Chapter 9
Landscape Applications of Machine Learning: Comparing Random Forests and Logistic Regression in Multi-Scale Optimized Predictive Modeling of American Marten Occurrence in Northern Idaho, USA

Samuel A. Cushman and Tzeidle N. Wasserman

9.1 Introduction

The American marten (martes americana) is a species that is dependent on old conifer forest at middle to high elevations and is highly sensitive to habitat loss and fragmentation in a scale dependent fashion (e.g., Hargis et al. 1999; Wasserman et al. 2012a, b), and forest management is often influenced by considerations of how management will affect extent and pattern of marten habitat. Due to their dependence on extensive, unfragmented forest landscapes and microhabitat structures associated with late successional forest (Buskirk and Ruggiero 1994; Hargis et al. 1999), American marten are sensitive to fragmentation of late seral forest habitats, such as that resulting from timber harvest and associated extraction routes and road building (e.g., Cushman et al. 2011). Previous studies have consistently shown that American marten habitat requirements include forests with high canopy cover (Hargis and McCullough 1984; Wynne and Sherburne 1984), abundant near ground structure (Chapin et al. 1998; Godbout and Ouellet 2008), high prey densities (Fuller and Harrison 2005), and sufficient snow depth to provide subnivean spaces during winter (Wilbert et al. 2000). These habitats are thought to provide opportunities for foraging, resting, denning, thermoregulation, and avoiding predation. Perturbations, such as timber harvest, remove canopy cover, reduce coarse woody debris, change mesic sites into xeric sites, remove riparian dispersal zones, and change prey communities (Buskirk and Ruggiero 1994). American marten avoid

S. A. Cushman (✉)
U.S. Forest Service, Rocky Mountain Research Station, Flagstaff, AZ, USA
e-mail: scushman@fs.fed.us

T. N. Wasserman
School of Forestry, Northern Arizona University, Flagstaff, AZ, USA
e-mail: tnw23@nau.edu

G. R. W. Humphries et al. (eds.), *Machine Learning for Ecology and Sustainable Natural Resource Management*, https://doi.org/10.1007/978-3-319-96978-7_9

areas with even relatively low levels of forest fragmentation and rarely use sites where more than 25% of forest cover has been removed (Hargis et al. 1999). Highly contrasting edge habitats, such as borders between late successional forest and harvested patches, and areas of open canopy are strongly avoided (Buskirk and Ruggiero 1994; Hargis et al. 1999; Cushman et al. 2011).

Recently, Wasserman et al. (2012a, b) predicted and mapped habitat suitability for American marten in northern Idaho, U.S.A. They used multiple scale habitat suitability modeling with logistic regression on a set of marten presence-absence locations collected non-invasively using genetic (hair) samples across a 3884 square kilometer region to quantify the relative importance of topographical, vegetation, and landscape metric variables in predicting marten occurrence. The Wasserman et al. (2012a, b) model identified strong and consistent relationships with various measures of landscape fragmentation: marten occurrence was positively associated with landscapes that contained high canopy closure, low density of all roads (including small forest roads), few past clear-cuts, and extensive late seral forest. Several of these variables had maximum influence on marten probability of occurrence at fairly broad spatial scales. At scales approximately the size of marten home ranges (500–1000 m radius; Tomson et al. 1999) within our study area, the Wasserman et al. (2012a, b) model showed that American marten select landscapes with high average canopy closure, low road density, and low forest fragmentation. Within these low-fragmentation landscapes, the model showed marten select foraging habitat at a fine scale (90 m) within middle-elevation, late-seral, mesic forests. This is consistent with the results of previous studies, which have shown high sensitivity to landscape fragmentation and perforation by non-stocked clear-cuts (Hargis et al. 1999; Cushman et al. 2011), and strong preference of American marten in northern Idaho for mesic riparian forest conditions in unfragmented watersheds (Tomson 1999; Shirk et al. 2014).

For a decade, logistic regression has been the dominant method in multi-scale habitat modeling (Hegel et al. 2010; McGarigal et al. 2016). Random forests (RF; Breiman 2001a, b) is increasingly used in a range of applications including digital soil mapping (Grimm et al. 2008), forest biomass mapping (Baccini et al. 2012), species distribution modeling (Evans and Cushman 2009), land cover change prediction (Cushman et al. 2017) and others given its often superior performance compared to other methods (Evans et al. 2011; Mi et al. 2017). However, there have been relatively few formal comparisons of the performance of multi-scale modeling between logistic regression and random forests. Recently, Cushman et al. (2017) compared the performance of logistic regression with random forests in a multi-scale optimized predictive modeling study of deforestation risk across Borneo. As found in virtually all of such investigations, the authors found that random forests substantially outperformed logistic regression. Our interest in this study is to conduct a similar comparison of logistic regression and random forests in multi-scale optimized predictive model of occurrence of a forest-dependent mammal species, the American marten (*Martes americana*) in northern Idaho USA.

The main purpose of this chapter is to compare the predictive power and the ecological interpretation of the Wasserman et al. (2012a, b) logistic regression model

with a model produced on the same data using the same multi-scale optimization approach, but using random forests instead of logistic regression. Based on past work showing that random forests often outperforms other predictive modeling approaches (e.g. Evans et al. 2011; Cushman et al. 2017), we predicted that the random forests model would outperform the logistic regression model based on AUC (area under the receiver operator curve). Also, previous work has shown that marten habitat selection is highly scale dependent (e.g., Hargis et al. 1999; Wasserman et al. 2012a, b), and a recent review has demonstrated that multi-scale optimization is important for habitat modeling in general (McGarigal et al. 2016). Accordingly, an additional goal of this chapter is to see if the inferences about what variables are important and at what scales they are operative differ between models developed with random forests and GLM logistic regression.

9.2 Methods

9.2.1 Study Area

The study area is a 3884 km^2 section of the Selkirk, Purcell, and Cabinet Mountains, encompassing the Bonners Ferry and Priest River Ranger Districts of the Idaho Panhandle National Forest (2282 km^2) and adjacent non National Forest System lands, including private land (986 km^2), State (508 km^2), tribal- and other federally managed land (Fig. 9.1). The topography is mountainous, with steep ridges, narrow

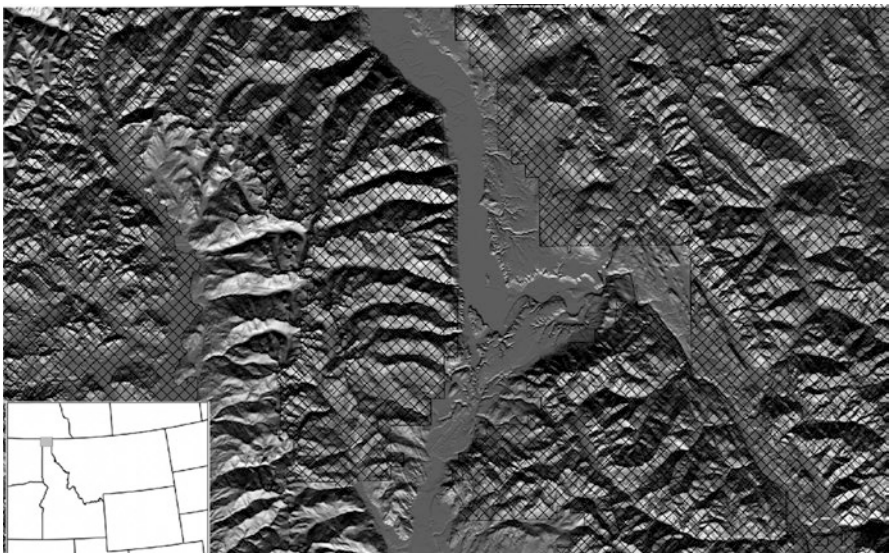

Fig. 9.1 Study area orientation map. Idaho Panhandle National Forest lands are cross-hatched

valleys, and many cliffs and cirques at the highest elevations. Elevation ranges from approximately 700 m to 2400 m above sea level. The climate is characterized by cold, moist winters and dry summers. The average daily maximum temperature at Bonners Ferry, the largest town in the study area, in the coldest month (January) is 0.2 °C, while that of the warmest month (July) is 27.8 °C. Average precipitation in the wettest month (December) amounts to 7.84 cm, while that of the driest month (July) is 2.33 cm, with an average annual total of 56.4 cm.

The area is heavily forested, with subalpine fir (*Abies lasiocarpa*) and Engelmann Spruce (*Picea engelmannii*) co-dominant above 1300 m, and a diverse mixed forest of Douglas-fir (*Pseudotsuga menziesii*), lodgepole pine (*Pinus contorta*), ponderosa pine (*Pinus ponderosa*), western white pine (*Pinus monticola*), grand fir (*Abies grandis*), western hemlock (*Tsuga heterophylla*), western red cedar (*Thjua plicata*), western larch (*Larix occidentalis*), paper birch (*Betula papyrifera*), quaking aspen (*Populus tremuloides*), and black cottonwood (*Populus trichocarpa*) dominating below 1300 m.

9.2.2 Occurrence Data and Logistic Regression Model

We decided to utilize a multi-scale habitat suitability model produced by (Wasserman et al. 2012a, b), who used multi-scale logistic regression modeling to predict habitat suitability from a presence/absence dataset collected non-invasively through hair snaring (e.g., Wasserman et al. 2010). To obtain data on American marten presence, Wasserman et al. (2010) deployed hair snare stations at 361 locations well distributed across a representative sample of topographical and ecological gradients over three winter seasons (2005, 2006, and 2007; 1 survey per site). Recently, Robinson et al. (2017) showed that this kind of non-invasive genetic sampling is consistent and has high success for species and individual identification across seasons and weather patterns. Genetic analysis confirmed the detection of American marten at 159 individual hair snare stations. (Wasserman et al. 2012a, b) selected variables *a priori* assumed to be related to American marten occurrence based on previous research (Buskirk and Ruggiero 1994; Hargis et al. 1999; Tomson 1999), including elevation, percent canopy closure, road density, patch density, percentage of the local landscape surrounding survey sites occupied by late seral forests, percentage of the landscape occupied by non-stocked clear-cuts, and probability of occurrence of each major tree species (western red cedar and six other species) in each cell across the landscape.

The first step undertaken by (Wasserman et al. 2012a, b) was to use bivariate scaling (Thompson and McGarigal 2002; Grand and others 2004) to identify the scale at which each of these independent variables was most strongly related to American marten occurrence. Given that environmental factors may be related to deforestation at a range of spatial scales (Wiens 1989), and given the critical importance of multi-scale optimization to correct inferences about habitat selection

(McGarigal et al. 2016), Wasserman et al. (2012a, b) calculated all predictor variables at 12 spatial extents including focal radii of 90 to 990 m at 90 m increments. This resulted in reduction of the model to seven variables significantly related to marten occurrence (Table 9.1). Wasserman et al. (2012a, b) then used logistic regression to test all combinations of these predictor variables, without interactions, and used model averaging, based on AIC weights, to produce parameter estimates for a final model predicting probability of marten occurrence. This model was then used to evaluate the impacts of past timber harvest and road building on the extent and quality of available marten habitat.

9.2.3 Predictor Variables for Analysis

A priori, we proposed several environmental and anthropogenic variables as predictors of marten occurrence. Following Wasserman et al. (2012a, b) we included road density and canopy closure, as well as a number of topographical and landscape composition and configuration metrics. Topographical variables included elevation and several terrain complexity measures produced using the Geomorphometry and Gradient Metrics Toolbox (ArcGIS 10.0; Evans et al. 2014). These included: topographical roughness, which measures the topographical complexity of the landscape

Table 9.1 Variables included in the Wasserman et al. 2012a, b habitat model used in the current analyses. There were seven variables in the habitat model, related to elevation, road density, canopy cover, patch density in the landscape mosaic, large saw timber, non-stocked clear cuts and western red cedar forest types. Each of these was included in the habitat model at a particular spatial scale (focal extent) at which it most strongly affected probability of occurrence. These scales ranged from 90 m in radius (western red cedar and large saw timber) to a maximum extent of influence of road density at a 1980 m radius. Each of these variables had different effects on marten probability of occurrence. Effect size in this table records the percent change in the probability of marten occurrence as the associated variable changes from the 10th to the 100th percentile value in the dataset, holding the other variables constant at their medians. Based on this measure of effect size, the most important predictors, in decreasing order of importance, are western red cedar forest type, percent canopy cover, road density, patch density, percent of the landscape in non-stocked clear cuts, elevation, and finally large saw timber

Predictor variable	Most significant scale (m)	Effect size
Elevation	1400	19.78
Road density	1980	−53.05
Percent canopy cover	990	61.05
Patch density	990	−46.26
Percentage of the focal landscape in large sawtimber	90	13.21
Percentage of the landscape in non-stocked conditions	990	−35.99
Western red cedar	90	77.21

within a defined focal extent (Blaszczynski 1997), relative slope position, which measures the relative position of the focal pixel within a defined extent on a gradient from valley bottom to ridge top (Evans et al. 2014), dissection index, which is the ratio between relative relief and to the absolute relief, curvature index, which measures the rate of change of local slope, heat load index, which predicts the total incident solar radiation as a function of latitude and topography, and compound topographical index, which models the cumulative aggregation of water flow through every cell in the landscape.

We also included FRAGSTATS metrics quantifying the extent and configuration of different land cover classes across a range of focal extents as predictor variables (McGarigal et al. 2012). The classes used in the analysis include: (1) large sawtimber (> 24 inches DBH), (2) small sawtimber (12–24 inches DBH), (3) pole timber (3–12 inches DBH), (4) sapling/seedling (< 3 inches DBH), (5) non-stocked forestland, and (6) non-forest (Wasserman et al. 2012a, b). For each of these classes we used FRAGSTATS 4.0 (McGarigal et al. 2012) to calculate five class-level (area-weighted mean patch size, Area_AM; edge density; ED, patch density, PD; percentage of the landscape, PLAND; area-weighed proximity index, PROXAM), and four landscape-level metrics (aggregation index, AI; contrast-weighted edge density, CWED; edge density, ED; patch density, PD). These metrics were chosen given that they measure several critical attributes of habitat extent and fragmentation that have been shown to have important influences on habitat selection (e.g., Chambers et al. 2016) and population connectivity (e.g., Cushman et al. 2013). Also, following Wasserman et al. (2012a, b) we calculated all variables within 12 focal scales ranging from 90 to 990 m radii around each sampling location to enable multi-scale model optimization.

9.2.4 Modeling Approaches

We used random forests machine learning and logistic regression to predict marten occurrence in the study landscape. We used the logistic regression model and results as published in Wasserman et al. (2012a, b). Random forests is a classification and regression tree (CART; De'ath and Fabricius 2000) - based bootstrap method that corrects many of the known issues in CART, such as over-fitting (Breiman 2001a, b; Cutler et al. 2007), multi-collinearity and variable interaction, and provides very well-supported predictions with large numbers of independent variables (Cutler et al. 2007). We used a modeling approach developed by Evans and Cushman (2009) to predict occurrence of marten using the random forests method (Breiman 2001a, b; Cutler et al. 2007) as implemented in the package 'randomForest' (Liaw and Wiener 2002) in R (R Development Core Team 2008).

We conducted the random forests in two steps, mirroring the approach Wasserman et al. (2012a, b) used in the original logistic regression model. We recognize that

using random forests like a GLM does not unleash all its powers, but our purpose was to conduct a strict comparison keeping as many parameters as similar as possible to see how random forests and GLM differed in their predictions in this context. First, we ran univariate models across the multiple scales to identify the scale at which each variable had the strongest ability to predict marten occurrence, as suggested by McGarigal et al. (2016) as a robust approach for multi-scale model optimization, and as shown to work well for random forests by Cushman et al. (2017). To accomplish this, we ran a series of single random forests analyses for each variable across the 12 scales in each nation and used the Model Improvement Ratio (MIR; Murphy et al. 2010) to measure the relative predictive strength of each scale of the variable. The MIR calculates the permuted variable importance, represented by the mean decrease in out-of-bag error, standardized from zero to one. We compared the MIR scores for all scales for each variable, and retained the scale that had the highest MIR score for further multivariate modeling.

In the second step we used random forests to develop multivariate models predicting probability of marten occurrence as a function of landscape condition across the suite of scale-optimized variables. To identify the most parsimonious random forests model we applied the Model Improvement Ratio (MIR; Murphy et al. 2010). In model selection using MIR, the variables were subset using 0.10 increments of MIR value, with all variables above the threshold retained for each model. This subset was always performed on the original model's variable importance to avoid over-fitting (Svetnik et al. 2004). We compared each subset model and selected the model that exhibited the lowest total out-of-bag error and lowest maximum within-class error.

Model predictions for the random forests model were created by using a matrix of the ratio of majority votes to create a probability distribution. Random forests makes predictions based on the plurality of votes across all bootstrap trees and not on a single rule set. This votes-matrix can be scaled and treated as a probability given the error distribution of the model (Evans and Cushman 2009; Murphy et al. 2010). We used the function that (Evans and Cushman 2009) added to GridAsciiPredict (Crookston and Finley 2008) which uses the votes-probability function to write the probabilities to ASCII grids.

9.2.5 Model Assessment

There are a multitude of ways to assess the performance of predictions of the random forests and logistic regression models, and most previous studies have used the Kappa statistic (Cohen 1968) and similar measures of improvement of predicted classification compared to random assignment (e.g., based on the confusion matrix). However, following Ponitus and Milones (2011), we avoided the Kappa statistic given that it does not report a meaningful statistical measure of predictive success, even when corrected to address the two different aspects of prediction related to

predicted amount and predicted location (Pontius and Si 2014). In addition, since the predictions we produced using random forests and logistic regression are in the form of predicted probabilities, it is more meaningful to assess the continuous pattern or predicted probability in comparison to the actual observed changes than to cross-tabulate observed vs. predicted change (Pontius and Si 2014) We chose this approach because transforming predicted probabilities into categorical responses requires using a threshold cut-point or probabilistic function, which loses information on the actual quality of the prediction (Pontius and Milones 2011; Pontius and Si 2014). We assessed the performance of the random forests and logistic regression predictions using area under the Total Operating Characteristic curve (Pontius 2014), as suggested by Pontius and Si (2014) and Pontius and Parmentier (2014). We also produced predicted probability of occurrence maps for both models and visually compared these to describe the differences in the pattern of predicted habitat suitability.

9.3 Results

9.3.1 Random Forests Univariate Scaling

The first step in the modeling approach was to identify the best scale for each individual variable out of the 12 scales considered (90–990 m, by 90 m increments), based on Model Improvement Ratio. For each variable we chose the scale with the largest Model Improvement Ratio, except in some cases we retained two scales if the second had an MIR value over 0.75 and differed substantially in scale from the scale with the highest MIR value (Table 9.2). There was a relatively broad range of scales selected across all variables (Fig. 9.2), with an apparent bimodal pattern where more variables were selected at either the broadest (greater than 630 m radius), or finest (less than 270 m radius) scales.

9.3.2 Random Forests Multivariate Model

The multivariate random forests model used the Model Improvement Ratio as a variable selection approach. The final model included 14 variables (Fig. 9.3). Five of these were selected at the broadest scale of 990 m, showing a stronger pattern of dominance by broad-scale relationships in the multivariate reduced model than in the univariate scaling.

We produced LOWESS splines of the pattern of presence vs. absence across the sampled range of each of the top eight variables. LOWESS (locally weighted scatterplot smoothing) is a non-parametric regression method that combine multiple regression models in a k-nearest-neighbor-based meta-model to produce non-linear

Table 9.2 Variables included in the random forests modeling and the scales retained in the univariate scaling step. Land – Landscape-level FRAGSTATS variable; Class – Class-level FRAGSTATS variable

Variable	Acronym	Top scale	Second scale retained
Agregation index (Land)	AI	630	180
Road density	AR	180	1440
Area-weighted mean patch size (Class)	areaam1	900	180
Area-weighted mean patch size (Class)	areaam2	720	
Area-weighted mean patch size (Class)	areaam3	540	
Area-weighted mean patch size (Class)	areaam4	990	
Area-weighted mean patch size (Class)	areaam5	450	
Area-weighted mean patch size (Class)	areaam6	990	
Mean canopy cover	canopy	180	630
Topographical curvature index	crv	810	630
Compound topographical index	cti	270	90
Contrast-weighted edge density (Land)	cwed	360	
Topographical dissection index	dis	90	
Edge density (Land)	ed	90	990
Edge density (Class)	ed1	630	
Edge density (Class)	ed2	90	
Edge density (Class)	ed3	180	
Edge density (Class)	ed4	990	
Edge density (Class)	ed5	450	
Edge density (Class)	ed6	900	
Elevation	elev90	720	
Heat load index	hil	720	270
Patch density (Land)	pd	990	630
Patch density (Class)	pd1	990	
Patch density (Class)	pd2	810	
Patch density (Class)	pd3	720	
Patch density (Class)	pd4	810	
Patch density (Class)	pd5	360	
Patch density (Class)	pd6	270	90
Percentage of the landscape (Class)	pland1	990	180
Percentage of the landscape (Class)	pland2	720	
Percentage of the landscape (Class)	pland3	990	
Percentage of the landscape (Class)	pland4	990	
Percentage of the landscape (Class)	pland5	360	
Percentage of the landscape (Class)	pland6	630	900
Proximity index (class)	proxam1	450	990
Proximity index (Class)	proxam2	450	
Proximity index (Class)	proxam3	540	
Proximity index (class)	proxam4	810	
Proximity index (Class)	proxam5	900	
Topographical roughness	r	90	900
Slope position	sp	810	

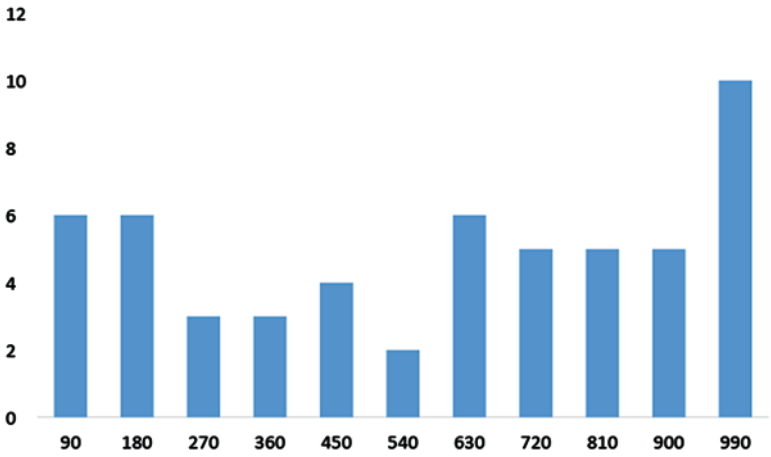

Fig. 9.2 Frequency of selected scales (in meters) across all variables for the random forests model

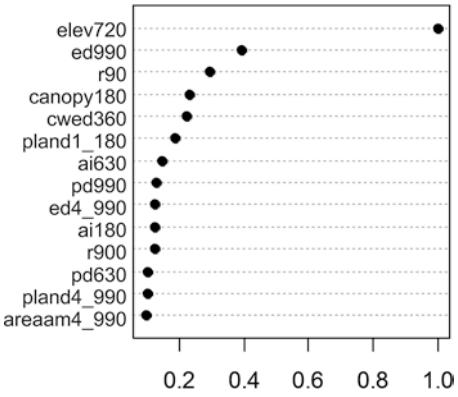

Fig. 9.3 Model improvement ratio plot for the selected variables. The most important variable is mean elevation within a 720 m focal radius (elev90720). The other variables are listed in order of their importance relative to elevation, with the x-axis indicating the relative additional model improvement when adding each successive variable

splines showing the response pattern in a bivariate scatter plot. The most important variable by far, based on the MIR, was mean elevation within a 720 m focal radius (Fig. 9.3). Marten occurrence has a strongly non-linear relationship with elevation; detections are very rare below 1000 m, rising rapidly to an apparent unimodal peak at approximately 1280 m, and then slowly declining at the highest elevations

(Fig. 9.4a). The second most important variable based on MIR was edge density within a 720 m focal radius. Marten occurrence has a nonlinear relationship with edge density as well, with the highest detection rates generally occurring at low edge densities (Fig. 9.4b). The third most important variable was topographical roughness at a 90 m focal radius, with marten occurrence increasing monotonically but nonlinearly with increasing topographical roughness (Fig. 9.4c). Mean canopy cover within a 180 m focal radius was the fourth most important variable in the random forests model, with marten detections increasing strongly, but again nonlinearly, at high levels of local canopy cover (Fig. 9.4d). The fifth most important variable based on MIR was contrast-weighted edge density, with marten occurrence declining with increasing density of high-contrast edges in the landscape mosaic

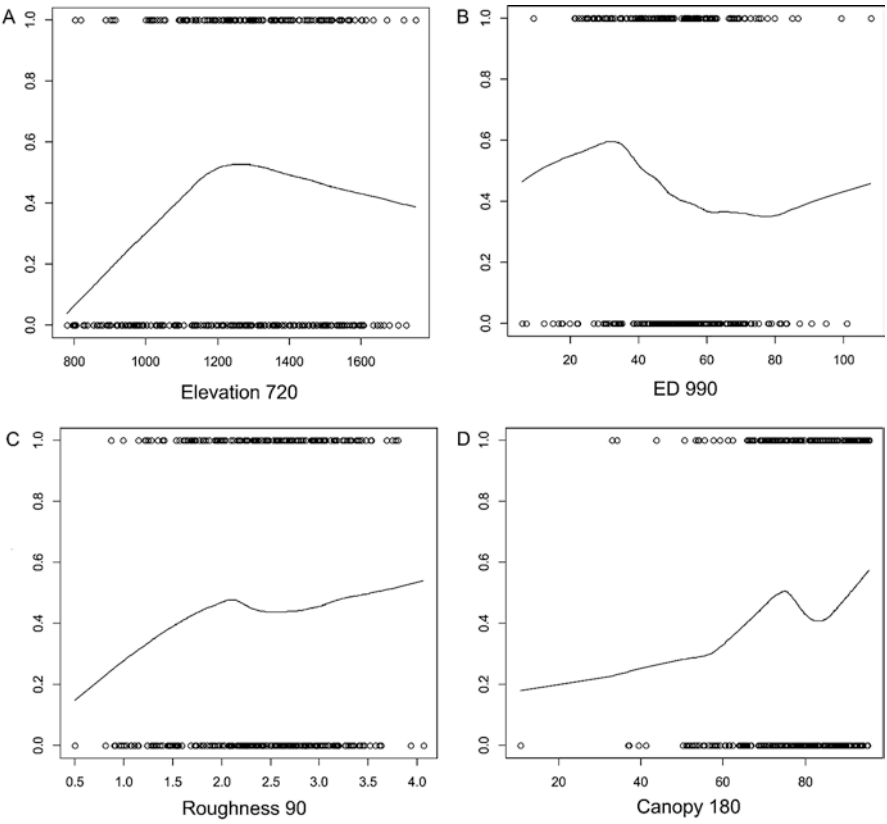

Fig. 9.4 (**a**–**d**) – part 1. Scatter plots of presence and absence and fitted LOWESS splines for the first four variables selected by the Model Improvement Ratio variable selection process for the multivariate random forests model

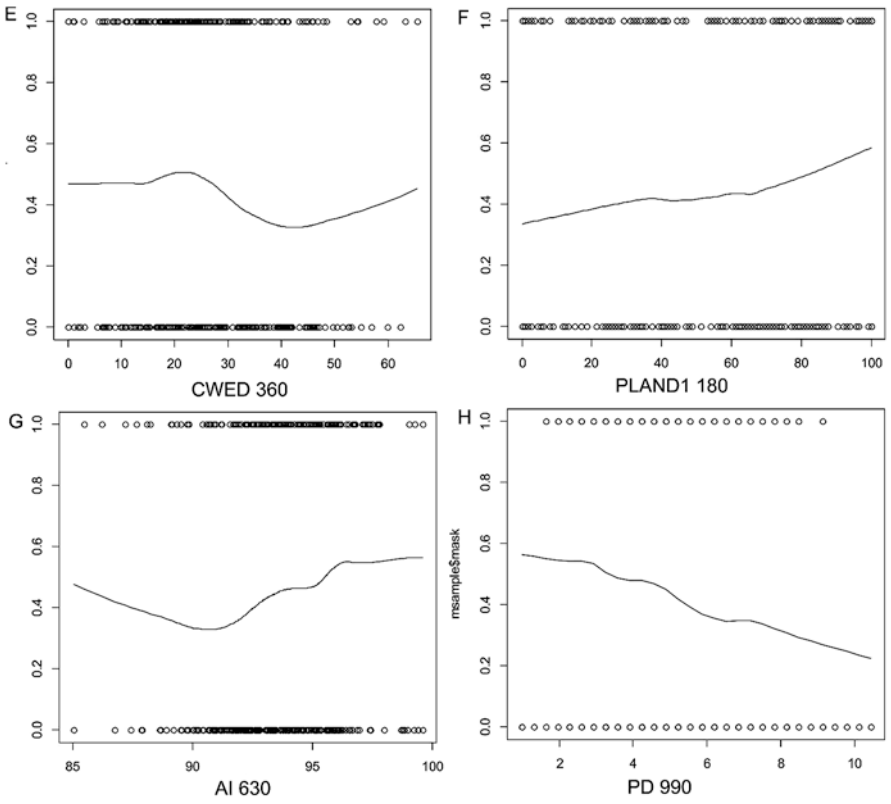

Fig. 9.4 (**e**–**h**) – part 2. Scatter plots of presence and absence and fitted LOWESS splines for the fifth through eighth variables selected by the Model Improvement Ratio variable selection process for the multivariate random forests model

(Fig. 9.4e). The percentage of a 180 m radius focal landscape occupied by large saw timber was the sixth most important variable, with occurrence frequency with occurrence frequency increasing monotonically with the amount of large, old forest in the local landscape (Fig. 9.4f). Landscape-level aggregation index was the seventh most important variable, with the frequency of marten occurrence increasing non-linearly but monotonically with increasing landscape aggregation within a 630 m focal radius (Fig. 9.4g). Landscape-level patch density within a 990 m focal landscape was the eighth most important variable, with monotonically decreasing frequency of marten as patch density increased (Fig. 9.4h).

9.3.3 Model Comparison

There was substantial similarity in the qualitative interpretation of the Wasserman et al. (2012a, b) logistic regression and the random forests model produced for this chapter. In both models occurrence was strongly predicted by a unimodal function

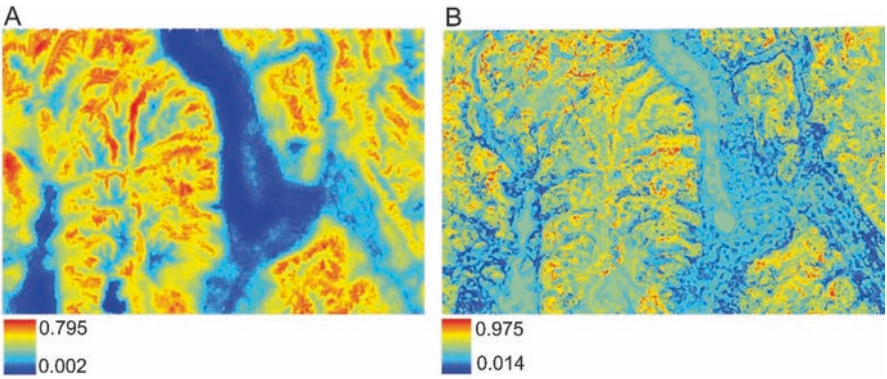

Fig. 9.5 Comparison of predicted probability of marten occurrence from the logistic regression (**a**) and random forests (**b**) models across the study area

of elevation, a non-linear function of canopy cover, a non-linear function of patch density, and the extent of the landscape in large conifer forest. However, there also were some important differences. First and foremost in performance (inference from predictions). Secondly, when looking at the predictors, road density and percentage of the landscape in non-stocked forestland were included in the final model-averaged logistic regression prediction, while these variables were not selected by the MIR in the random forests model.

In addition, a number of other variables were included in the random forests model that were not included in the logistic regression model, notably edge density, topographical roughness, contrast-weighted edge density and aggregation index. Together, these variables provide a substantially stronger "fragmentation signal" in the random forests model than the logistic regression model, with stronger identification of the negative effects of landscape heterogeneity than indicated by the logistic regression model.

In both models, extent of large sawtimber forest was a strong predictor at a fine spatial scale, while patch density was a strong predictor at the broadest spatial scale tested. This suggests that both models predict that optimal American marten habitat consists of patches of large, old forest within broad forested landscapes that have low levels of heterogeneity or fragmentation. However, the logistic regression model identified canopy cover as having the strongest effect at the broadest scale, while the random forests model identified a relatively fine scale effect of canopy cover.

A visual comparison of the predicted probability maps (Fig. 9.5) shows three main differences in the spatial prediction of marten habitat between the logistic regression and the random forests model. As also seen by Cushman et al. (2017), random forests produces predictions that are more discriminatory, with higher range of predicted probability and higher spatial heterogeneity than logistic regression. Logistic regression fits smooth linear functions of a linear combination of variables, which results in simple and smooth patterns of predicted occurrence. The logistic

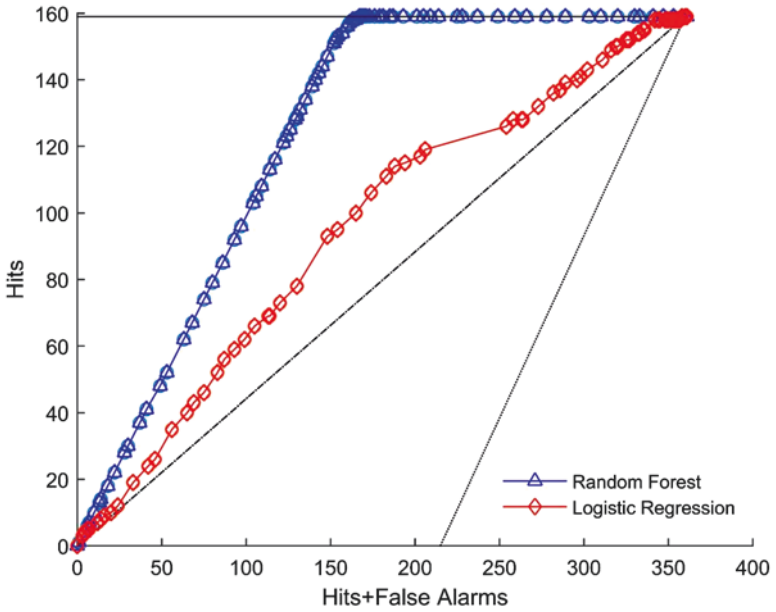

Fig. 9.6 TOC curves showing comparative model performance among the logistic regression, random forests, random forests without fragmentation variables, and the naïve model. Higher model performance is indicated by stronger convex curvature toward the upper left corner of the plot space. The AUC for the two models are 0.983 for random forests and 0.701 for logistic regression

regression map highlights areas of high canopy cover, high extent of old forest, and low fragmentation at middle to upper elevations. In contrast, the random forests model shows much higher heterogeneity of predictions, with steeper and stronger gradients of habitat quality across the landscape. Both models indicate that marten habitat quality is highest in middle to upper elevation areas with high canopy cover, low fragmentation, and high cover of old forest, but the random forests model shows that habitat quality varies more across space, with more areas of predicted very high occurrence probability, interspersed with areas of predicted lower quality, which are not seen in the logistic regression model predictions.

9.3.4 Model Performance

We assessed model performance based on the area under the TOC curve (Fig. 9.6). The logistic regression model has an AUC of 0.701, as previously reported by Wasserman et al. (2012a, b), indicating moderately good success in predicting presence vs. absence in the training dataset. By comparison, the random forests model had an AUC of 0.981, indicating very high predictive ability, and a much stronger ability to predict presences and absences in the training dataset than the logistic regression model. Expressed as a percentage, the random forests model had 28% higher performance, leading to much better prediction of habitat suitability, better

inferences about habitat variables influencing marten occurrence, improved identification of scale dependency, and ultimately, therefore, better guidance to conservation and management (Breiman 2001b).

9.4 Discussion

Consistent with the results of other researchers who found that random forests outperforms other methods for prediction and classification (e.g., Cushman et al. 2010; Evans et al. 2011; Drew et al. 2010; Rodriguez-Galiano et al. 2012; Schneider 2012; Cushman et al. 2017), we expected that random forests would outperform logistic regression in predicting marten occurrences. Consistent with this expectation, random forests greatly outperformed logistic regression based on AUC measures of predictive success. This confirms the superiority of random forests as a modeling tool for habitat modeling, and we suggest that future studies use this powerful technique as a baseline.

Our analysis provided insight into patterns of scale-dependent habitat selection in American marten. The results were generally consistent with those found by Wasserman et al. (2012a, b). Specifically, the models show that American marten occurrence is highest in middle to upper elevation forested landscapes with high local canopy closure and high local cover of old-growth forest, and low levels of landscape heterogeneity and fragmentation at broader scales. In essence, our model reconfirms the description of Wasserman et al. (2012a, b) for optimal American marten habitat in northern Idaho: "*… at the scale of home ranges, marten select landscapes with high average canopy closure and low fragmentation. Within these low-fragmentation landscapes, marten select foraging habitat at a fine scale within late-seral, middle-elevation mesic forests. In northern Idaho, optimum American marten habitat, therefore, consists of landscapes with low road density and low density of non-forest patches with high canopy closure and large areas of middle-elevation, late successional mesic forest.*" Our analysis augments this interpretation with further emphasis on the importance of landscape heterogeneity at intermediate (CWED at 360 m) to broad scales (AI at 630 m, PD at 990 m), suggesting perhaps a larger importance of landscape fragmentation than suggested by the Wasserman et al. (2012a, b) analysis.

The random forests and logistic regression models were also quite different in their spatial predictions, with logistic regression producing smooth, monotonic patterns of predicted suitability, while random forests produced a map with higher heterogeneity and discrimination, showing stronger identification of areas of high suitability for marten. These differences are highly relevant if predictions from models are to be used effectively for management and conservation. Conservation prioritization based on habitat suitability would likely be quite different when based on either of these two maps, with the logistic regression producing coarse recommendations to protect middle elevation, unfragmented, old-growth forest in general,

while the random forests would suggest the same general habitat niche but provide much stronger delineation of high priority areas.

Our analysis also provides an ability to assess patterns of scale-dependency in habitat relationships across a large number of predictor variables. This is an area of ongoing and increasing interest in landscape ecology (McGarigal et al. 2016). Relatively few studies have comprehensively evaluated patterns of scale dependence across pools of predictor variables. For example, Chambers et al. (2016) evaluated scale dependence of habitat associations and scaling patterns of landscape metrics in relation to bat occurrence or capture rate in forests of southwestern Nicaragua. They found that that edge density and patch density were the most important configuration variables across species, and percentage of the landscape was the most important class-level variable. In addition, they found that certain landscape and configuration metrics were most influential at fine (100 m) and/or broad (1000 m) spatial scales. Our results echo the importance of patch density and edge density as configuration predictor variables (the most important configuration variables in our analysis) and PLAND as a composition predictor variable (PLAND1 was the only composition variable in the random forests model).

One of the most important comparative differences between the logistic regression and the random forests models was their interpretation of scale dependence among the different predictor variables. In general both models found that landscape heterogeneity and forest fragmentation affected marten habitat suitability at broad scales, but the random forests analysis showed that fragmentation effects are active at both fine and broad scales, in contrast to the logistic regression which only identified these effects at broad scales. Also, the scales at which canopy cover and extent of old forest most strongly affected predictions were different between the models, indicating that the optimal scale of influence is highly sensitive to the method of modeling.

Random forests (Breiman 2001a, b) is a tree-based method based on "bagging" that is executed by bootstrapping (with replacement) 63% of the data and generating a weak learner based on a CART for each bootstrap replicate. Within the pre-set specification (e.g., node depth and number of samples per node) each CART is unconstrained (grown to fullest) and prediction is accomplished by tallying the 'majority votes' across all nodes in all random trees (Hegel et al. 2010). Independent variables are randomly selected at each node, with the number of variables selected at each node defined by m [sqrt(number of independent variables)]. These attributes provide several reasonable explanations for why random forests proved so much more powerful in predicting marten occurrence patterns in our northern Idaho dataset than did logistic regression. As seen in the LOWESS splines, there are strongly non-linear, often unimodal or multi-modal patterns of frequency of marten occurrence across the range of values of independent variables. Such complex non-linearity and non-monotonicity is a massive challenge to GLM modeling, such as logistic regression, even when, as in Wasserman et al. (2012a, b), nonlinear transformations are applied to the data. In contrast, the bootstrapping of CART within random forests provides the generation of a large number of trees which are combined across all nodes in all random trees. This enables immense flexibility to deal with

non-linearity and multi-modality of response, resulting in random forests models predicting patterns of presence and absence in the training data much more tightly than is possible with GLM or similar functional relationship methods. This also enables random forests to accurately reflect complex multi-variate non-linear interactions among predictor variables, which are typically completely ignored in GLM modeling (see Chap. 10 by Baltensperger).

In our case the logistic regression model was fair at prediction (AUC = 0.7) while the random forests model was excellent (AUC = 0.98), even though both were applied to the same data, and included largely the same predictor variables. This suggests that the difference in prediction is primarily due to random forest's superior ability to reflect the complex non-linear relationships and multi-variate interactions in the American marten habitat relationships in northern Idaho.

9.5 Conclusion

Random forests is shown here to substantially outperform logistic regression in predicting patterns of marten occurrence. This suggests, consistent with other research, that random forests may generally be a superior approach when the goal is obtaining high predictive power. It should be by now the starting platform for any analysis of this sort. The random forests model produced an ecological understanding that was generally similar to that provided by the logistic regression model, but with some additional detail and clarity regarding variables and scales of influence. However, given the much higher predictive success, applications of the random forests model for mapping habitat quality and assessing the extent and pattern of habitat is likely to produce much more accurate and useful information.

References

Baccini A, Goetz SJ, Walker WS et al (2012) Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. Nat Clim Chang 2(3):182–185

Blaszczynski JS (1997) Landform characterization with geographic information systems. Photogramm Eng Remote Sens 63(2):183–191

Breiman L (2001a) Random Forests. Mach Learn 45(1):5–32

Breiman L (2001b) Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci 16:199–231

Buskirk SW, Ruggiero LF (1994) The American marten. In: Ruggiero LF, Aubry KB, Buskirk SW, Lyon LJ, Zielinski WJ (eds) American marten, fisher, lynx, and wolverine in the western United States. Gen. Tech. Rep. RM-254. U.S. Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins

Chambers CL, Cushman SA, Medina-Fitoria A, Martinez-Fonesca J (2016) Influences of scale on bat habitat relationships in a forested landscape in Nicaragua. Landsc Ecol 31:1299–1318

Chapin TG, Hamson DJ, Katnik DD (1998) In audience FH Is that the correct title? of landscape pattern on habitat use by American marten in an industrial forest. Conserv Biol 12:1327–1337

Cohen J (1968) Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 70(4):213–220
Crookston NL, Finley AO (2008) yaImpute: An r package for knn imputation. J Stat Softw 23(10):1–16
Cushman SA, Gutzwiller K, Evans JS, McGarigal K (2010) The gradient paradigm: a conceptual and analytical framework for landscape ecology. In: Cushman SA, Huettman F (eds) Spatial complexity, informatics, and wildlife conservation. Springer, Tokyo, pp 83–108
Cushman SA, Macdonald EA, Landguth EL, Halhi Y, Macdonald DW (2017) Multiple-scale prediction of forest-loss risk across Borneo. Landsc Ecol 32:1581–1598
Cushman SA, Raphael MG, Ruggiero LF, Shirk AJ, Wasserman TN, O'Doherty EC (2011) Limiting factors and landscape connectivity: The American marten in the rocky mountains. Landsc Ecol 26:1137–1149
Cushman SA, Shirk AJ, Landguth EL (2013) Landscape genetics and limiting factors. Conserv Genet 14:263–274
Cutler DR, Edwards TC, Beard KH et al (2007) Random forests for classification ecology. Ecology 88(11):2783–2792
De'ath G, Fabricius KE (2000) Classification and Regression Trees: A powerful yet simple technique for ecological data analysis. Ecology 81(11):3178–3192
Drew CA, Wiersma YF, Huettmann F (eds) (2010) Predictive species and habitat modeling in landscape ecology: concepts and applications. Springer Science & Business Media, New York
Evans JS, Cushman SA (2009) Gradient modeling of conifer species using random forests. Landsc Ecol 24(5):673–683
Evans JS, Murphy MA, Holden ZA, Cushman SA (2011) Modeling species distribution and change using random forest. In: Drew CA (ed) Predictive species and habitat modeling in landscape ecology: concepts and applications. Springer, New York
Evans JS, Oakleaf J (2012) Geomorphometry & Gradient Metrics Toolbox (ArcGIS 10.0)
Evans JS, Oakleaf J, Cushman SA, Theobald DM (2014) An ArcGIS toolbox for surface gradient and geomorphometric modeling, version 2.0-0. Accessed:2015 Dec 2nd. http://evansmurphy.wix.com/evansspatial
Fuller AK, Harrison DJ (2005) Influence of partial timber harvesting on American martens in north-central Maine. J Wildl Manag 69:710–722
Godbout G, Ouellet JP (2008) Habitat selection of American marten in a logged landscape at the southern fringe of the boreal forest. Ecoscience 15:332–342
Grand J, Buonaccorsi J, Cushman SA, Griffin CR, Neel MC (2004) A multiscale landscape approach to predicting bird and moth rarity hotspots in a threatened pitch pine–scrub oak community. Conserv Biol 18(4):1063–1077
Grimm R, Behrens T, Märker M, Elsenbeer H (2008) Soil organic carbon concentrations and stocks on Barro Colorado Island—digital soil mapping using random forests analysis. Geoderma 146(1):102–113
Hargis CD, Bissonette JA, Turner DL (1999) The influence of forest fragmentation and landscape pattern on American martens. J Appl Ecol 36:157–172
Hargis CD, McCullough DR (1984) Winter diet and habitat selection of marten in Yosemite National Park. J Wildl Manag 48:140–146
Hegel TM, Cushman SA, Evans J, Huettmann F (2010) Current state of the art for statistical modelling of species distributions. In: Cushman SA, Huettman F (eds) Spatial complexity, informatics and wildlife conservation. Springer, Tokyo, pp 273–312
Liaw A, Wiener M (2002) Classification and regression by random. Forest R news 2(3):18–22
McGarigal K, Cushman SA, Ene E (2012) FRAGSTATS v4: Spatial Pattern Analysis Program for Categorical and Continuous Maps. Computer software program produced by the authors at the University of Massachusetts, Amherst. Available at the following web site: http://www.umass.edu/landeco/research/fragstats/fragstats.html
McGarigal K, Wan HY, Zeller KA, Timm BC, Cushman SA (2016) Multi-scale habitat modeling: A review and outlook. Landsc Ecol 31:1161–1175

Mi C, Huettmann F, Guo Y, Han X, Wen L (2017) Why to choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. Peerj 5:e2849

Murphy MA, Evans JS, Storfer A (2010) Quantifying *Bufo boreas* connectivity in Yellowstone National Park with landscape genetics. Ecology 91(1):252–261

Pontius RG Jr, Milones M (2011) Death to Kappa: Birth of quality disagreement and allocation disagreement for accuracy assessment. Int J Remote Sens 32:4407–4429

Pontius RG Jr, Parmentier B (2014) Recommendations for using the relative operating characteristic (ROC). Landsc Ecol 29:367–382

Pontius RG Jr, Si K (2014) The total operating characteristic to measure diagnostic ability for multiple thresholds. Int J Geogr Inf Sci 28:570–583

Pontius RG Jr, Walker R, Yao-Kumah R, Arima E, Aldrich S, Caldas M, Vergara D (2014) Accuracy assessment for a simulation model of Amazonian deforestation. Ann Assoc Am Geogr 97:677–695

R Development Core Team (2008) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Robinson L, Cushman SA, Lucid M (2017) Winter bait stations as a multi-species survey tool. Ecol Evol 7:6826–6838

Rodriguez-Galiano VF, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez JP (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J Photogramm Remote Sens 67:93–104

Samuel A. Cushman, Nicholas B. Elliot, Dominik Bauer, Kristina Kesch, Laila Bahaa-el-din, Helen Bothwell, Michael Flyman, Godfrey Mtare, David W. Macdonald, Andrew J. Loveridge (2018). Prioritizing core areas, corridors and conflict hotspots for lion conservation in southern Africa. July 5, https://doi.org/10.1371/journal.pone.0196213

Schneider A (2012) Monitoring land cover change in urban and peri-urban areas using dense time stacks of Landsat satellite data and a data mining approach. Remote Sens Environ 124:689–704

Shirk AS, Raphael MG, Cushman SA (2014) Spatiotemporal variation in resource selection: Insights from the American Marten (Martes americana). Ecol Appl 24:1434–1444

Svetnik V, Liaw A, Tong C, Wang T (2004) Application of breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. In: Roli F, Kittler J, Windeatt T (eds) Multiple classifier systems, lecture notes in computer science. Springer, Berlin/Heidelberg, pp 334–343

Thompson CM, McGarigal K (2002) The influence of research scale on bald eagle habitat selection along the lower Hudson River, New York. Landsc Ecol 17:569–586

Tomson SD (1999) Ecology and summer/fall habitat selection of American marten in northern Idaho. University of Montana. Thesis, Missoula, p 80

Wasserman TN, Cushman SA, Schwartz MK, Wallin DO (2010) Spatial scaling and multimodel inference in landscape genetics: *Martes americana* in northern Idaho. Landsc Ecol 25:1601–1612

Wasserman TN, Cushman SA, Wallin DO, Hayden J (2012a) Multi scale habitat relationships of *Martes americana* in northern Idaho, USA. Research Paper RMRSRP-94. USDA Forest Service, Rocky Mountain Forest and Range Experimental Station, Fort Collins

Wasserman TN, Cushman SA, Shirk AS, Landugth EL, Littell JS (2012b) Simulating the effects of climate change on population connectivity of American marten (Martes americana) in the northern Rocky Mountains, USA. Landsc Ecol. https://doi.org/10.1007/s10980-011-9653-8

Wiens JA (1989) Spatial scaling in ecology. Funct Ecol 3(4):385–397

Wilbert CJ, Buskirk SW, Gerow KG (2000) Effects of weather and snow on habitat selection by American martens (*Martes americana*). Can J Zool 78:1691–1696

Wynne KM, Sherburne JA (1984) Summer home range use by adult marten in northwestern Maine. Can J Zool 62:941–943