Wildl. Biol. 15: 352-369 (2009)
DOI: 10.2981/08-057

www.wildlifebiology.com

Original article

Internal validation of predictive logistic regression models for decision-making in wildlife management

Justin A. Gude, Michael S. Mitchell, David E. Ausband, Carolyn A. Sime & Edward E. Bangs

Predictive logistic regression models are commonly used to make informed decisions related to wildlife management and conservation, such as predicting favourable wildlife habitat for land conservation objectives and predicting vital rates for use in population models. Frequently, models are developed for use in the same population from which sample data were obtained, and thus, they are intended for internal use within the same population. Before predictions from logistic regression models are used to make management decisions, predictive ability should be validated. We describe a process for conducting an internal model validation, and we illustrate the process of internal validation using logistic regression models for predicting the number of successfully breeding wolf packs in six areas in the US northern Rocky Mountains. We start by defining the major components of accuracy for binary predictions as calibration and discrimination, and we describe methods for quantifying the calibration and discrimination abilities of a logistic regression model. We also describe methods for correcting problems of calibration and future predictive accuracy in a logistic regression model. We then show how bootstrap simulations can be used to obtain unbiased estimates of prediction accuracy when models are calibrated and evaluated within the same population from which they were developed. We also show how bootstrapping can be used to assess coverage rates and recalibrate the endpoints of confidence intervals for predictions from a logistic regression model, to achieve nominal coverage rates. Using the data on successfully breeding wolf packs in the northern Rocky Mountains, we validate that predictions from a model developed with data specific to each of six analysis areas are better calibrated to each population than a global model developed using all data simultaneously. We then use shrinkage of model coefficients to improve calibration and future predictive accuracy for the area-specific model, and recalibrate confidence interval endpoints to provide better coverage properties. Following this validation, managers can be confident that logistic regression predictions will be reliable in this situation, and thus that management decisions will be based on accurate predictions.

Key words: calibration, Canis lupus, confidence interval coverage, decision-making, discrimination, internal validation, logistic regression, northern Rocky Mountains

Justin A. Gude & Carolyn A. Sime, Montana Fish, Wildlife & Parks, 1420 E. 6th Avenue, Helena, Montana 59620, USA - e-mail addresses: jgude@mt.gov (Justin A. Gude); casime@mt.gov (Carolyn A. Sime)
Michael S. Mitchell & David E. Ausband, Montana Cooperative Wildlife Research Unit, 205 Natural Sciences Building, University of Montana, Missoula, Montana 59812, USA - e-mail addresses: michael.mitchell@mso.umt.edu (Michael S. Mitchell); david.ausband@mso.umt.edu (David E. Ausband)
Edward E. Bangs, US Fish and Wildlife Service, 585 Shepard Way, Helena, Montana 59601, USA - e-mail: ed_bangs@fws.gov

Corresponding author: Justin A. Gude

Received 11 August 2008, accepted 3 July 2009

Associate Editor: Nigel G. Yoccoz

Logistic regression models are developed often in wildlife management and management-related research (Keating & Cherry 2004, Johnson et al. 2006, Guthery & Bingham 2007). A common use of these models is to make informed decisions, where models generated by data collected in the past are used to

make predictions. Numerous applications of predictive logistic models are related to wildlife habitat, such as predicting favourable habitat for species (Mladenoff et al. 1995, O'Brien et al. 2005), predicting critical habitat for endangered species (Turner et al. 2004), and predicting species occurrence (Scott et al. 2002). Other common applications include estimating vital rates for use in predictive models of population dynamics (White 2000, Williams et al. 2002:143-161 and 343-347).

Applications of logistic models include making predictions about future observations from the same study area or study population. Using models generated with historical data to predict future events can be problematic. Though models may fit historical data well, future circumstances may not be sufficiently similar to allow for reliable predictions from the logistic model. A danger also exists that a model is overfit, which leads to biased predictions (Harrell 2001:60-64, Randin et al. 2006). Furthermore, nothing assures that estimates of uncertainty for model predictions will accurately reflect uncertainty in the future. The ability of logistic regression models to make good predictions should be evaluated before they are used in practice. Otherwise, such models may lead to biased predictions and misguided management decisions.

Fitting a model to data to estimate parameters and test hypotheses about processes that may account for observed patterns in a data set differs fundamentally from developing a model intended to make accurate predictions about the future (Copas 1983). To determine reliability of a model for making predictions with new data, an unbiased evaluation is needed to determine the prediction error rate, and identify shortcomings in a model that lead to poor predictions (Miller et al. 1991). When this type of evaluation is conducted for a model that will be used to make predictions for the same population from which it was developed, the process is called internal validation (Harrell et al. 1996).

In our paper, we describe a process for internally validating a predictive logistic regression model, and illustrate the process with an example based on the estimation of successfully breeding wolf packs (BP) in the northern Rocky Mountains (NRM; Mitchell et al. 2008). A wolf BP has been legally defined as at least two adult wolves and two pups from the same pack that survive until December 31 of the year of reproduction (USFWS 1994). Mitchell et al. (2008) developed logistic regression models designed to estimate the number of BPs in the NRM based on the sizes of wolf packs, using data collected through intensive monitoring from 1981 to 2005. Recovery criteria for wolves in the NRM require monitoring BPs (USFWS 1994). The capacity to monitor the number of BPs intensively is expected to diminish after federal delisting due to reduced funding in state wildlife agencies, so these models were developed to assist wolf managers in predicting the number of BPs in the NRM using more easily obtained information on pack sizes. Mitchell et al. (2008) found that the relationship between pack size and the probability of a pack containing a BP varied across six analysis areas within NRM, and was influenced by levels of human-caused mortality and wolf population dynamics unique to each area. Mitchell et al. (2008) argued that a global model that made use of all data from all areas to develop a predictive equation would not predict the number of BPs accurately across the NRM. The logistic BP models which they developed were fit to historical data and were intended to make predictions for the same population using future data. However, the relative predictive abilities of area-specific and global models were only evaluated using the same data that were used to develop the models, and the predictive ability of the models was not validated. We conducted an internal validation of the predictive abilities of the models presented in Mitchell et al. (2008).

Methods

Example context

Mitchell et al. (2008) developed two predictive logistic regression models for estimating the number of BPs in each of six areas when the number and size of wolf packs in those areas are known. One of these models uses a common intercept and slope term to estimate the number of BP (global model), while the other uses intercept and slope terms unique to each area to estimate the number of BP (area-specific model, Table 1, Fig. 1). Both of these models can be used to rank the relative probabilities that multiple wolf packs are BPs, and to estimate the total number of wolf BPs in an area, $\hat{T}$, using:

$$\hat{T} = \sum_{i=4}^{k} (n_i{}^{*}\hat{\pi}_i),$$

where n_i is the number of wolf packs of a given size (for $i \geq$ four wolves in order to meet the legal BP

Table 1. Estimated model coefficients and SEs for global and area-specific predictive logistic regression models presented in Mitchell et al. (2008). Model parameters are on the logit scale and were estimated using maximum likelihood. Idaho was used as the reference area for generating the global model and area-specific models. Parameter estimates for area-specific models other than Idaho represent adjustments to the intercept and slope for the Idaho reference model, e.g. the model for NW Montana would be (-0.90 -1.19) + ((0.38 + 0.10) * pack size). SW Montana-CIEPA refers to the Central Idaho Experimental Population Area, as it overlaps into Montana. SW Montana-GYEPA refers to the Greater Yellowstone Experimental Population Area, as it overlaps into Montana.

		Global model		Area-specific model	
Parameter	Analysis area	Parameter estimate	SE	Parameter estimate	SE
Reference intercept	Idaho	-1.7	0.40	-0.90	0.75
Area-specific intercept adjustment	NW Montana	.	.	-1.19	1.12
	SW Montana-CIEPA	.	.	-5.41	3.59
	SW Montana-GYEPA	.	.	-1.61	1.78
	Wyoming	.	.	-1.16	1.54
	Yellowstone NP	.	.	-0.78	1.14
Reference slope, pack size	Idaho	0.43	0.06	0.38	0.11
Area-specific slope adjustment	NW Montana	.	.	0.10	0.17
	SW Montana-CIEPA	.	.	0.84	0.68
	SW Montana-GYEPA	.	.	0.12	0.29
	Wyoming	.	.	0.09	0.23
	Yellowstone NP	.	.	-0.01	0.16

definition), $\hat{\pi}_i$ is the predicted probability that a pack of size i is a BP from the logistic model, and the summation is over all observed pack sizes greater than three wolves. The confidence interval (CI) estimators for probabilities that wolf packs are BPs presented in Mitchell et al. (2008) used the standard normal method as an approximation for binomial data on the logit scale, then back-transformed CI endpoints to the probability (0-1) scale. This method is known to have coverage rates that do not match the 100*(1 - α) nominal coverage rate, where α is defined as the type I error rate, particularly for estimated probabilities closer to 0 and 1 than to 0.5 (Jennings 1987). However, we chose to evaluate this method because Mitchell et al. (2008) recommended its use with the BP estimators.

The model must be able to discriminate among packs with high and low probabilities of being a BP, have well-calibrated predicted probabilities across the range of predictions, and have CIs for the number of BPs in each area with accurate endpoints. In this case, two CIs are relevant. The two-sided 95% CI will provide a standard and interpretable measure of overall precision in the estimate for each area. One-sided 95% lower CIs can be used for management purposes to provide a level of confidence as to whether management activities have reduced the number of wolf BPs below recovery criteria, given uncertainty in the point estimates. If the number of BPs is reduced below recovery criteria, management authority for wolves will be transferred from state agencies to the federal government.

Mitchell et al. (2008) recommended that the area-specific model be used for prediction rather than the global model. We quantified the merits of this recommendation based on validation of predictive abilities of the two models. Our validation consisted of three steps: 1) choosing between global and area-specific models based on prediction accuracy in each analysis area, 2) improving calibration and predictive ability of the selected model across the range of predicted probabilities, and 3) calibration of the CI endpoints of the selected model to match desired coverage rates.

Internal validation of logistic models

Our method for internally validating the accuracy of predictive logistic regression models consists of five steps (Fig. 2). The first two steps are aimed at estimating two different measurements of accuracy, calibration and discrimination, and the next three steps involve correcting for bias in estimates of accuracy, improving accuracy, and improving confidence interval coverage rates.

Accuracy of point estimates from logistic models can be divided conceptually into calibration and discrimination. Calibration describes whether predicted probabilities are too high or too low relative to true population values, whereas discrimination refers to the correct relative ranking of predicted probabilities (Justice et al. 1999, Pearce & Ferrier 2000). For example, a perfectly calibrated model might predict a survival probability of 0.2 for a given animal when 0.2 of the animals in the population with the

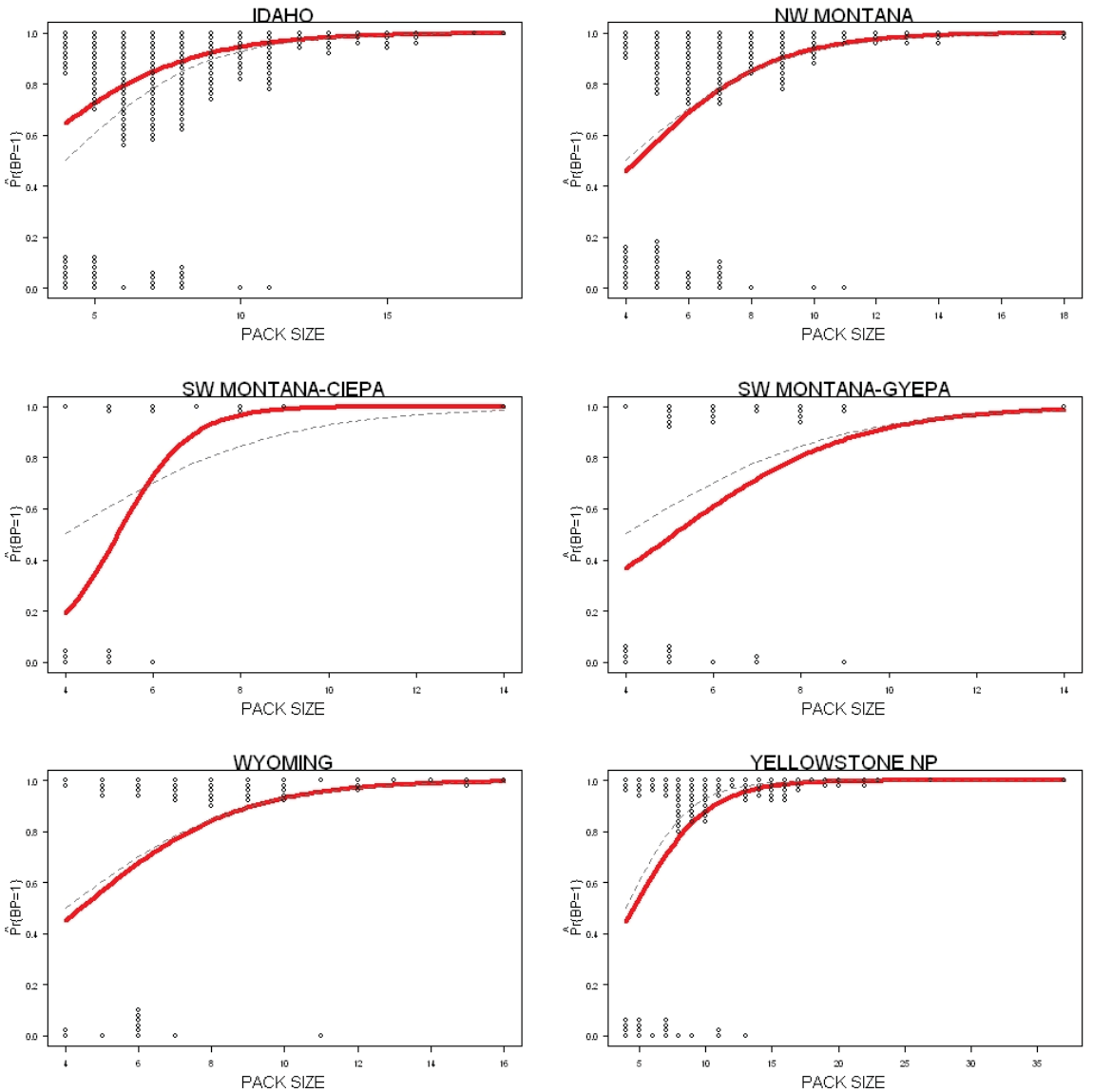

Figure 1. Predicted probabilities that a wolf pack contains a successful breeding pair (BP) in the northern Rocky Mountains. Predictions are from the global (dashed line, repeated on every panel) and area-specific (solid lines) models fit by maximum likelihood, as in Table 1. The dot plots across the top and bottom of the panels represent the distribution and quantity of '1' and '0' data used to fit the models in each area, respectively. SW Montana-CIEPA and the SW Montana-GYEPA refer to the Central Idaho experimental population area and Greater Yellowstone experimental population area, respectively, as they overlap into Montana.

same characteristics actually survive. The same model might be poorly calibrated at higher predicted probabilities if it predicts a survival probability of 0.9 for a given animal when 0.5 of the animals in the population with the same characteristics actually survive. This model would, however, have accurate discrimination ability as higher probabilities of survival are predicted for animals with higher observed survival rates. Methodologies for more precisely quantifying calibration and discrimination accuracy are described in the following.

Calibration

Methodology for assessing overall average calibration of a logistic model was introduced by Cox (1958), and consists of developing a logistic model using one data set. Calibration of the model can be assessed using an independent data set of observed values. Cox (1958) suggested modeling the observed independent values as a function of their predicted values, $\hat{\pi}_i$, obtained by applying the fitted model to the independent data, using another logistic regression. The two estimated parameters in this model, β_0

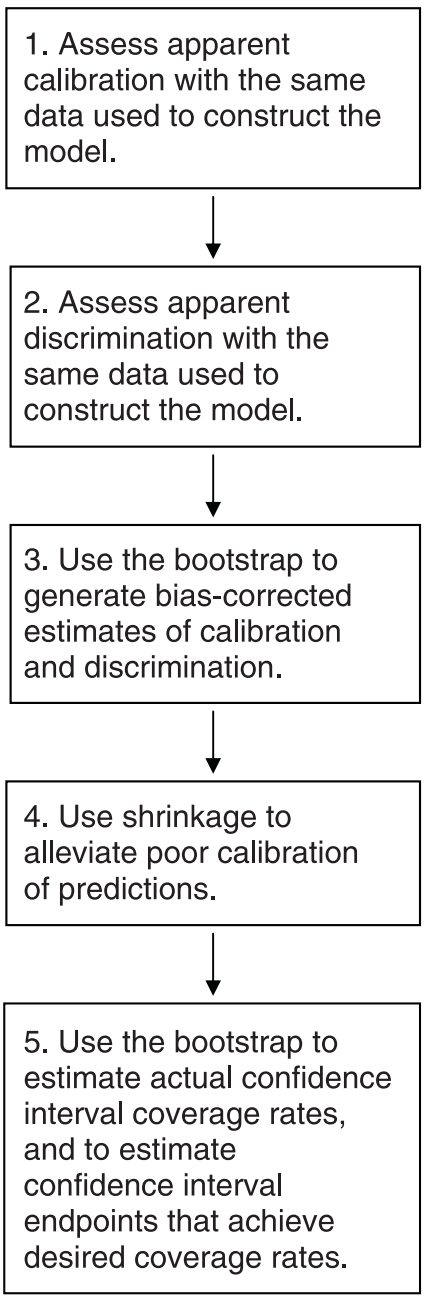

Figure 2. Flow diagram illustrating the steps of interval validation of predictive logistic regression models that are described in the text. Calibration and discrimination are measures of predictive accuracy and are defined in the text.

and β_1, are useful as measures of calibration of the model to the independent data set, specifically to examine the hypothesis that the observed proportions in the independent data set are equal to the predicted probabilities from the original model (i.e. $\Pr(Y_i = 1) = \hat{\pi}_i$). The slope, β_1, is a measure of the direction and spread of the predicted probabilities. When $\beta_1 = 1$, the $\hat{\pi}_i$ are generally correct. If $\beta_1 > 1$, the $\hat{\pi}_i$ show the correct direction but do not vary enough. If $0 < \beta_1 < 1$, the $\hat{\pi}_i$ vary too much. If $\beta_1 < 0$, the $\hat{\pi}_i$ show the wrong direction, and if $\beta_1 = -1$, the $\hat{\pi}_i$ are exact complements of the true probabilities (Pearce & Ferrier 2000). The intercept, β_0, is a measure of the overall calibration of the model if $\beta_1 = 1$, and of the calibration at $\hat{\pi}_i = 0.5$ if $\beta_1 \neq 1$, because the Cox model assumes β_0 is a function of β_1. If $\beta_0 = 0$, the $\hat{\pi}_i$ are generally correct, the $\hat{\pi}_i$ are too low if $\beta_0 > 0$, and the $\hat{\pi}_i$ are too high if $\beta_0 < 0$ (Pearce & Ferrier 2000).

The methods created by Cox are useful measures of the average calibration of the model. However, as with any regression, the fitted model may be correct on average even though there are areas within the range of predictor variables where the model does not fit well. χ^2-type tests have been developed to test the average calibration for logistic regression models (Hosmer & Lemeshow 2000: 147-156). These methods do not necessarily illustrate what range(s) of predicted probabilities have poor fit to the observed data (Hosmer & Lemeshow 2000:151). An accurate fit to the data across the range of predicted probabilities from a logistic regression is often required, and the ranges of predicted probabilities and covariates for which the fit of the model poorly matches the observed data are of interest. Hosmer & Lemeshow (2000:167-186) provide diagnostic measures for the leverage of individual covariate patterns on model parameter estimates and goodness of fit statistics to identify covariate patterns that lead to poor model fit. Focusing instead on the calibration of predicted probabilities, Harrell et al. (1996) identify smoother functions on scatterplots of observed vs predicted probabilities as simple, graphical measures of the calibration of model predictions. One function that is readily available in many statistical packages is locally-weighted regression, the LOWESS smoother (Cleveland 1979, 1981). This function provides a visual representation of how well the predicted probabilities match the observed data. Because both average model calibration and adequate calibration across the range of predicted probabilities are necessary for predicting wolf BPs, we applied both the Cox method and a LOWESS smoother to assess calibration of the global and area-specific models for predicting wolf BPs. We used the Design package (Harrell 2005) in R 2.5.0 (R Development Core Team 2007) to conduct these analyses.

If problems of calibration are identified using the Cox regression or the LOWESS smoother approaches, the analyst has several options for addressing the problems, including shrinkage of predictions. Shrinkage describes the degree to which a validation fit falls short of the original model fit (Copas 1983), or equivalently the flattening of the predicted vs observed plot away from the ideal 45° line caused by overfitting (Harrell et al. 1996). Overfitting is a characteristic common to maximum likelihood estimation, particularly for small data sets where estimated regression coefficients can be influenced by noise (Copas 1983, van Houwelingen & Le Cressie 1990). Overfitting refers to the problem of estimated regression coefficients that are too large, which causes model predictions to be too extreme for future data (Steyerberg et al. 2000). Closely related to the concept of overfitting is the concept of regression to the mean, in which low predictions will be too low, high predictions will be too high, and predictions closer to the overall mean will be more accurate in future data (Efron & Morris 1977, Harrell 2001: 62). Overfitting of model parameters and regression to the mean can be addressed by 'shrinking' model coefficients toward zero (Hastie et al. 2003:55-75).

For models containing multiple slope parameters, certain parameters can lead to poorer fit than others, and these parameters can be targeted for shrinkage more than others to improve predictive ability (Harrell 2001:64). Penalized maximum likelihood estimation of model parameters is often used for this purpose, and it consists of adding a penalty term to the likelihood function relative to the size of the coefficients in the logistic model (Harrell 2001: 64). Firth (1993, 2006) demonstrated that for logistic regression, this penalty is equivalent to Jeffreys (1946) invariant prior, and can be imposed iteratively in the maximization of the likelihood function, with larger penalties added to observations that have more influence. Because each observation has more influence on parameter estimates in smaller data sets, parameter estimates will be shrunken more when there are less data available. Prediction bias in maximum likelihood model estimates is highest in smaller data sets (Efron & Morris 1977, Harrell 2001:60-61). Hence, the shrinkage method presented by Firth (1993) is commonly used to reduce prediction bias by shrinking model coefficients. The package brlr (Firth 2006) for R (R Development Core Team 2007) provides access to this method.

Discrimination

A widely accepted measure of discrimination ability of a predictive model is the c index (for concordance), which applies to predictions that are continuous, dichotomous, ordinal, and censored time-to-event outcome predictions (Harrell et al. 1996). In binary cases, c is equivalent to the area under the Receiver Operating Characteristic (ROC) curve, which is a common method of measuring the predictive ability of logistic regression models (Harrell et al. 1996, Fielding & Bell 1997, Hosmer & Lemeshow 2000:160-164). We used c to measure the discrimination ability of both the global and area-specific models to predict wolf BPs. We used the Design package (Harrell 2005) in R 2.5.0 (R Development Core Team 2007) to conduct these analyses.

Unlike calibration, poor discrimination ability of a particular model cannot be corrected analytically. For this reason, discrimination ability should be a focus of predictive model selection (Harrell et al. 1984). Hosmer & Lemeshow (2000:162) suggested a sensible model should have $c > 0.7$, as $c = 0.5$ is the discrimination ability that would be expected with random guessing.

Internal validation based on calibration and discrimination

For assessing measures of calibration and discrimination, the apparent accuracy of the model measured by calculating accuracy measures on the exact data set that was used to develop the model, is overly optimistic even for new observations from the same population (Copas 1983, Steyerberg et al. 2001). Several options for conducting an internal validation procedure have been developed and compared for logistic regression, including k-fold cross-validation, the jackknife, split-sample validation and bootstrapping (Lobo et al. 2008, Verbyla & Litvaitis 1989, Boyce et al. 2002, Steyerberg et al. 2001). The most efficient procedure for logistic regression is bootstrapping, because it makes use of the full data set for training and testing models while providing estimates of prediction error with relatively low variability and minimal bias (Harrell 2001:95, Steyerberg et al. 2001). The bootstrap validation algorithm consists of (Efron & Tibshirani 1993:248-249): 1) Estimate the apparent predictive ability (A) using the original sample used to fit the model, 2) Draw B bootstrap samples from the original sample, 3) For each bootstrap sample, fit the model to the bootstrap sample and measure the apparent predictive ability (a), 4) Test the accuracy measure by applying the

bootstrap model to the original sample and measuring the accuracy (t), 5) Calculate the optimism (o) in predictive ability for this bootstrap model as o = a - t, 6) Obtain a stable estimate of optimism as the mean optimism from the B bootstrap samples

$$O = \frac{\sum_{i=1}^{B} o_i}{B},$$

and 7) Obtain an internally validated estimate of predictive accuracy by subtracting the estimated optimism from the apparent predictive ability: V = A - O.

This algorithm can be used to internally validate any measure of model calibration or discrimination. For example, A might be defined as the intercept term in the Cox (1958) measure of calibration described above. Then, a and t would measure the same quantity in the individual bootstrap samples and the test (original) sample, o would measure the optimism in the models developed from the individual bootstrap samples, and O would measure the optimism of the model calibration, as measured by the intercept term in the Cox method, in the underlying population that generated the sample. Note that if this algorithm is used, the Cox (1958) and other measures of calibration or discrimination can be estimated without an independent data set. The bootstrap can also be used to estimate uncertainty in the internally validated estimate of predictive accuracy. For example, bootstrap percentile CIs (Efron & Tibshirani 1993:170-174) for V can be obtained by repeating the above algorithm in another bootstrap resampling process, and using the desired percentiles from the resulting distribution of O. In our example, we applied this bootstrap algorithm to validate measures of calibration and discrimination in each analysis area, including the Cox regression parameters, the LOWESS smoother, and c for both the global and area-specific models. We used the Design package (Harrell 2005) in R 2.5.0 (R Development Core Team 2007) to conduct these analyses.

Assessing CI Coverage Rates

Probabilities predicted by logistic models are point estimates, with uncertainty commonly depicted using CI estimation. CIs provide a measure of certainty that a population value, μ, falls in a range specified by the interval, I. Intervals are defined based on an acceptable level of error, α, such that $\Pr\{\mu \in I\} = 1\text{-}\alpha$, where $100*(1 - \alpha)$ defines the nominal coverage percent (Thompson 2002:29). However, when a CI is calculated, actual coverage rates for the intervals may not be specified by $100*(1 - \alpha)$, as the CI endpoints can be biased. For standard normal CIs, this bias decreases at a rate of $\frac{1}{\sqrt{n}}$, where n is the sample size used to estimate the interval (Efron & Tibshirani 1993:187).

A strategy for assessing actual coverage rates for CI estimators that includes both estimating and calibrating the accuracy of CI coverage rates makes use of the bootstrap. The desired probability characteristics of CI endpoints are

$$\text{Two-sided CI}: \Pr\left\{\theta \le \hat{\theta}_L\right\} = \frac{\alpha}{2},\ \Pr\left\{\theta \le \hat{\theta}_U\right\} = 1 - \frac{\alpha}{2} \text{ and}$$

$$\text{One-sided lower CI}: \Pr\left\{\theta \le \hat{\theta}_L\right\} = \alpha,$$

where θ is the true population value, $\hat{\theta}$ is the point estimator of that value, and $\hat{\theta}_L$ and $\hat{\theta}_U$ are estimators of the lower (L) and upper (U) CI endpoints for θ. Based on these definitions, it may be possible to construct a CI using a value $\lambda \neq \alpha$ that gives the desired coverage rate with error level α (Efron & Tibshirani 1993:264). Selection of λ is accomplished with the following bootstrap algorithm (Efron & Tibshirani 1993:263-266): 1) Generate B bootstrap samples from the original sample, 2) Compute the λ-level confidence point, $\hat{\theta}^*_\lambda(b)$, for a grid of λ values that includes α for each bootstrap sample, 3) For each λ, calculate the probability that θ will be missed by the interval endpoint using

$$\Pr_*\left\{\hat{\theta} \le \hat{\theta}^*_\lambda(b)\right\} = \frac{\sum_{i=1}^{B} C_i}{B},$$

where $C_i = 1$ if $\hat{\theta} \le \hat{\theta}^*_\lambda(b)$ and $C_i = 0$ otherwise, $\hat{\theta}$ is the estimated probability from the predictive logistic model, and the summation is over the B bootstrap samples, and 4) Find the value of λ satisfying the desired level of error, α.

Note that because the range of values for λ includes α, the results of the above algorithm also contain information about the calibration of the CI endpoint. If it is well calibrated, then the bootstrap algorithm will choose $\lambda = \alpha$. If the procedure for estimating CIs is not well calibrated, then a solution is offered for a correctly calibrated procedure. CIs with bias decreasing at a rate of $\frac{1}{\sqrt{n}}$ that are calibrated using this procedure will have bias decreasing at a rate of $\frac{1}{n}$ (Efron & Tibshirani 1993:268). In our example, following comparisons of the predictive accuracy of the global and area-specific models to

predict wolf BPs and selection of one model to use for predictions with future data, we used this bootstrap algorithm to calibrate two-sided and one-sided lower CI endpoints for estimates of the number of BPs in each analysis area. We used the boot package (Canty & Ripley 2006) in R 2.5.0 (R Development Core Team 2007) to conduct these analyses.

All analyses were done in R 2.5.0 (R Development Core Team 2007). In addition to the analysis packages mentioned above, we also made use of the MASS (Venables & Ripley 2002) and stats (R Development Core Team 2007) packages.

Results

Choosing between global and area-specific models

The area-specific model and global models had similar internally validated discrimination ability according to the bootstrap validation algorithm (Fig. 3). In general, c ranged within 0.7-0.8 for both models. This indicates acceptable discrimination ability for both models (Hosmer & Lemeshow 2000:162). Bootstrap validation indicated that the ranges of predicted probabilities by both the global and area-

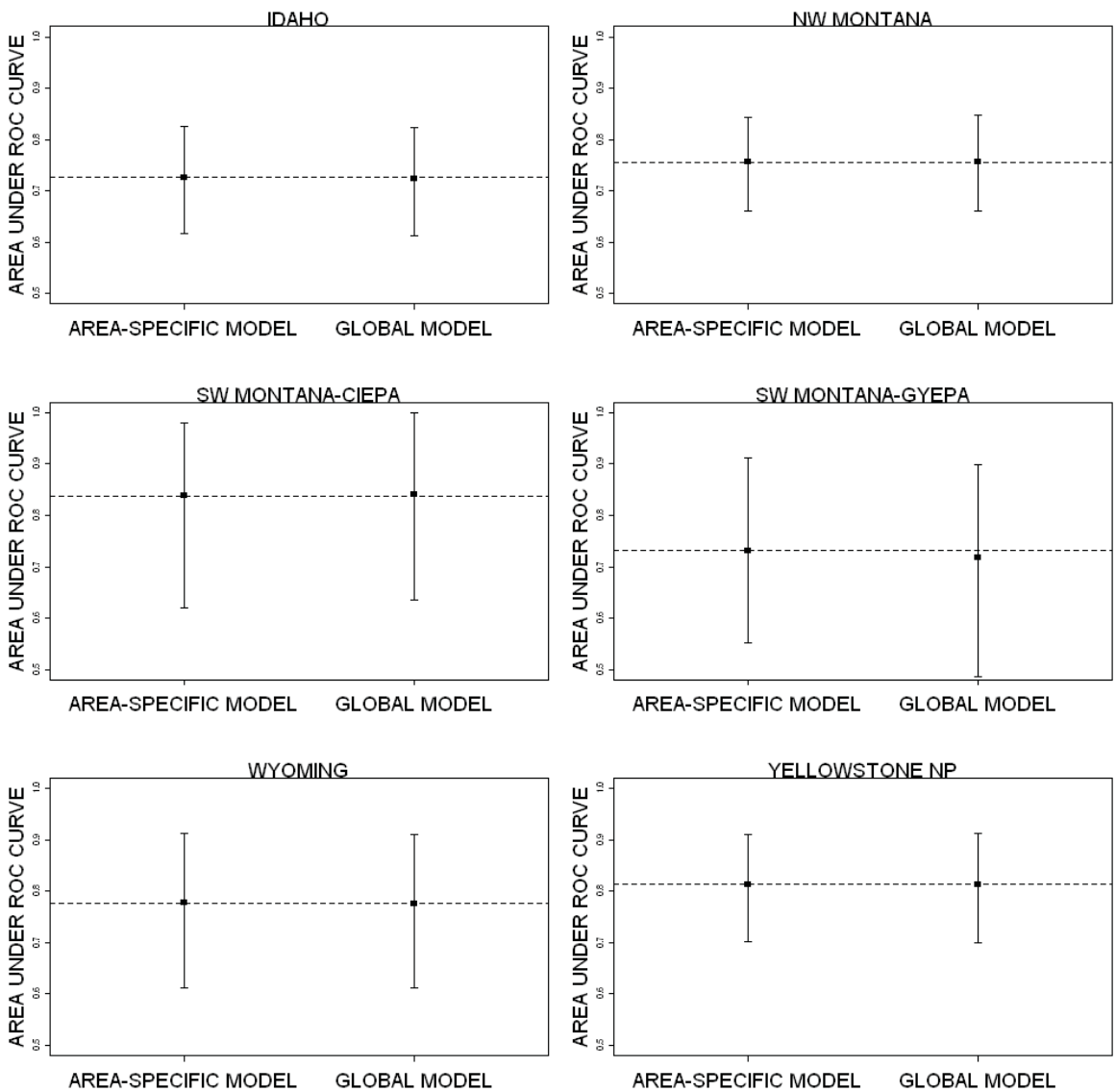

Figure 3. Internally validated discrimination ability for the global and area-specific models presented in Mitchell et al. (2008), as measured by c (or equivalently, the area under the ROC curve). Dashed line is a visual reference of c for the area-specific model. Bars represent 95% bootstrap percentile CIs.

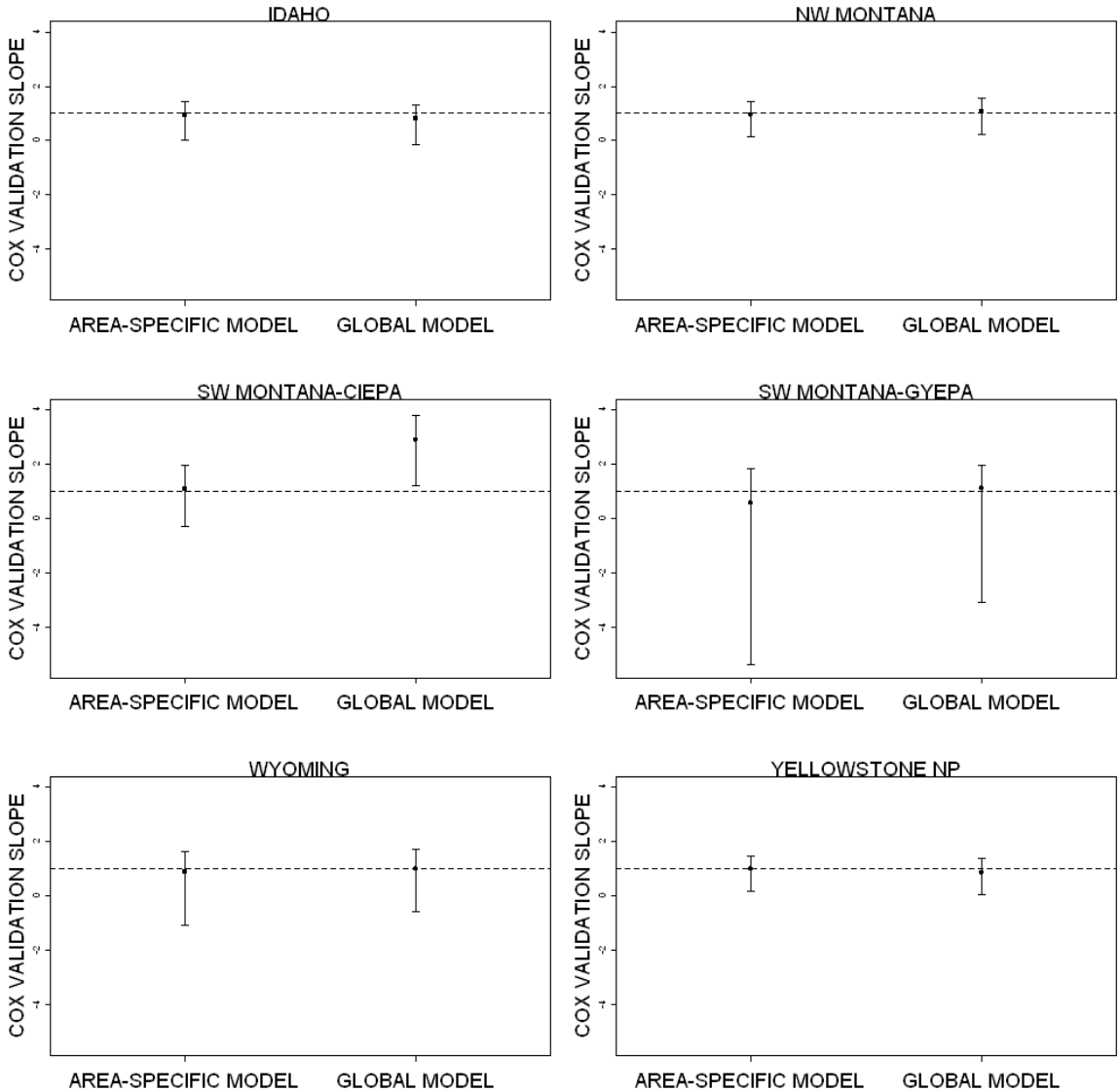

Figure 4. Internally validated calibration for the global and area-specific models presented in Mitchell et al. (2008), as measured by the Cox regression slope. Dashed line is the ideal value of 1. Bars represent 95% bootstrap percentile Cis.

specific models, as indexed by the Cox regression slope, were similar and close to 1 in each area, suggesting good calibration (Fig. 4). However, the range of predicted probabilities in one area in SW Montana was too narrow under the global model, as indicated by a slope >1 in the Cox regression. Bootstrap validation indicated that overall calibration of the area-specific model, as measured by the Cox regression intercept term estimate, was generally closer to 0 than the estimates for the global model (Fig. 5). The global model had overall predicted probabilities that were considerably too low for Idaho (intercept > 0) and considerably too high (intercept < 0) for the two areas in SW Montana.

Calibration of selected predictive model

The internally validated scatterplot smoother of the observed vs predicted probabilities showed that Idaho, NW Montana and Yellowstone NP had remarkably well-calibrated model predictions from the area-specific model (Fig. 6). There were ranges of the data that were not well-calibrated in the Greater Yellowstone Experimental Population Area in SW Montana and Wyoming (see Fig. 6). Finally, although model predictions seemed well-calibrated to observations in the Central Idaho Experimental Population Area in SW Montana, sample sizes for model development were small, increasing concerns about regression to the mean for future predictions

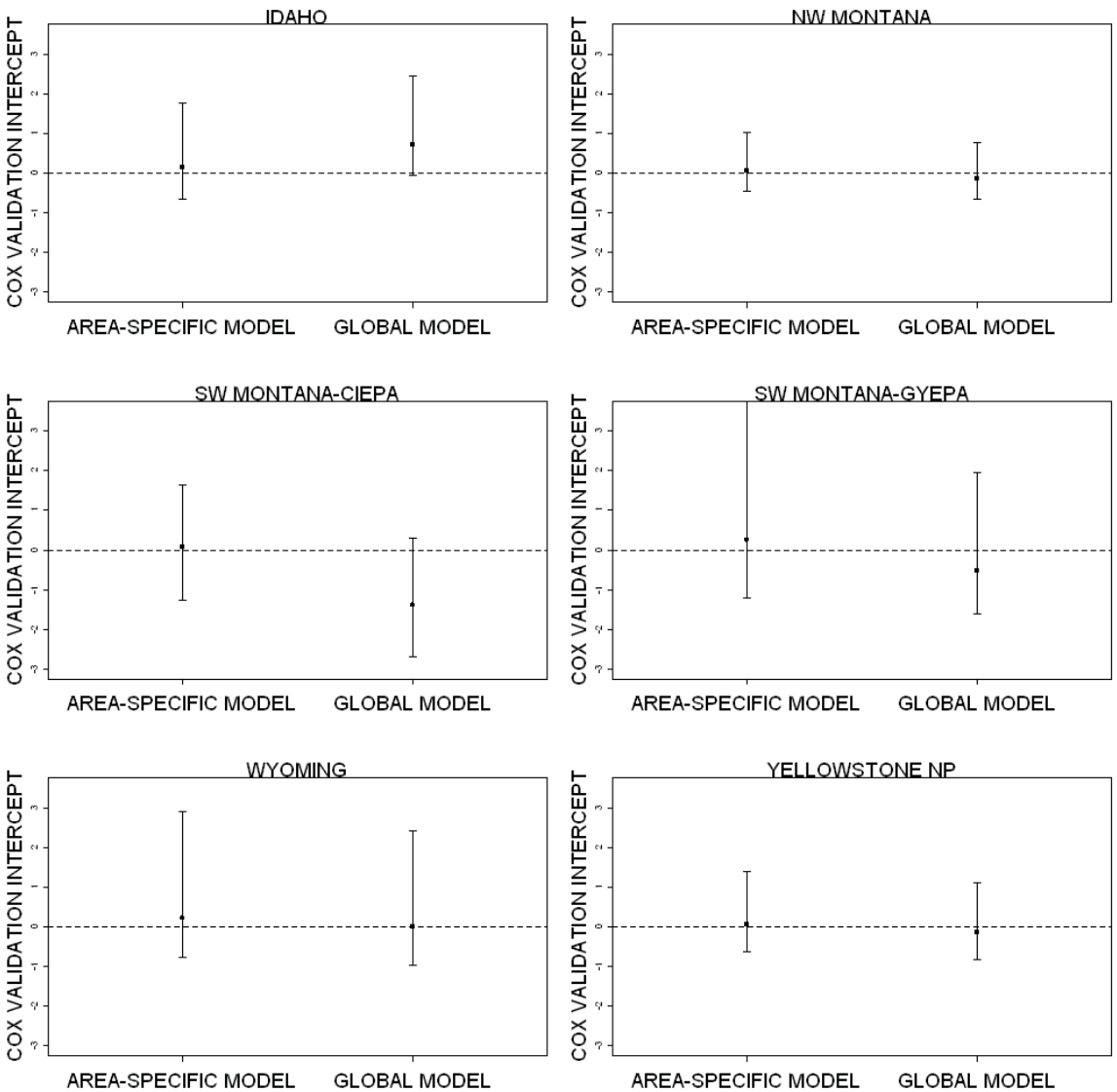

Figure 5. Internally validated calibration for the global and area-specific models presented in Mitchell et al. (2008), as measured by the Cox regression intercept. Dashed line is the ideal value of 0. Bars represent 95% bootstrap percentile CIs.

(see Fig. 6). Penalized maximum likelihood shrinkage resulted in many model predictions that were closer to the ideal 45° line in the two SW Montana and the Wyoming areas (Fig. 7), and smaller model parameter estimates (Table 2). Some of the higher predicted probabilities in these areas were also shrunken to lower values, which were closer to the global mean predictions and further from the ideal 45° line, consistent with regression to the mean (see Fig. 7). This pattern was more evident in the SW Montana areas than in the Wyoming area. Shrunken model predictions in the Idaho, NW Montana, and Yellowstone NP closely matched the maximum likelihood estimates for these areas (see Fig. 7).

Calibration of CI endpoints

CI coverage rates for both the two-sided and one-sided lower CI on predicted probabilities from the shrunken, area-specific model were higher than the nominal rates using the standard normal procedure from Mitchell et al. (2008). This is illustrated by the bootstrap procedure identifying that $\lambda > \alpha$ is necessary to provide nominal coverage rates for all three CI points (see Table 2, Fig. 8). For the one-sided lower and two-sided upper CI endpoints, λ was approximately 2-3 times larger than the desired α level, while for the two-sided lower CI endpoint, λ was approximately 3-5 times larger than the desired α level (see Table 2).

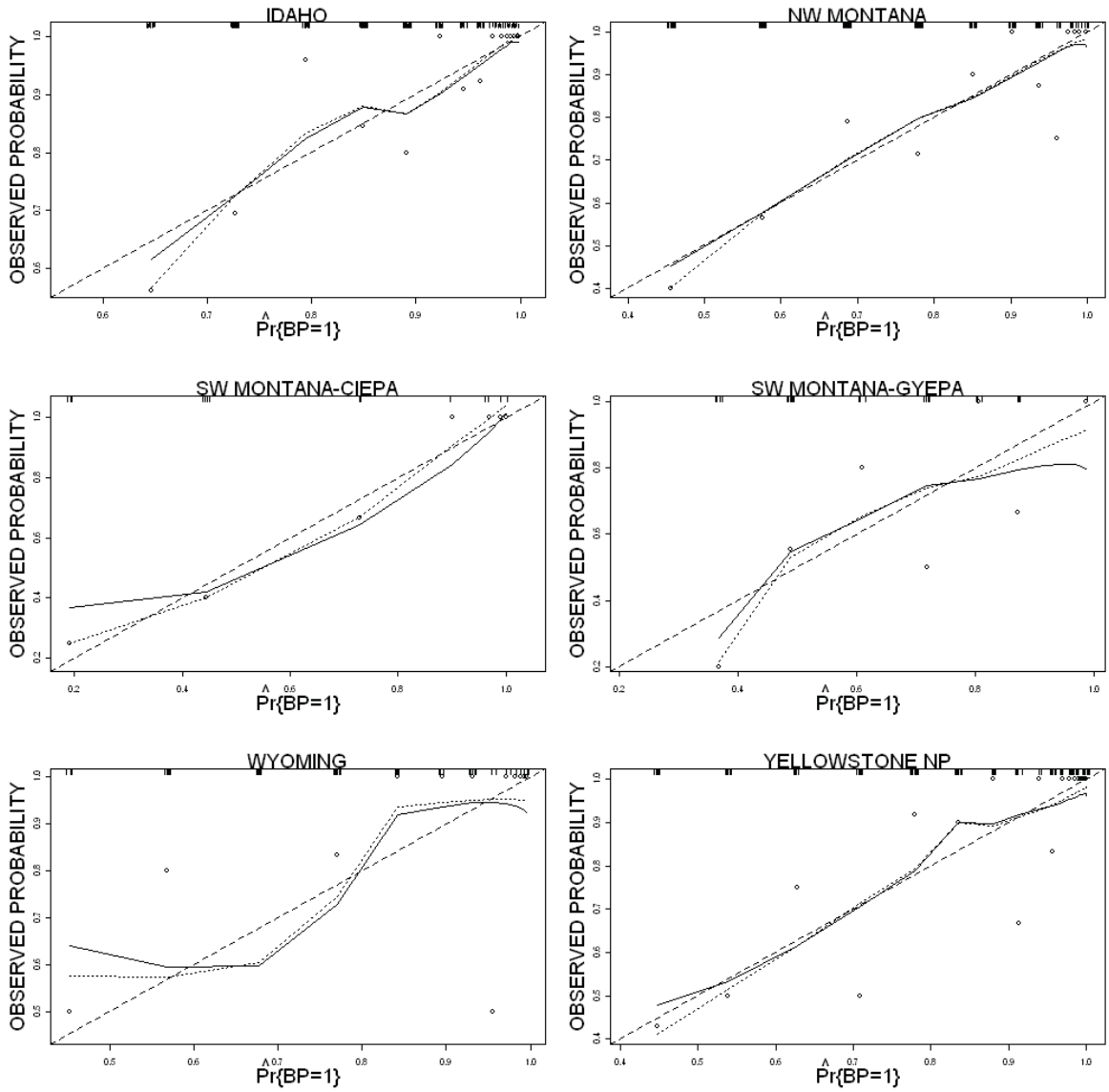

Figure 6. Internally validated calibration for the area-specific model presented in Mitchell et al. (2008), as measured by a LOWESS smoother on a scatterplot of observed vs predicted values. The rug plots across the top of the figures show the distribution and quantities of data used to fit the model in each area. The dotted line is the apparent calibration curve (fit by the LOWESS smoother), and the solid line is the bias-corrected calibration curve (i.e. validated calibration curve). The dashed line is the ideal 45-degree line.

Discussion

We have described and demonstrated procedures for validating a logistic model that will be used to make predictions in the future for the same underlying population that generated the sample data. An accurate predictive logistic model is one that is able to discriminate between high and low probability observations, one that produces predicted probabilities close to the observed probabilities in the population, and one that has accurate measures of uncertainty. In our example, we were able to show that an area-specific model for estimating wolf BPs in the NRM had superior accuracy compared to a global model. Mitchell et al. (2008) recommended that area-specific models be used to make predictions about the BP probabilities in each area, because the processes that generated the data in each area were too different to be captured by a global model. Our analysis reinforced this recommendation in that the predictions from the area-specific model are better calibrated to the populations in each area than are the predictions from the global model. In both models, the only predictor of BP status for wolf

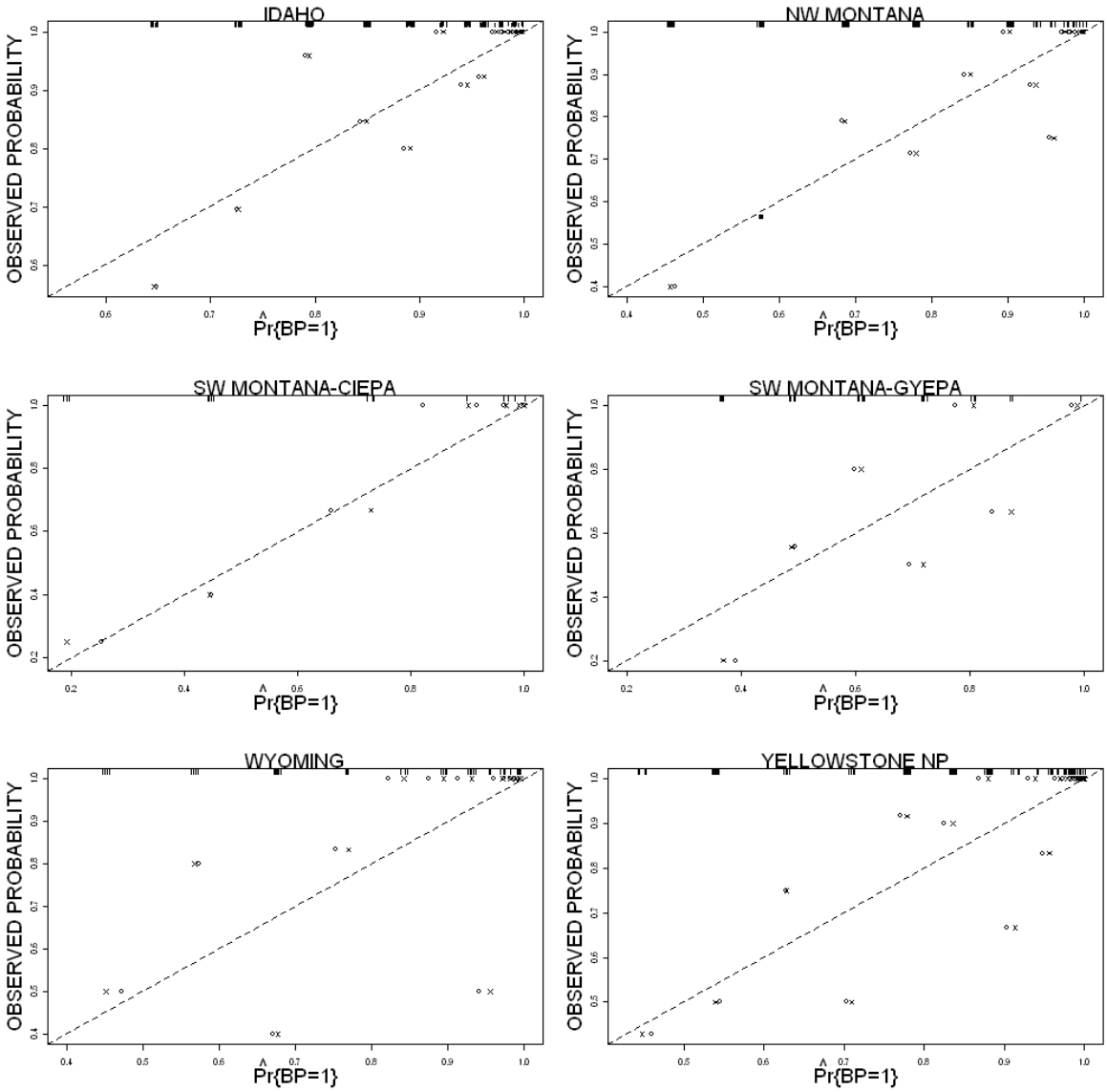

Figure 7. Predictions for each area from the area-specific model fit by maximum likelihood and the shrunken area-specific model (fit by penalized maximum likelihood). The rug plots across the top of the figures show the distribution and quantities of data used to fit the model in each area. X's are the predictions from the original maximum likelihood model, and open circles are the predictions from the shrunken model. Dashed line is the ideal 45-degree relationship.

packs was the size of individual packs. In the area-specific model, the estimated relationship was unique to each analysis area, whereas in the global model, the estimated relationship was the same in every analysis area. Mitchell et al. (2008) discussed other biological variables that likely influenced the relationship between BP and pack size in each analysis area. Therefore, there may be more complex models that would further improve the predictive accuracy, because they would more fully capture the biological processes in each area. When and if these models are developed, the methods presented in this paper could be used to validate their predictive accuracy. The purpose of our paper was not to develop or evaluate more complex, biological-process based models. It was rather to demonstrate and apply a method for evaluating the predictive accuracy of existing models, when these models are to be used to make decisions for wildlife conservation and management programs. Further, through detailed investigation of area-specific model calibration, shrinkage of model coefficients, and recalibration of the CI α-

Table 2. Model coefficients and CI error rates (α) to achieve 95% nominal coverage, for the maximum likelihood and penalized maximum likelihood area-specific models to predict wolf BPs in the northern Rocky Mountains presented in Mitchell et al. (2008). Model parameters are on the logit scale. Idaho was used as the reference area for generating the area-specific models. Parameter estimates for area-specific models other than Idaho represent adjustments to the intercept and slope for the Idaho reference model, e.g. the ML estimates for NW Montana would be (-0.90 -1.19) + ((0.38 + 0.10) * pack size). ML estimates = Maximum likelihood estimates with back-transformed, standard normal CIs. PML estimates = Penalized maximum likelihood estimates with calibrated back-transformed, standard normal CIs. SW Montana-CIEPA refers to the Central Idaho Experimental Population Area, as it overlaps into Montana. SW Montana-GYEPA refers to the Greater Yellowstone Experimental Population Area, as it overlaps into Montana.

			Model	
	Parameter	Management area	ML estimates	PML estimates
Model coefficient vector	Reference intercept	Idaho	-0.90	-0.81
	Area-specific intercept adjustment	NW Montana	-1.19	-1.15
		SW Montana-CIEPA	-5.41	-3.74
		SW Montana-GYEPA	-1.61	-1.31
		Wyoming	-1.16	-0.94
		Yellowstone NP	-0.78	-0.72
	Reference slope, pack size	Idaho	0.38	0.36
	Area-specific slope adjustment	NW Montana	0.10	0.10
		SW Montana-CIEPA	0.84	0.51
		SW Montana-GYEPA	0.12	0.06
		Wyoming	0.09	0.05
		Yellowstone NP	-0.01	-0.01
Alpha level for 95% confidence interval endpoints	Lower 1-sided interval	Idaho	0.05	0.11
		NW Montana	0.05	0.11
		SW Montana-CIEPA	0.05	0.11
		SW Montana-GYEPA	0.05	0.11
		Wyoming	0.05	0.12
		Yellowstone NP	0.05	0.11
	Lower 2-sided interval	Idaho	0.05	0.20
		NW Montana	0.05	0.16
		SW Montana-CIEPA	0.05	0.24
		SW Montana-GYEPA	0.05	0.16
		Wyoming	0.05	0.19
		Yellowstone NP	0.05	0.16
	Upper 2-sided interval	Idaho	0.05	0.10
		NW Montana	0.05	0.14
		SW Montana-CIEPA	0.05	0.10
		SW Montana-GYEPA	0.05	0.16
		Wyoming	0.05	0.10
		Yellowstone NP	0.05	0.10

level, the area-specific predictive model was changed substantially from the model that was estimated by maximum likelihood procedures.

Despite the superiority of the area-specific model, there were still some predictions that were not well calibrated to the observed probabilities in some areas. Model predictions were also based on a small amount of data (< 50 observations) for some areas. The greatest of these problems, and therefore the largest amount of shrinkage in the penalized maximum likelihood model coefficients, were in the areas with the least data in SW Montana. Overfitting and regression to the mean with maximum likelihood estimation are most likely in small data sets (Efron & Morris 1977, Steyerberg et al. 2000). While some predictions for the Wyoming data were also corrected in the penalized maximum likelihood shrinkage procedure, these data still showed substantial deviations from the ideal 45° line of observed vs predicted probabilities. This likely occurred because the penalized maximum likelihood model had similar assumptions as the maximum likelihood model, in that a logistic equation was estimated to represent the process generating the observations. Mitchell et al. (2008) assumed that the process generating the observed BP data was a smooth logistic function of pack size. Small packs in Wyoming, however, contained more BPs than were predicted by the logistic models and large packs contained fewer BPs than were predicted. This suggests that the relationship in Wyoming between pack size and the probability that a pack contains a BP is not well represented by

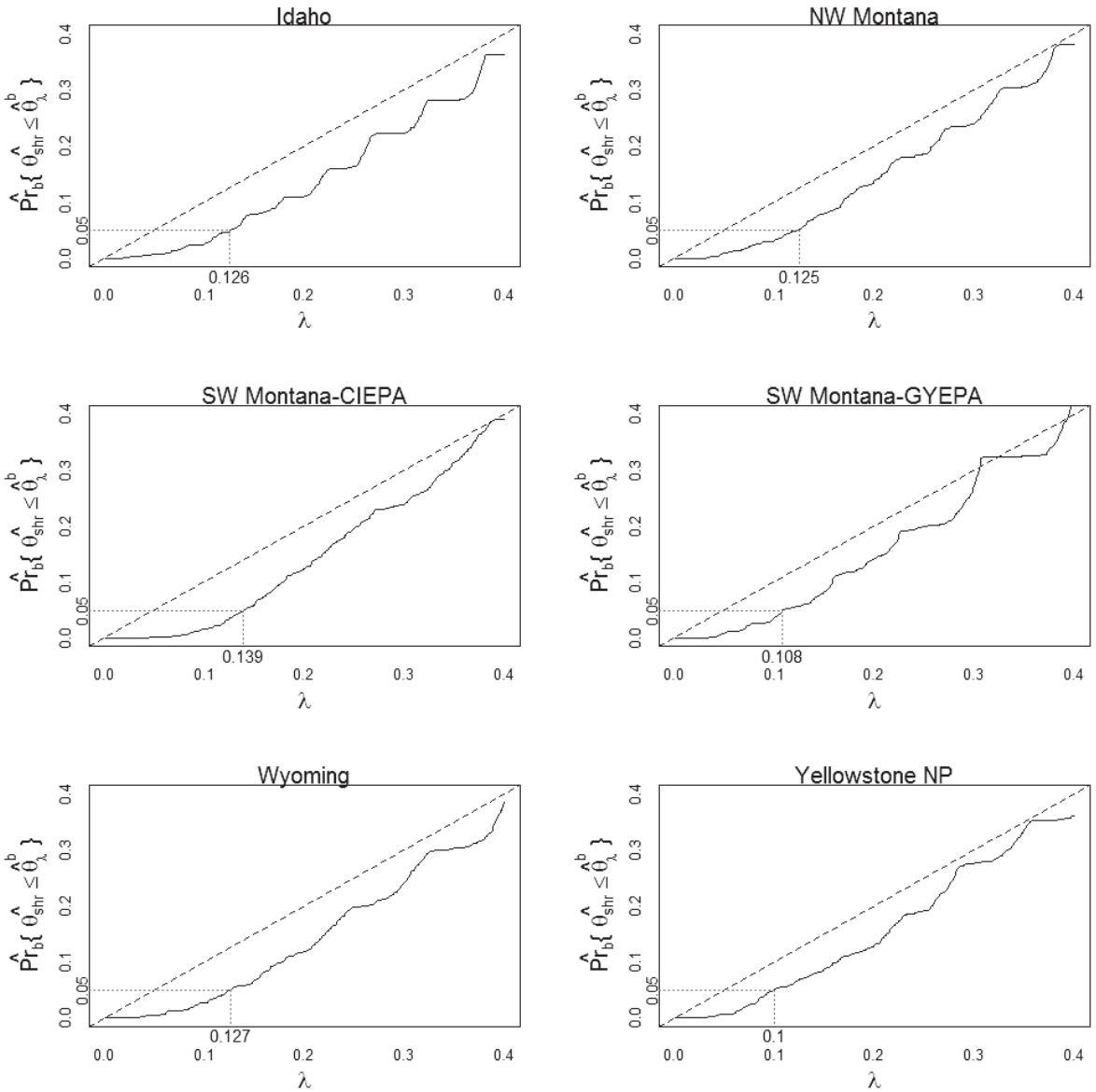

Figure 8. Example of a CI endpoint calibration for the lower 1-sided 95% endpoint. The solid line represents the actual coverage rates for the range of λ values depicted on the x-axis, as determined by the bootstrap. The dotted line displays the selection of a λ value to achieve a coverage error rate of $\alpha = 0.05$. The dashed line represents the ideal line of perfect calibration of the CI calculation procedure.

a smooth logistic function, and perhaps a function more flexible than a logistic function might result in better predictions for this area. Whether or not a more flexible function would result in better predictions for these data depends on how unique these observations are to this data set, or conversely, how common such observations are in the underlying population.

In our example, we employed shrinkage to improve the predictive ability of the area-specific predictive model. Shrinkage methods are not panaceas that produce perfect predictions for inadequate models, but shrinkage estimation will produce more accurate predictions under the assumed model for future data within the same pack size range dealt with here (Copas 1983, Firth 1993). A feature of the shrinkage predictions in our example is that larger predicted probabilities were shrunken closer to the global mean predictions as compared to maximum likelihood model predictions, particularly in the

areas with the fewest data. This occurred because of the generally higher predicted BP probabilities for packs in the size range of 8-13 animals in these areas. These predictions were shrunken to be closer to the mean predicted probabilities for these pack sizes, as represented by the global model. This illustrates how shrinkage predictions deal with the ubiquitous regression to the mean problem, in which observed probabilities in the future will be closer to the mean than what was observed in the data set at hand (Efron & Morris 1977, Harrell et al. 1996, Harrell 2001: 62). Efron & Morris (1977), White et al. (2002), and others have identified that this is more common in small data sets due to higher sampling variances.

Actual coverage rates for CIs presented in Mitchell et al. (2008) were higher than the nominal rates that we were trying to achieve. This result is not surprising given that coverage rates for the CI procedure described in Mitchell et al. (2008) is known to have poor properties for predicted probabilities close to 0 and 1 (Jennings 1987). Predicted probabilities close to 1 were present in every analysis area under the area-specific model.

The outcome of the internal validation which we conducted on the logistic models presented by Mitchell et al. (2008) was to maximize the utility of the area-specific model for making predictions with future data. We used shrinkage to help ensure that model predictions will be well calibrated to the observed proportions of packs that contain BPs, across the full range of pack sizes considered here. We have also made sure that the model can reliably predict packs that have high and low probabilities of containing BPs, and that estimates of uncertainty accurately reflect the true amount of uncertainty. This validation is no substitute, however, for evaluating and improving the models using data collected in the future to ensure the models are performing adequately. This would also serve to identify unanticipated situations in which the models might not work well. An assumption in our analysis is that historical observations used to generate the models and predictions from the models continue to apply to the same underlying populations. This assumption would be violated if the processes determining the relationship between pack size and the probability of a pack containing a BP changed in the future, resulting in differences between the wolf populations used to generate the models and those for which models are being used to make predictions. Disease outbreaks, major changes in prey density, and extensive hunting and trapping harvest are examples of processes that might influence the validity of this assumption. Evaluating model predictions using future data from a continued intensive monitoring program will help identify such problems. Model predictions can be re-evaluated and calibrated to match the new observations more closely using temporal validation procedures (Justice et al. 1999, Altman & Royston 2000). Further, the accuracy of the CI calibration we performed might diminish if the distribution of pack sizes changes. The normal approximation to the binomial CI has better coverage properties for predicted probabilities closer to 0.5 (Jennings 1987). Therefore, if increased human-caused mortality following delisting results in smaller pack sizes, using α might result in better coverage rates than what we observed.

Management implications

Both the analyses of the example presented in our paper and the analysis methods themselves have implications for wildlife management. Our analysis of the models for estimating BPs presented by Mitchell et al. (2008) suggests that an area-specific model will provide robust and reliable estimates for the NRM wolf population into the future. We recommend that managers use the shrunken, area-specific model to estimate the number of BP, and rank the relative probabilities that several packs are BPs, provided that the underlying wolf population can be considered the same population that was used to generate the model. To temporally validate this assumption, we recommend that the predictive accuracy of the model be evaluated as the intensive field-monitoring program continues in the near future. Further, we recommend that this assumption be evaluated as major changes in the population dynamics of wolves in these areas occur in the future, possibly resulting from disease outbreaks, major changes in prey density, extensive hunting and trapping harvest, or other factors.

More generally, the issues of prediction accuracy that arose in our example are the same issues that are faced by any prediction method that depends on statistical models. Some of the methods we have presented can be directly extended to other parametric as well as non-parametric predictive models (Hastie et al. 2002, Hochachka et al. 2007). For example, the measures of discrimination we used can be applied to any prediction method for continuous, dichotomous, ordinal and time-to-event outcomes (Harrell et al. 1996). The calibration concepts that we described were specific to ordinary

logistic regression, but they can be applied to time-to-event predictions (Harrell 2001:493-494) and Resource Selection Functions (Johnson et al. 2006) with some modification. Other measures of model calibration also exist for linear regression (Harrell 2001: 91) and case-control sampling designs in logistic regression (Hosmer & Lemeshow 2000:248-250). The bootstrap validation methods we used can be used for any prediction method that is based on data (Efron & Tibshirani 1993:237-255, Harrell 2001:95). However, when it is known that data will follow a particular parametric structure, estimation of prediction error and accuracy can be more efficiently achieved using covariance-based penalties rather than nonparametric, resampling-based optimism estimates (Efron 2004). Such penalties include Akaike's Information Criterion (AIC), which is a tool for selecting the best predictive model from a set of alternative models (Akaike 1981). These methods offer more accurate estimates of future predictive ability at the cost of more assumptions, which may be appropriate for some situations. Further, the accuracy of the covariance-based penalties is asymptotic, and these methods are subject to the many biases that can arise in small data sets collected under realistic field conditions. Conversely, the methods we have presented focus on a practical application of predictive models, and provide a solid basis to examine if a predictive model fit to real data will suffice for the purpose for which it was developed. If a model is to be used for predictive purposes related to wildlife management, we propose that a predictive logistic model selected using covariance penalties be evaluated with the methods we present here. Such efforts will enhance the credibility of management decisions, and more fundamentally will maximize the odds that decisions based on model predictions will achieve management goals.

Acknowledgements - we are very grateful for all the contributions made to wolf monitoring efforts by hundreds of people since 1979. Employees and volunteers affiliated with the US Fish and Wildlife Service, the National Park Service, US Forest Service, the US Bureau of Land Management, USDA Wildlife Services, academic institutions, Nez Perce Tribe, Blackfeet Nation, Confederated Salish and Kootenai Tribe, Turner Endangered Species Fund, Montana Fish, Wildlife & Parks, Idaho Department of Fish and Game, and Wyoming Game and Fish contributed greatly to the data set we used for our analyses. Vanna Boccadori, Steve Cherry, Matthew Gray, Nigel Yoccoz and three anonymous reviewers provided useful comments that improved this paper.

References

Akaike, H. 1981: Likelihood of a model and information criteria. - Journal of Econometrics 16: 3-14.

Altman, D.G. & Royston, P. 2000: What do we mean by validating a prognostic model? - Statistics in Medicine 19: 453-473.

Boyce, M.S., Vernier, P.R., Neilson, S.E. & Schmiegelow, F.K.A. 2002: Evaluating resource selection functions. - Ecological Modelling 157: 281-300.

Canty, A. & Ripley, B.D. 2006: boot: Bootstrap R (S-Plus) Functions. R package version 1.2-27. - Available at: http://cran.r-project.org/web/packages/boot/index.html (Last accessed on 13 October 2009).

Cleveland, W.S. 1979: Robust locally weighted regression and smoothing scatterplots. - Journal of the American Statistical Association 74: 829-836.

Cleveland, W.S. 1981: LOWESS: A program for smoothing scatterplots by robust locally weighted regression. - The American Statistician 35: 54.

Copas, J.B. 1983: Regression, prediction, and shrinkage. - Journal of the Royal Statistical Society, Series B, Methodological 45: 311-354.

Cox, D.R. 1958: Two further applications of a model for binary regression. - Biometrika 45: 562-565.

Efron, B. 2004: The estimation of prediction error: covariance penalties and cross-validation. - Journal of the American Statistical Association 99: 619-622.

Efron, B. & Morris, C. 1977: Stein's paradox in statistics. - Scientific American 236: 119-127.

Efron, B. & Tibshirani, R.J. 1993: An introduction to the bootstrap. - Chapman and Hall, Boca Raton, Florida, USA, 436 pp.

Fielding, A.H. & Bell, J.F. 1997: A review of methods for the assessment of prediction errors in conservation presence/absence models. - Environmental Conservation 24: 38-49.

Firth, D. 1993: Bias reduction of maximum likelihood estimates. - Biometrika 80: 27-38.

Firth, D. 2006: brlr: Bias-reduced logistic regression. R package version 0.8-8. - Available at: http://www.warwick.ac.uk/go/dfirth (Last accessed on 6 July 2009).

Guthery, F.S. & Bingham, R.L. 2007: A primer on interpreting regression models. - Journal of Wildlife Management 71: 684-693.

Harrell, F.E., Jr. 2001: Regression modeling strategies, with application to linear models, logistic regression, and survival analysis. - Springer-Verlag, New York, New York, USA, 568 pp.

Harrell, F.E., Jr. 2005: Design: Design Package. R package version 2.0-12. - Available at: http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/Design (Last accessed on 13 October 2009).

Harrell, F.E. Jr., Lee, K.L., Califf, R.M., Pryor, D.B. & Rosati, R.A. 1984: Regression modeling strategies for

improved prognostic prediction. - Statistics in Medicine 3: 143-152.

Harrell, F.E., Jr., Lee, K.L. & Mark, D.B. 1996: Tutorial in biostatistics; multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. - Statistics in Medicine 15: 361-387.

Hastie, T., Tibshirani, R.J. & Friedman, J. 2003: The elements of statistical learning; data mining, statistical inference, and prediction. - Springer-Verlag, New York, New York, USA, 533 pp.

Hochachka, W.M., Caruana, R., Fink, D., Munson, A., Riedewald, A., Sorokina, D. & Kelling, S. 2007: Data-mining discovery of pattern and process in ecological systems. - Journal of Wildlife Management 71: 2427-2437.

Hosmer, D.W. & Lemeshow, S. 2000: Applied logistic regression. - John Wiley and Sons, New York, New York, USA, 375 pp.

Jeffreys, H. 1946: An invariant form for the prior probability in estimation problems. - Proceedings of the Royal Society, Series A: 186: 453-461.

Jennings, D.E. 1987: How do we judge confidence-interval adequacy?. - The American Statistician 41: 335-337.

Johnson, C.J., Nielsen, S.E., Merrill, E.H., Mcdonald, T.L. & Boyce, M.S. 2006: Resource selection functions based on use-availability data: theoretical motivation and evaluation methods. - Journal of Wildlife Management 70: 347-357.

Justice, A.C., Covinsky, K.E. & Berlin, J.A. 1999: Assessing the generalizability of prognostic information. - Annals of Internal Medicine 130: 515-524.

Keating, K.A. & Cherry, S. 2004: Use and interpretation of logistic regression in habitat-selection studies. - Journal of Wildlife Management 68: 774-789.

Lobo, J.M., Jiménez-Valverde, A. & Real, R. 2008: AUC: a misleading measure of the performance of predictive distribution models. - Global Ecology and Biogeography 17: 145-151.

Miller, M.E., Hui, S.L. & Tierney, W.M. 1991: Validation techniques for logistic regression models. - Statistics in Medicine 10: 1213-1226.

Mitchell, M.S., Ausband, D.E., Sime, C.A., Bangs, E.E., Gude, J.A., Jiminez, M.D., Mack, C.M., Meier, T.J., Nadeau, M.S. & Smith, D.W. 2008: Estimation of successful breeding pairs for wolves in the U.S. northern Rocky Mountains. - Journal of Wildlife Management 72: 881-891.

Mladenoff, D.J., Sickley, T.A., Haight, R.G. & Wydeven, A.P. 1995: A regional landscape analysis and prediction of favorable wolf habitat in the Northern Great Lakes region. - Conservation Biology 9: 279-294.

O'Brien, C.S., Rosenstock, S.S., Hervert, J.J., Bright, J.L. & Boe, S.R. 2005: Landscape-level models of potential habitat for Sonoran pronghorn. - Wildlife Society Bulletin 33: 24-34.

Pearce, J. & Ferrier, S. 2000: Evaluating the predictive performance of habitat models developed using logistic regression. - Ecological Modelling 133: 225-245.

Randin, C.F., Dirnböck, T., Dullinger, S., Zimmerman, N.E., Zappa, M. & Guisan, A. 2006: Are niche-based species distribution models transferable in space? - Journal of Biogeography 33: 1689-1703.

R Development Core Team 2007: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. - Available at: http://www.R-project.org. (Last accessed on 13 October 2009).

Scott, J.M., Heglund, P.J., Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.A. & Samson, F.B. (Eds.) 2002: Predicting species occurrence: issues of accuracy and scale. - Island Press, Washington, D.C., USA, 868 pp.

Steyerberg, E.W., Eijkemans, M.J.C., Harrell, F.E. Jr. & Habbema, J.D.F. 2000: Prognostic modeling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. - Statistics in Medicine 19: 1059-1079.

Steyerberg, E.W., Harrell, F.E. Jr., Borsboom, G.J.J.M., Eijkemans, M.J.C., Vergouwe, Y. & Habbema, J.D.F. 2001: Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. - Journal of Clinical Epidemiology 54: 774-781.

Thompson, S.K. 2002: Sampling. 2nd edition. - John Wiley and Sons, New York, New York, USA, 367 pp.

Turner, J.C., Douglas, C.L., Hallum, C.R., Krausman, P.R. & Ramey, R.R. 2004: Determination of critical habitat for the endangered Nelson's bighorn sheep in southern California. - Wildlife Society Bulletin 32: 427-448.

US Fish and Wildlife Service (USFWS) 1994: The reintroduction of gray wolves to Yellowstone National Park and central Idaho. - Final Environmental Impact Statement. Denver, Colorado, Appendix 9, 414 pp. Available at: http://westerngraywolf.fws.gov/EIS_1994.pdf (Last accessed on 13 October 2009).

van Houwelingen, J.C. & Le Cressie, S. 1990: Predictive value of statistical models. - Statistics in Medicine 9: 1303-1325.

Venables, W.N. & Ripley, B.D. 2002: Modern Applied Statistics with S. 4th edition. - Springer, New York, New York, USA, 495 pp.

Verbyla, D.L. & Litvaitis, J.A. 1989: Resampling methods for evaluating classification accuracy of wildlife habitat models. - Environmental Management 13: 783-787.

White, G.C. 2000: Population viability analyses: data requirements and essential analyses. - In: Boitani, L. & Fuller, T.K. (Eds); Research techniques in animal ecology: controversies and consequencies. - Columbia

University Press, New York, New York, USA, pp. 288-331.
White, G.C., Franklin, A.B. & Shenk, T.M. 2002: Estimating parameters of PVA models from data on marked animals. - In: Beissinger, S.R. & McCullough, D.R. (Eds); Population viability analysis. University of Chicago Press, Chicago, Illinois, USA, pp. 169-190.
Williams, B.K., Nichols, J.D. & Conroy, M.J. 2002: Analysis and management of animal populations. - Academic Press, San Diego, California, USA, 817 pp.