),$$

where $\hat{\underline{\mu}}$ and $\hat{\Sigma}$ are, for the habitat covariates under consideration, the vector of the mean values and the variance–covariance matrix at presence locations, respectively. The variable $\underline{x}$ is the vector of values for each habitat variable for a given cell. Cells with smaller D^2 values have habitat values more similar to the average of the training data and so should be more likely to be occupied. The D^2 values are continuous with a minimum of zero. If the training data meet the assumption of multivariate normality, then the D^2 values are chi-square distributed and can be rescaled to probabilities. Even when this assumption is violated, there is a monotonic relationship between the D^2 values and dissimilarity from the mean, with equal scores being equally distant from the mean in multivariate space. Thus, D^2 values rank habitat in terms of suitability rather than providing a probability of occupancy for each map cell. Follow-up surveys guided by model predictions can provide estimates of probability of occupancy (Boetsch et al. 2003).

For defining suitable habitat, a threshold D^2 value is usually identified. Map cells with D^2 values lower than that threshold are considered suitable for the study organism and the remaining cells are considered unsuitable (Thatcher et al. 2006). The threshold may be set so that all occupied points are classified as being within suitable habitat or such that some lesser proportion of the occupied locations are classified as suitable (Podruzny et al. 2002, Boetsch et al. 2003, van Manen et al. 2005, Thatcher et al. 2006, Thompson et al. 2006). When the proportion of occupied map cells with D^2 values below the threshold is much greater than the proportion of random map cells with D^2 values below that same value, or when distribution of D^2 scores of occupied test locations is similar to those of training data, models are considered to perform well (Boetsch et al. 2003, Browning et al. 2005, van Manen et al. 2005).

Comparison of Mahalanobis Distance Models

Computing D^2 values for several models for every 25-m × 25-m cell in an entire landscape would be cumbersome, but computing D^2 values of a few locations is easily done using Matlab (The Mathworks, Inc., Natick, MA) or similar software. Therefore, we randomly selected 1,000 map cells

from within the study area to represent the total landscape in our comparisons. We used presence locations, or bootstrapped samples from these presence locations, to represent suitable habitat.

As a metric with which to rank the models, we determined the proportion of random locations classified as suitable habitat under each of the models given that 80% of occupied locations were classified as suitable (we term this metric P_{r80}). That is, for each model we identified the smallest D^2 value such that 80% of training data (in this case, a bootstrap sample of presence locations) fell in map cells with values less than or equal to that value. We classified all map cells with equivalent or smaller D^2 values as suitable and cells with larger D^2 values as unsuitable. We could then rank models based on P_{r80}, with lower values indicating superior models. We were, in effect, seeking to minimize the error of commission rate, or maximize specificity, while holding the omission error rate constant at 20%. The 80% threshold for defining suitable habitat was somewhat arbitrary and the optimal point of comparison will vary depending on the purpose of the model (e.g., van Manen et al. 2005, Thatcher et al. 2006).

Because $\underline{\hat{\mu}}$, $\hat{\Sigma}$, and ultimately P_{r80} for a given model depend on the particular sample of presence points, we compared the mean of P_{r80} from 2,500 bootstrap replicates of size 376 from our presence data rather than relying on the results of one sample. We calculated P_{r80} for each replicate under each of the $2^n - 1$ models based on $\underline{\hat{\mu}}$ and $\hat{\Sigma}$ for that bootstrap sample. We ranked all models based on the mean values ($\bar{P}_{r80}$).

To determine how much the models varied in which of the 1,000 randomly selected map cells they predicted to be suitable, we calculated the Sorenson's similarity coefficient, Ssc (Sorenson 1948), between the set of points selected by the top model and those from each of the other models. The statistic Ssc measures the degree of overlap in 2 groups, with a value of one indicating total overlap (e.g., all points classified identically by the 2 models) and a zero indicating that there is no overlap in classification. We wrote Matlab code to calculate the D^2 scores and $\bar{P}_{r80}$ from bootstrapped samples and to calculate the Ssc values (Appendix SA, <www.wildlifejournals.org>).

For the highest ranked model based on $\bar{P}_{r80}$ from the bootstraps, we built one model from all presence locations. With this model, we computed D^2 scores for the 1,000 random points and visually compared the cumulative frequency curve of these scores to that from the scores of occupied data. We also computed and plotted D^2 scores for 114 abandoned locations (see Sampling Abandoned Habitat below). Similarity between abandoned and occupied data would provide additional support for the model, although differences in occupied and abandoned locations could indicate differences in the 2 types of sites, rather than a poorly fitting or over-fit model. Finally, we randomly selected 1,000 locations from the unoccupied survey polygons and compared the cumulative frequency curves of these to occupied locations. Although we only drew these unoccupied locations from a subset of the study area, they do provide an interesting comparison because they represent open, high-elevation habitat that did not show any sign of recent marmot use. If the Mahalanobis distance model represented a substantial improvement over the preliminary model upon which we based our surveys, a large proportion of these unoccupied points should be classified as unsuitable.

Collection and Subsampling of Location Data

We collected location data across the marmots' range during the course of habitat surveys and other activities. From 2002 to 2005, we surveyed 811 polygons of possible marmot habitat throughout Olympic National Park. We had identified these polygons in a preliminary GIS model (hereafter, the 2002 model) and polygon selection and survey protocol are described in detail elsewhere (Griffin et al. 2008). Briefly, we subdivided by aspect patches of meadow or rock (a cover class that included bare ground) >1,400 m elevation. We removed patches <0.56 ha because preliminary surveys suggested that smaller patches would not support marmots. We surveyed on foot groups of 1–5 of resulting polygons according to a stratified random sampling design. We classified polygons as occupied, abandoned, or without sign of marmots (no sign) based on presence of marmots, active or inactive burrows, and other evidence.

The polygons provided an effective approach to sampling the habitat but because marmots occupied only a small portion of many occupied polygons, they were unsuitable as a sampling unit in a presence–absence model. However, in most occupied and abandoned polygons we did collect representative locations of marmots or burrows (active or abandoned), using a handheld Global Positioning System (GPS) unit (usually accurate to $\leq$10 m). Hereafter, occupied location indicates these recorded locations of occupied burrows or marmots and abandoned location indicates recorded locations of abandoned burrows. That is, location refers to a point on the ground rather than a marmot colony or a meadow and provided precise, reliable data suitable for use in a presence-only model. In addition to recording locations during polygon surveys, we recorded opportunistically encountered marmots or burrows outside polygons; we conducted trapping, resighting, and radiotelemetry studies in 3 areas of the park (hereafter, demographic sites; Griffin et al. 2007, 2008); we collected hair samples for genetic analyses from marmots throughout the park; and we investigated several written and oral reports of marmots (Griffin 2007). During each of these activities, we collected additional location data. We included in the occupied data set locations from 3 colonies that were known to have been abandoned during the study period.

We collected >10,000 occupied locations, each of which was not necessarily unique; in the course of radiotracking, trapping, and genetic sampling, burrows were recorded multiple times and individuals were often represented by multiple locations within their home ranges. To reduce the influence of the more heavily sampled areas, we subsampled from the data as follows: we randomly ordered locations and then sequentially compared each on the list to all prior points. We accepted locations >125 m from all previously

Table 1. Names and descriptions of habitat covariates we used in developing habitat models for Olympic marmots. Value ranges for covariates are based on marmot locations collected in 2002–2005 in Olympic National Park, Washington, USA.

Variable	Description	Classes or value range	Window
Elevation[a]	Elevation (m)	1,300–2,430	Focal map cell
Slope[b]	Slope steepness (°)	0–89	Focal map cell
Rock[c]	No. of rock or sparse ground map cells within window	0–25	25 map cells
Meadow[c]	No. of meadow map cells within window	0–25	25 map cells
May insolation[d]	Modeled incoming daily solar radiation for 21 May (10^6 Kj/m^2)	23.0–41.1	25 map cells
Aspect NE/SW[b]		0/1	Focal map cell
Aspect NW/SE[b]		0/1	Focal map cell
Tree[c]	Trees present within focal map cell	0/1	Focal map cell
SD of slope[d]	Measure of topographic variability (SD)	0.7–21.0	25 map cells
Planiform curvature[b]	Slope curvature in horizontal plane (unitless)	−24.6–30.8	25 map cells
Profile curvature[b]	Slope curvature in vertical plane (unitless)	−20.6–18.4	25 map cells

[a] U.S. Geological Survey (2000).
[b] We used standard ArcGIS tools to derive these variables from the elevation layer.
[c] Pacific Meridian Resources (1996).
[d] Hetrick et al. (1993).

accepted locations into the data set we used to build the models; we discarded all other locations. We also removed 8 locations that fell within cells classified as 71–100% canopy closure because these were likely the result of GPS, recording, or map classification error, and these locations fell outside the defined study area. Following subsampling, the marmot presence data set was composed of 376 occupied points (Fig. 1b).

We chose 125 m as the buffer distance around each location because the 4.9 ha thus encompassed approximates the median minimum convex polygon (MCP) home range of Olympic marmots (Griffin 2007). However, due to concerns that the demographic study sites might be overrepresented despite subsampling, we explored the effect of a larger buffer size (200 m) on distributions of the habitat covariates used to build the model. As expected, subsampling with a 200-m buffer reduced the final data set by 31%, to 258 cases. The proportion of the included locations that represented the demographic study sites decreased from 23% to 16%. However, neither the mean nor variance of any of the 11 habitat covariates (see Habitat Covariates) changed in a statistically or biologically significant manner (t-tests, F-tests, and χ^2 tests of association as appropriate, all $P > 0.05$). We expected that the distributions of the habitat covariates would be similar at the 2 scales because most (77%) locations in the 125-m buffer data set were not from the demographic sites and 165 of 219 (75%) known occupied polygons were represented by ≥ 1 locations. There was no indication that the 165 represented polygons differed from the 54 that were occupied but not represented with respect to region, aspect, area class, or slope class (χ^2 tests, all $P > 0.05$). Furthermore, the demographic sites collectively represented all aspects and a range of elevations, colony sizes, and vegetation types (Griffin et al. 2008). Thus, the 125-m buffer appeared to have adequately reduced the influence of the demographic study sites and we used that data set in all further analyses.

We restricted the data set of abandoned locations to burrows that showed no signs of recent use and that were located in areas that contained many such burrows or where there were historical records of marmots. We did not include burrows that were <200 m from occupied habitat and any others that we felt might not represent truly abandoned habitat, initially leaving 175 abandoned burrow locations. We subsampled from these as for occupied locations. The 114 points (Fig. 1c) ultimately composing the abandoned data set represented 78 of the 111 (70%) of known abandoned polygons.

Habitat Covariates

Habitat models included up to 11 explanatory variables that described topography and vegetation within either the focal 25-m × 25-m map cell or a 25-cell window centered on the focal cell (Table 1), depending on the scale at which we thought variables might be important. For example, by considering the proportion of cells that were meadow or rock within a 25-cell window rather than individual cells, we intended to allow inclusion of small sparsely vegetated or rocky areas if appropriate while penalizing those cells that were surrounded by extensive rock or scree with no forage or digging substrate available. Use of the larger moving window also served to reduce the influence of classification error in the GIS layer. Continuous variables such as elevation and aspect should be adequately characterized by the value of one map cell. We attempted to use Beers' transformations (Beers et al. 1966) to linearize the aspect covariate but the result was a bimodal distribution. A bimodal distribution is not desirable for the Mahalanobis distance calculations because locations at the modes are far from the mean and thus penalized, although the modes represent the most typical locations for marmots. Instead, we used 2 binary variables, NE/SW and NW/SE, to describe aspect.

Although we identified 5 types of meadow in the GIS cover-type layer and marmots probably prefer some types over others, we lumped all meadow types into one classification meadow because the rarity of each meadow type resulted in the mean proportion of the 25-cell window occupied by each type being small (<0.10) at the 376 occupied locations. Thus, locations with a high proportion of any given meadow type would have been penalized because they were unusual (the best habitat is rare in this

case and so not represented by the average of the occupied locations). The same problem would have resulted if we used one cell instead of the 25-cell window.

Final Model Refinements and Evaluation

We used the best performing model to produce a final habitat map for Olympic marmots. We used the $\hat{\underline{\mu}}$ and $\hat{\Sigma}^{-1}$ from all 376 occupied locations to compute D^2 values for each map cell in the entire landscape. For this purpose, we wrote an Arc Macro Language code for use in ARC/INFO GIS. We identified the D^2-value at which 80% of the 376 occupied locations were correctly classified and used this as the upper limit for designating habitat as suitable for Olympic marmots.

We then made several refinements to facilitate comparison with the 2002 habitat model and to make the final model more useful for Olympic Park monitoring and management efforts. First, we eliminated several patches of permanent snow and ice that were predicted, because these are clearly inhospitable to marmots. Second, as in the 2002 model, we removed patches of predicted habitat <0.56 ha, assuming them to be too small to support marmots (the min. recorded MCP home range of an adult Olympic marmot is 0.60 ha [n = 33]; Griffin 2007). Finally, we eliminated areas outside the park boundaries because the 2002 model did not include these areas and monitoring by park personnel would be confined to the park.

We then determined the distance that each occupied and each abandoned location fell from predicted habitat in this refined model, to confirm that removed patches were not important marmot habitat, which also allowed us to determine whether points falling outside the predicted area were at least close to these areas. We also examined the amount of rock predicted by our new model, because one objective of the modeling process was to eliminate some of the 20,454 ha of often inhospitable rock identified as habitat by the 2002 model.

RESULTS

Model Comparison

We compared $\bar{P}_{r80}$, the mean proportion of random locations predicted to be suitable habitat, from 2,500 bootstrap replicates for 2,047 models representing all possible combinations of 11 habitat covariates. The best performing model included 7 covariates and predicted 29.2% of random locations to be suitable. Of the 2,500 bootstrap replicates, this 7-variable model was top ranked in 624 (25%). For all other variable combinations, $\bar{P}_{r80}$ was >30% (Fig. 2) and the full model ranked 8 with a $\bar{P}_{r80}$ of 30.8%. Bootstrap replicates demonstrated considerable variability in the proportion of random locations classified as suitable due to sampling. The fifth and 95th percentiles of this distribution for the top ranked model were 25.3% and 33.6%, respectively. This range of variability was less than that of 92.9% of the 2,047 models.

Several covariates appeared at high frequency in the top 100 models (Table 2). Elevation proved to be the most critical, appearing in all of the top 459 models. Rock, meadow, and May insolation each appeared in >90 of the top 100 models. These variables were followed in apparent importance by aspect NE/SW, aspect NW/SE, and profile curvature. Slope, planiform curvature, trees, and standard deviation of slope all appeared in <55 of the top ranked 100 models. The top ranked model included only the 7 most frequently occurring covariates.

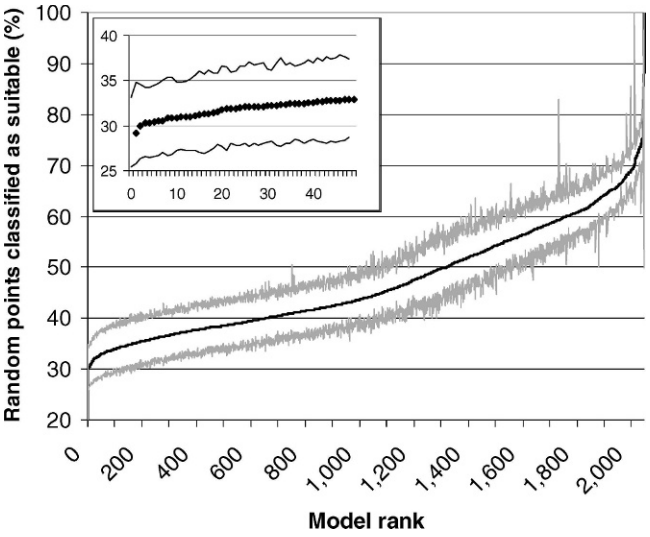

Figure 2. Mean percentage of locations randomly selected from Olympic National Park, Washington, USA, study area that were classified as suitable for Olympic marmots (with the fifth and 95th percentiles of 2,500 bootstrap samples) by each of the 2,047 models, ordered in descending order of performance. The results for the top 50 models are shown in the inset. We developed models from Olympic marmot location data collected in 2002–2005.

A model's Ssc relative to the best model generally decreased with model rank (Fig. 3), but there was considerable variation in scores even among models that predicted almost the same proportion of random points to be suitable. This variation in Ssc scores indicates that although the highest ranked models all predicted about 30% of random points to be suitable habitat, not all of these highly ranked models predicted the same map cells as those predicted by the top-ranked model. In particular, it appeared that inclusion of trees as a predictor variable led to greater dissimilarity in configuration of selected habitat (Fig. 3).

The cumulative frequency curve for occupied locations showed a high degree of overlap with abandoned locations (Fig. 4). The cumulative frequency curve for points selected from unoccupied polygons lay considerably below that of occupied and abandoned locations and only 37% of these had D^2 values less than the threshold value of 8.54.

Final Model Refinements and Evaluation

The unrefined final model, using 7 habitat covariates and all 376 occupied locations, identified 22,624 ha (28.9% of the entire study area) as containing 80% of suitable habitat (D^2 < 8.54). Within the park itself, 21,143 ha had D^2 values ≤8.54; when we removed snowfields and polygons <0.56 ha, the predicted area was further reduced to 18,579 ha (Fig. 5). This predicted area may be compared

Table 2. Means and standard deviations of habitat covariates at 376 locations occupied by Olympic marmots, 2002–2005, and at 1,000 random points within the Olympic National Park, Washington, USA, study area, number of times each covariate appeared in the highest ranking 100 models, and whether the covariate was included in the final habitat model for Olympic marmots.

Covariate	Occupied locations ($\bar{x}$)	Occupied locations SD	Proportion of occupied locations	Random locations ($\bar{x}$)	Random locations SD	Proportion of random locations	No. times included in top 20 models	No. times included in top 100 models	Included in top-ranked model?
Elevation (m)	1,680	119		1,594	194		20	100	Yes
Slope (°)	25.37	11.18		30.55	11.71		8	53	No
Rock	6	7		8	9		20	93	Yes
Meadow	10	8		4	6		20	99	Yes
May insolation	37.32	3.06		35.64	4.39		20	91	Yes
Aspect NE/SW (proportion NE)			0.37			0.48	19	65	Yes
Aspect NW/SE (proportion SE)			0.56			0.57	20	78	Yes
Tree (proportion with trees)			0.18			0.33	8	50	No
SD of slope	5.60	2.90		5.75	2.94		9	44	No
Planiform curvature	−0.82	6.10		−0.03	7.67		7	38	No
Profile curvature	0.88	10.14		−0.37	12.63		15	62	Yes

to our 2002 model, which was similarly restricted to patches >0.56 ha within the park. The 2002 model predicted an area of 28,275 ha.

Elimination of the small patches did little to reduce specificity of the model. Of the 376 occupied locations, 77.7% fell within predicted habitat. An additional 19.6% of the 376 occupied locations were ≤100 m from the predicted area and so would likely be detected during a survey of the area. Of abandoned locations, 71.9% fell within predicted habitat and 21.1% were ≤100 m from predicted habitat. We consider 100 m to be the approximate maximum distance at which active marmots or large burrows would be visible to an attentive surveyor from the edge of the predicted area; in fact, that distance will vary according to topographic and vegetative features in an area. Only 4,653 ha of rock were classified as suitable for marmots in our new model, as compared to 20,454 ha identified as potential habitat in the original model. However, the reduction in over-prediction of rock came with a cost; 6,575 ha of trees were predicted as suitable in the final model, whereas map cells with trees were excluded from the 2002 model.

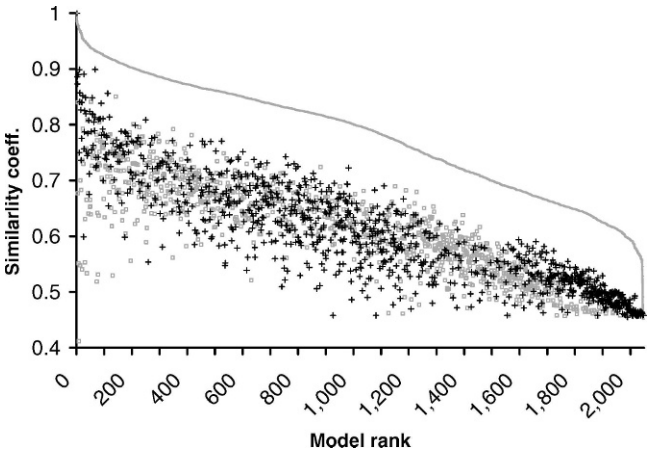

Figure 3. Sorenson's similarity coefficient relative to the top-ranked model for each of 2,047 models of Olympic marmot habitat in Olympic National Park, Washington, USA. Models that included trees as a covariate are shown with black crosses and models that did not include trees are shown with gray squares. The gray line indicates the maximum possible value of the similarity coefficient for each model, given the proportion of the random points classified as suitable by the top model and the model under consideration. We developed models from Olympic marmot location data collected in 2002–2005.

DISCUSSION

Mahalanobis distance-based models are useful for identifying suitable habitat but without a formal approach to variable selection, one had to make decisions about what variables to include in a final model based only on expert opinion or by experimenting with different variable combinations. Because inclusion of redundant or uninformative variables may reduce specificity of the model, the lack of a method for identifying useful covariates represented an important limitation on the use of the Mahalanobis distance to model habitat. Partitioning variance in the data set has been proposed as a way to identify important variables (Dunn and Duncan 2000, Browning et al. 2005,

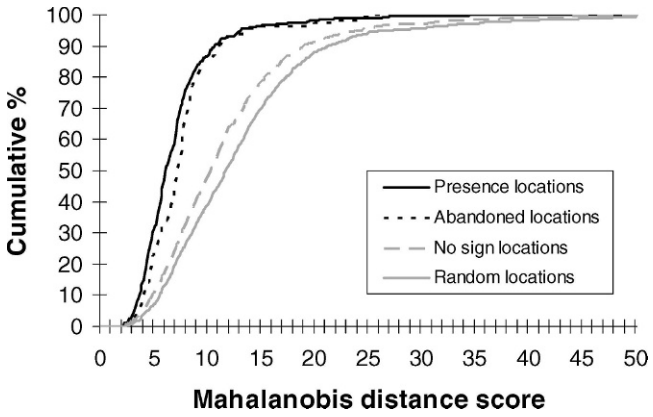

Figure 4. Cumulative frequency curves of Mahalanobis distance values from the highest ranking of 2,047 models of Olympic marmot habitat in Olympic National Park, Washington, USA, for 4 data sets. We developed models from Olympic marmot location data collected in 2002–2005.

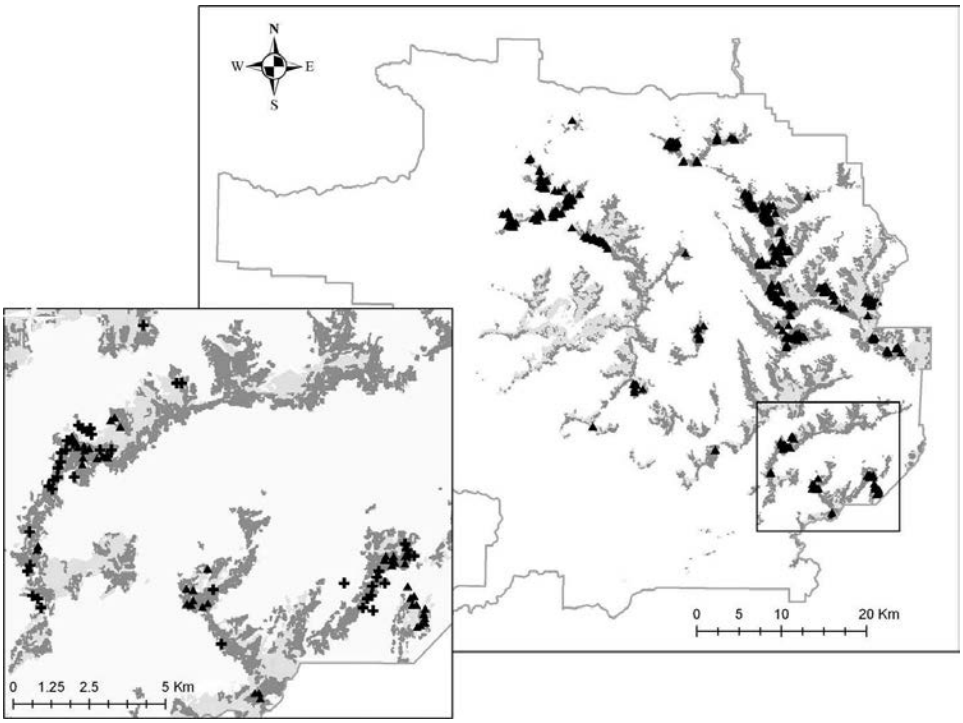

Figure 5. Predicted distribution of suitable habitat for Olympic marmots (dark gray) within Olympic National Park, Washington, USA, based on the top ranked, 7-variable model, clipped to the extent of the park, from which we removed snowfields and patches <0.56 ha. Also shown are additional areas predicted by the 2002 model that were not predicted by the current model (light gray) and the 376 occupied locations used to build the model (black triangles). A number of the abandoned locations used in evaluating the fit of the model are also shown in the inset (black crosses); >90% of both abandoned and occupied locations fell within 100 m of the predicted habitat. Occupied and abandoned locations were collected in 2002–2005.

Rotenberry et al. 2006), but has statistical and practical weaknesses when used to model habitat.

Our model comparison procedure allowed us to overcome a major limitation of the D^2 statistic as a means of modeling habitat, the lack of robust and practical variable selection procedure. For our data set, the comparison process indicated that to obtain the highest specificity (that is, minimize $\bar{P}_{r80}$) in identifying habitat, only 7 of the original 11 habitat covariates should be used. Only a handful of models approached the top model in specificity and most were similar in structure to the top model, indicating that the comparison procedure had identified the most useful variables. Inclusion of trees in several of the top models did represent a deviation from this pattern. In particular, as indicated by low Ssc values, models including trees tended to predict somewhat different habitat configurations from models without trees. The highest ranked model outperformed the others (including those with trees) in multiple metrics. Despite the many models we compared, there was no indication that the top-ranked model over-fit the data. The cumulative frequency curve of the abandoned locations was almost identical to that of the training data even though these abandoned locations included many mountains or ridge systems not represented in the occupied data set and were generally further south and west than occupied locations.

The model comparison procedure was able to identify a model that we would not have identified otherwise. Covariance among explanatory variables and, thus, the amount of additional information each variable brought to the model, was considered much more efficiently than casual inspection could do. Whereas several variables were obvious candidates for inclusion—for example meadow, on which marmots have a well-established and obvious dependence (Barash 1989, Armitage 2000), or May insolation, for which there were large differences between means and variances of these variables at occupied and random locations (Table 2)—inclusion or exclusion of others was less obvious. Based only on observed differences in distributions of occupied and random locations, we would probably have included trees but not aspect NW/SE. However, the model comparison procedure identified trees as less informative than aspect NW/SE, which was included in the top model and 78 of the top 100 models.

Some may object that comparison of all possible models is data-dredging. However, the primary objective of our habitat modeling effort, like that of many others, was not to test hypotheses but to produce the best possible predictive model. The more relevant concern was that the resulting best-model would over-fit the data and would have little predictive power when applied to new data. In fact, there was considerable variability in which model was top-ranked among the different bootstrapped sample replicates, confirming that results based on one data set could be idiosyncratic. However, the highest ranked model was the best fit for 25% of sample replicates and transferred well to the abandoned location data, indicating that we had surmounted the over-fitting problem.

Finally, it is important to recognize that although computation of the D^2 statistic for each cell is independent of choice of study area, our model comparison metric was not. To produce the most specific model possible, we advocate defining the study area as narrowly as can be biologically justified; thus, we placed a lower limit on elevation and removed heavily treed map cells from consideration. If our study area had been the entire park, it is probable that the top-ranked model would have included lack of trees as a predictor, because most of the park was forested. Similarly, without a priori removal of map cells with closed canopy from the study area, it is probable that trees would have been included in all the top models. Habitat identified as suitable by a given model would not have changed but the models would have ranked differently.

Olympic Marmots

By applying our approach to Olympic marmots, a declining endemic species of concern, we confronted the method with problems typical of sampling habitat for small and declining species. The final habitat model successfully identified suitable habitat for the Olympic marmot, based on quantitative and qualitative criteria. The model accurately identified currently occupied locations used in building and testing the model and also identified as suitable most abandoned areas of which we were aware and several unsurveyed areas where backcountry travelers have since reported marmots or suitable habitat. Although by construction only 80% of the 376 occupied marmot locations used to build the final model fell within the predicted habitat, >96% of occupied locations and >90% of abandoned locations were <100 m from predicted habitat and, thus, were likely to be found by observant workers surveying predicted habitat. The variables included in the best model are biologically meaningful and several have been previously suggested as being important to marmots.

The final model did predict lightly forested areas more often than we would have preferred, although marmots are found close to or in lightly treed areas on occasion. In the original 10,000 locations, there were undoubtedly poor GPS fixes and transcription errors, as well as marmots that were in atypical habitat. Although we believe that errors were rare and we removed the most obvious ones, the subsampling process by which we reduced our initial database of presence locations likely increased the frequency of erroneous locations in the data set. Errors that resulted in presence data being located in inappropriate habitat were likely to result in those locations also being >125 m from other occupied locations; thus, these erroneous and misleading points would have been retained during subsampling. In fact, increasing the subsampling buffer from 125 m to 200 m increased the frequency of these treed map cells from 17.8% to 19.1%. We considered using only hibernacula locations (Borgo 2003), or hibernacula and natal burrows, but our data for these were restricted to the northeast and sample sizes were limited. We also considered removing all locations in map cells classified as treed but we knew many of these to be valid.

Because D^2 models emphasize the most commonly occupied habitat type, rare but preferred types may not be identified as highly suitable. If currently rare but preferred types are thought to be important components of potential changes in species distributions resulting from habitat improvement projects, climate change, or human-mediated introductions, D^2 models may need to be supplemented (Knick and Rotenberry 1998). For example, preference-based models (e.g., Lele and Keim 2006) may be useful in these conditions, although these models remain considerably more difficult to implement.

MANAGEMENT IMPLICATIONS

A critical step in any modeling effort is to select analytical techniques appropriate for both the question at hand and the available data. In situations where absence data are unavailable, presence-only approaches such as the Mahalanobis distance may be the obvious choice. Even when presence–absence data are available, we recommend that these data be carefully evaluated before they are used to model habitat for species of conservation concern, because absence data may not be a reliable indicator of habitat nonsuitability. If a presence-only method is justified, we recommend the Mahalanobis distance approach because it is both effective and intuitive. In cases where biology of the species is less than perfectly described, as is often the case for species of conservation concern, our model comparison method can be used to identify the most useful set of variables to include in habitat models and may shed light on the species biology as well. We also envision situations in which a set of habitat covariates known to be important are included in all models and our approach is used to select among those covariates whose importance is less clear, which would limit the number of models compared, although top models should always include the most important covariates regardless of how many extraneous covariates are tested. In the common situation where presence-only data are available for conservation decision-making, our method will identify the model from that set that best minimizes over-prediction while still accurately identifying suitable habitat.

ACKNOWLEDGMENTS

National Science Foundation (NSF; DEB-0415604, DEB-0415932), The Canon National Parks Science Scholars Program, Mazama, Northwest Scientific Association, The American Society of Mammalogists, and the American Museum of Natural History provided funding. S. C. Griffin was supported by a United States Environmental Protection Agency Graduate Student Fellowship, an NSF Graduate Student Fellowship, a Budweiser Conservation Scholarship from the Anheuser-Busch Corporation and the National Fish and Wildlife Foundation, and the University of Montana College of Forestry and Conservation. We thank J. Boetsch for introducing S. C. Griffin to the D^2 method of modeling habitat; S. Cherry for useful discussions regarding D^2 habitat ranking and for generously providing his unpublished manuscript; P. Griffin, D. Naugle, D. Van

Vuren, D. Euler, and 2 anonymous reviewers for useful comments on the manuscript; Olympic National Park for providing vehicles and logistical support; and S. Pagacz, K. Sterling, and others for help in the field.

LITERATURE CITED

Armitage, K. B. 2000. The evolution, ecology, and systematics of marmots. Oecologia Montana 9:1–18.

Barash, D. P. 1989. Marmots. Social behavior and ecology. Stanford University Press, Stanford, California, USA.

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

Boetsch, J., F. Van Manen, and J. Clark. 2003. Predicting rare plant occurrence in Great Smoky Mountains National Park, USA. Natural Areas Journal 23:229–237.

Borgo, A. 2003. Habitat requirements of the Alpine marmot *Marmota marmota* in re-introduction areas of the Eastern Italian Alps. Formulation and validation of habitat suitability models. Acta Theriologica 48:557–569.

Browning, D. M., S. J. Beaupre, and L. Duncan. 2005. Using partitioned Mahalanobis $D^2(K)$ to formulate a GIS-based model of timber rattlesnake hibernacula. Journal of Wildlife Management 69:33–44.

Busby, J. 1991. BIOCLIM—a bioclimate analysis and prediction system. Plant Protection Quarterly 6:8–9.

Clark, J. D., J. E. Dunn, and K. G. Smith. 1993. A multivariate model of female black bear habitat use for a geographic information system. Journal of Wildlife Management 57:519–526.

Corsi, F., E. Dupre, and L. Boitani. 1999. A large-scale model of wolf distribution in Italy for conservation planning. Conservation Biology 13:150–159.

Dunn, J. E., and L. Duncan. 2000. Partitioning Mahalanobis D^2 to sharpen GIS classification. Pages 195–204 *in* C. A. Brebbia and P. Pascolo, editors. Management information systems 2000: GIS and remote sensing. WIT Press, Southhampton, United Kingdom.

Fortin, M., T. Keitt, B. Maurer, M. Taper, D. Kaufman, and T. Blackburn. 2005. Species' geographic ranges and distributional limits: pattern analysis and statistical issues. Oikos 108:7–17.

Griffin, S. C. 2007. Demography and ecology of a declining endemic: the Olympic marmot. Dissertation, University of Montana, Missoula, USA.

Griffin, S. C., M. L. Taper, R. Hoffman, and L. S. Mills. 2008. The case of the missing marmots: are metapopulation dynamics of range-wide declines responsible? Biological Conservation 141:1293–1309.

Griffin, S. C., T. Valois, M. L. Taper, and L. S. Mills. 2007. Effects of tourists on behavior and demography of Olympic marmots. Conservation Biology 21:1070–1081.

Guisan, A., and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8:993–1009.

Guisan, A., and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135:147–186.

Hanski, I. 1998. Metapopulation dynamics. Nature (London) 396:41–49.

Hetrick, W. A., P. M. Rich, and S. B Weiss. 1993. Modeling insolation on complex surfaces. Thirteenth Annual ESRI User Conference 2:447–458.

Hirzel, A. H., J. Hausser, D. Chessel, and N. Perrin. 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology 83:2027–2036.

Johnson, C. J., and M. P. Gillingham. 2005. An evaluation of mapped species distribution models used for conservation planning. Environmental Conservation 32:117–128.

Knick, S. T., and D. L. Dyer. 1997. Distribution of black-tailed jackrabbit habitat determined by GIS in southwestern Idaho. Journal of Wildlife Management 61:75–85.

Knick, S. T., and J. T. Rotenberry. 1998. Limitations to mapping habitat use areas in changing landscapes using the Mahalanobis Distance Statistic. Journal of Agricultural, Biological, and Environmental Statistics 3:311–322.

Lele, S. R., and J. L. Keim. 2006. Weighted distributions and estimation of resource selection probability functions. Ecology 87:3021–3028.

MacKenzie, D. I., J. D. Nichols, J. E. Hines, M. G. Knutson, and A. B. Franklin. 2003. Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. Ecology 84:2200–2207.

MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollack, L. L. Bailey, and J. E. Hines. 2006. Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. Academic Press, New York, New York, USA.

Manel, S., J. M. Dias, S. T. Buckton, and S. J. Ormerod. 1999. Alternative methods for predicting species distribution: an illustration with Himalayan river birds. Journal of Applied Ecology 36:734–747.

McArdle, B. H. 1990. When are rare species not there? Oikos 57:276–277.

Melcher, J. C., K. B. Armitage, and W. P. Porter. 1990. Thermal influences on the activity and energetics of yellow-bellied marmots (*Marmota flaviventris*). Physiological Zoology 63:803–820.

Mladenoff, D. J., T. A. Sickley, and A. P. Wydeven. 1999. Predicting gray wolf landscape recolonization: logistic regression models vs. new field data. Ecological Applications 9:37–44.

Ozgul, A., K. B. Armitage, D. T. Blumstein, D. H. Van Vuren, and M. K. Oli. 2006. Effects of patch quality and network structure on patch occupancy dynamics of a yellow-bellied marmot metapopulation. Journal of Animal Ecology 75:191–202.

Pacific Meridian Resources. 1996. National Park Service—Pacific Northwest Region. Vegetation and landform database development study. Pacific Meridian Resources, Portland, Oregon, USA.

Pearce, J. L., and M. S. Boyce. 2006. Modelling distribution and abundance with presence-only data. Journal of Applied Ecology 43:405–412.

Peeters, E. T. H. M, and J. J. P. Gardeniers. 1998. Logistic regression as a tool for defining habitat requirements of two common gammarids. Freshwater Biology 39:605–615.

Podruzny, S. R., S. Cherry, C. C. Schwartz, and L. A. Landenburger. 2002. Grizzly bear denning and potential conflict areas in the Greater Yellowstone Ecosystem. Ursus 13:19–28.

Rotenberry, J. T., K. L. Preston, and S. T. Knick. 2006. GIS-based niche modeling for mapping species habitat. Ecology 87:1458–1464.

Sorenson, T. 1948. A method of establishing groups of equal amplitude in a plant based on similarity of species content and its applications to analysis of vegetation on Danish commons. Biologiske Skrifter 5:1–34.

Thatcher, C. A., F. T. Van Manen, and J. D. Clark. 2006. Identifying suitable sites for Florida panther reintroduction. Journal of Wildlife Management 70:752–763.

Thompson, L. M., F. T. van Manen, S. E. Schlarbaum, and M. DePoy. 2006. A spatial modeling approach to identify potential Butternut restoration sites in Mammoth Cave National Park. Restoration Ecology 14:289–296.

Tsoar, A., O. Allouche, O. Steinitz, D. Rotem, and R. Kadmon. 2007. A comparative evaluation of presence-only methods for modelling species distribution. Diversity and Distributions 13:397–405.

Türk, A., and W. Arnold. 1988. Thermoregulation as a limit to habitat use in alpine marmots (*Marmota marmota*). Oecologia 76:544–548.

U.S. Geological Survey. 2000. National elevation dataset. Earth Resources Observation and Science Data Center, Sioux Falls, South Dakota, USA.

van Manen, F. T., J. A. Young, C. A. Thatcher, W. B. Cass, and C. Ulrey. 2005. Habitat models to assist plant protection efforts in Shenandoah National Park, Virginia, USA. Natural Areas Journal 25:339–350.

Associate Editor: Euler.