The Journal of Wildlife Management 9999:1–13; 2011; DOI: 10.1002/jwmg.213

Research Article

Spending Degrees of Freedom in a Poor Economy: A Case Study of Building a Sightability Model for Moose in Northeastern Minnesota

JOHN H. GIUDICE,[1] *Biometrics Unit, Minnesota Department of Natural Resources, 5463C West Broadway Avenue, Forest Lake, MN 55025, USA*
JOHN R. FIEBERG, *Biometrics Unit, Minnesota Department of Natural Resources, 5463C West Broadway Avenue, Forest Lake, MN 55025, USA*
MARK S. LENARZ, *Minnesota Department of Natural Resources, 1201 East Highway 2, Grand Rapids, MN 55744, USA*

ABSTRACT Sightability models are binary logistic-regression models used to estimate and adjust for visibility bias in wildlife-population surveys. Like many models in wildlife and ecology, sightability models are typically developed from small observational datasets with many candidate predictors. Aggressive model-selection methods are often employed to choose a best model for prediction and effect estimation, despite evidence that such methods can lead to overfitting (i.e., selected models may describe random error or noise rather than true predictor–response curves) and poor predictive ability. We used moose (*Alces alces*) sightability data from northeastern Minnesota (2005–2007) as a case study to illustrate an alternative approach, which we refer to as degrees-of-freedom (df) spending: sample-size guidelines are used to determine an acceptable level of model complexity and then a pre-specified model is fit to the data and used for inference. For comparison, we also constructed sightability models using Akaike's Information Criterion (AIC) step-down procedures and model averaging (based on a small set of models developed using df-spending guidelines). We used bootstrap procedures to mimic the process of model fitting and prediction, and to compute an index of overfitting, expected predictive accuracy, and model-selection uncertainty. The index of overfitting increased 13% when the number of candidate predictors was increased from three to eight and a best model was selected using step-down procedures. Likewise, model-selection uncertainty increased when the number of candidate predictors increased. Model averaging (based on $R = 30$ models with 1–3 predictors) effectively shrunk regression coefficients toward zero and produced similar estimates of precision to our 3-df pre-specified model. As such, model averaging may help to guard against overfitting when too many predictors are considered (relative to available sample size). The set of candidate models will influence the extent to which coefficients are shrunk toward zero, which has implications for how one might apply model averaging to problems traditionally approached using variable-selection methods. We often recommend the df-spending approach in our consulting work because it is easy to implement and it naturally forces investigators to think carefully about their models and predictors. Nonetheless, similar concepts should apply whether one is fitting 1 model or using multi-model inference. For example, model-building decisions should consider the effective sample size, and potential predictors should be screened (without looking at their relationship to the response) for missing data, narrow distributions, collinearity, potentially overly influential observations, and measurement errors (e.g., via logical error checks).

KEY WORDS aerial survey, *Alces alces*, degrees-of-freedom spending, logistic regression, Minnesota, model averaging, model selection, moose, shrinkage, sightability, visibility bias.

It is widely recognized that even well-designed aerial surveys can result in population estimates that are biased low because <100% of animals present are counted (Caughley 1974, Pollock and Kendall 1987, Samuel et al. 1987, Steinhorst and Samuel 1989). Much of the work on aerial surveys over the last 30 years has concentrated on estimating visibility bias (Pollock and Kendall 1987), a measure of how many animals are missed. Approaches for estimating and adjusting for visibility bias in aerial surveys include distance sampling (Buckland et al. 2001), mark-resight and multiple-observer methods (Rice and Harder 1977, Bartmann et al. 1987, Potvin et al. 1992), double sampling (Gasaway et al. 1986), and sightability models (Steinhorst and Samuel 1989, Walsh et al. 2009). Sightability models are binary-logistic models applied to sighting data collected from

Received: 29 September 2010; Accepted: 21 April 2011

[1]*E-mail: john.giudice@state.mn.us*

marked individuals, and have been developed for a variety of large ungulates including moose (*Alces alces*) populations in Wyoming, USA (Anderson and Lindzey 1996), Michigan, USA (Drummer and Aho 1998), and British Columbia, Canada (Quayle et al. 2001).

Sightability models are typically developed by conducting repeated surveys of radio-marked animals (over 1 or a few years), taking measurements on many candidate predictor variables, and then using variable-selection methods to identify important sightability covariates and determine model complexity (for a nice exception, see Rice et al. 2009). Common variable-selection methods include stepwise procedures, selection based on information criterion (e.g., Akaike's Information Criterion [AIC]), and univariate screening (e.g., using plots or models of the response versus single predictors). Stepwise and other aggressive variable-selection procedures have been criticized in the statistical (Derksen and Keselman 1992, Steyerberg et al. 2000, Harrell 2001) and ecological literature (Whittingham et al. 2006, Mundry and Nunn 2009, Dahlgren 2010). In particular, Harrell (2001:56:57) noted the following problems:

1. Regression coefficient estimates will be biased high in absolute value. "The choice of the variables to be included depends on estimated regression coefficients rather than their true values, and so x_j is more likely to be included if its regression coefficient is overestimated than if its regression coefficient is underestimated" (Copas and Long 1991).
2. Stepwise selection yields R^2 values that are optimistic (i.e., as an estimate of response variation explained by the model).
3. The distribution of test statistics are difficult to determine and do not follow those typically assumed for pre-specified hypotheses (e.g., F and χ^2 distributions).
4. Standard errors of regression coefficient estimates are biased low; thus, confidence intervals are too narrow (i.e., the true coverage probability is less than the nominal coverage probability, which often is set at 0.95).
5. *P*-values are optimistic (i.e., biased toward rejecting the null hypothesis) and no longer have the standard interpretation, and corrections for multiple comparisons are not easy to apply.
6. Choice of variables is often arbitrary when predictors are linearly related (or collinear).
7. It avoids the need to think critically about the problem.

These results have been shown in many simulation studies, several of which are summarized in Harrell (2001:8, 28, 57, 59, 61) and Babyak (2004). In addition to problems associated with obtaining unbiased estimates of regression parameters and their uncertainty following model selection, models chosen using stepwise procedures tend to predict future observations poorly when the number of predictor variables considered during model fitting is large relative to information in the data (Harrell 2001). Allowing too much flexibility in the model-building process increases the chance that the chosen model(s) will fit noise in the data rather than the underlying relationship between response and predictors, which is termed overfitting (Harrell 2001, Babyak 2004, Mundry and Nunn 2009). At the extreme, a model with n free parameters can fit n unique data points exactly, but such a model will almost surely perform poorly when used to predict new observations.

Similar concerns apply to other aggressive model-building strategies including univariate screening, consideration of multiple transformations or distributional assumptions, and sequential model-building processes where model sets are constructed for groups of covariates or model components and then model selection is performed separately for each group or component with the best model(s) carried forward from one selection process to the next (Harrell 2001). A further disadvantage of univariate screening and sequential approaches is that they may fail to discover important predictors that can only be discovered after adjusting for other confounding variables (Sun et al. 1996). Regardless of the approach, one can measure the degree of overfitting by plotting new response data against predicted values, or if new or hold-out data are not available, one can use bootstrapping to mimic the process of model fitting and prediction of new responses (Harrell 2001:94). The typical pattern is that low predictions are too low and high predictions are too high. This statistical phenomenon is referred to as "regression to the mean" (Copas 1997, Harrell 2001). The intercept (γ_0) and slope (γ_1) of the regression line (y = new responses, x = predicted values) are referred to as calibration factors (since they could potentially be used to correct for overfitting); γ_1 will be <1 if there is overfitting and γ_0 will be different from zero to compensate. Thus, a bootstrap estimate of γ_1 not only quantifies overfitting but also provides an estimate of shrinkage (i.e., $1 - \gamma_1$ = the amount that regression coefficients need to be shrunk toward zero to make predictions more calibrated; Harrell 2001:95).

Harrell (2001) and Babyak (2004) discussed two alternative modeling strategies that attempt to overcome these pitfalls. The first strategy, degrees-of-freedom (df) spending, uses sample-size guidelines to roughly determine the amount of information in the data and, thus, the df one can afford to spend on the model (Appendix A). Degrees of freedom are then allocated to specific predictor variables and their interactions based on a priori expectations (multiple df may be allocated for nonlinear effects). Data-reduction techniques can be used to help decrease the number of candidate predictors before allocating the df (e.g., dropping variables that have many missing values or that vary little among sample units, or combining multiple, often correlated variables into a single index using techniques such as principal components analysis). Once df have been spent (i.e., a full model is fit), the statistical and biological significance of each predictor can be judged by examining the estimated regression parameters (or functions of these parameters) along with associated confidence intervals to determine patterns that can and cannot be ruled out by the data. In other words, a full, pre-specified model is used for inference; thus, confidence intervals, *P*-values, and R^2 measures all have their intended interpretation. No model selection is involved.

The second strategy involves shrinking parameter estimates of less important variables toward zero to help produce more reliable predictions, especially in cases where there are few observations per predictor or many variables with small effects (e.g., Harrell 2001:207–210; van Houwelingen 2001, Dahlgren 2010). Several statistical methods have been developed to accomplish this goal (e.g., penalized maximum likelihood estimation, ridge regression, and lasso regression); we refer the reader to van Houwelingen (2001) for an introduction to these methods. Similarly, Harrell (2001) provided examples of how one can use bootstrap-validation procedures to shrink regression coefficient estimates and predictions towards zero, after the fact, in cases where one fails to reduce the number of candidate predictors sufficiently beforehand. Although these techniques are not well-known among ecologists, they share similarities with model averaging and multi-model inference. In particular, Burnham and Anderson (2002:158–164) suggested adjusting standard errors to account for exploring multiple models, and they proposed model averaging for obtaining predictions; the latter can be shown to be a type of shrinkage estimator (Hoeting et al. 1999, Burnham and Anderson 2002:253–255, Lukacs et al. 2010).

Shrinkage methods (including model averaging) and fully pre-specified models with limited df both provide advantages relative to performing model selection and then proceeding as if the selected model had been pre-specified. In our consulting work, we tend to emphasize a df-spending approach because: 1) we often find it simpler to apply; 2) limiting the number of candidate variables to be in line with sample-size guidelines forces one to think hard about the problem and helps investigators focus on estimates of effect sizes and their uncertainty; 3) it explicitly recognizes the limitations of small datasets; and 4) it emphasizes the need to collect sufficient data with sample-size requirements dependent on the complexity of the problem. However, we acknowledge there is no universal best approach to model-based inference. In fact, we recommend a pragmatic and flexible approach to statistical inference, recognizing that different methods suit different problems (Chatfield 2002). Our goal is simply to illustrate an alternative approach to aggressive model-building strategies (based on df spending), which seems to be unfamiliar to many wildlife biologists. The df-spending approach we advocate is largely adopted from Harrell (2001). We list the main steps of the df-spending approach in Appendix A, but readers are encouraged to consult Harrell (2001) for more details.

As a case study, we illustrate the df-spending approach by building a sightability model for moose in northeastern Minnesota during 2005–2007. In addition to reviewing the implications of stepwise-selection strategies, our objectives were to: 1) illustrate an alternative approach to model development based on df spending and exploratory analyses that do not use the response (Y) to specify a list of candidate predictors, and 2) compare predictive ability and precision of population estimates from our "df model" (single-model inference) to estimates from models developed using more traditional model-selection methods.

STUDY AREA

Our study area included 15,500 km^2 of boreal forest in northeastern Minnesota (Lenarz 1998, 2007). The area was a low plateau of modest relief that rose abruptly from Lake Superior to a crest approximately 700 m above sea level (Heinselman 1996). The area was sparsely inhabited by people and most communities occurred along Lake Superior on the southeastern boundary of the study area. There were few paved roads and much of the area was accessible only from logging roads or snowmobile and all-terrain-vehicle trails.

The study area was a mosaic of conifer communities classified as the Northern Superior Upland section (Minnesota Department of Natural Resources [MNDNR] 2007). The landscape was characterized by northern white cedar (*Thuja occidentalis*), black spruce (*Picea mariana*), and tamarack (*Larix laricina*) on the lowlands and balsam fir (*Abies balsamea*), and jack, white, and red pines (*Pinus banksiana*, *P. strobus*, and *P. resinosa*, respectively) on uplands. Deciduous forests contained mostly quaking aspen (*Populus tremuloides*) and white birch (*Betula papyrifera*). Non-forested areas were characterized as upland and lowland deciduous shrub and sedge meadows. Wetlands, lakes, bogs, swamps, and small streams were interspersed among the rolling uplands.

METHODS

Sightability Trials

Fifty-five adult ($\geq$1.7-yr-old) male and 61 adult female moose were radio-marked during 2002–2005 as part of a survival study (Lenarz et al. 2009). However, annual mortality events and availability issues (e.g., animals moved outside the boundaries of survey plots or were located in plots that were not safe to survey) reduced our sightability sample for 2005–2007 to 55 radio-marked moose (19 males, 36 females). Sightability trials were conducted concurrently with operational surveys of moose in northeastern Minnesota (Lenarz 2007). Data from the sightability trials (surveys) were used to build predictive sightability models, whereas operational surveys were standard aerial counts of moose in randomly selected plots (with no attempt to differentiate marked and unmarked moose).

Surveys were initiated as soon as mean snow depth was $\geq$20 cm, which was early January in most years. Both operational and sightability surveys were conducted using a Bell OH-58A helicopter (Bell Helicopter Textron, Fort Worth, TX) flying 90–120 m above ground level at 75–110 km/hr on east-west transects spaced at 0.5 km within rectangular survey plots. We used two survey teams, each consisting of a pilot and two experienced observers (one seated behind the pilot). We used fixed-wing aircraft to locate all radio-collared moose at least weekly, and based on that information we constructed test plots (4.0 km $\times$ 4.3 km) for each sightability survey. The test plots were delineated solely to

provide areas likely to contain radiocollared moose. Individuals responsible for identifying test plots were not involved in sightability surveys. We surveyed test plots within 1–5 days of fixed-wing flights using the same procedures as in operational surveys (except that plots were 8.0 km × 4.3 km in operational surveys). When a moose (marked or unmarked) was sighted, the helicopter left the transect and circled the moose to determine group size, classify individuals according to sex and age (calf or adult), and identify any marked animals. We also recorded a suite of potential sightability covariates for each moose-group observation, including an ocular estimate of visual obstruction (VO) (Table 1). We defined VO as the amount of screening cover within four animal lengths (approx. 10-m radius circle) of the first animal seen, and it was measured from the location and angle of initial sighting. We used photographs of moose taken during previous surveys to train observers and standardize ocular estimates of screening cover. If we failed to detect a radiocollared moose during a sightability survey, we attempted to locate it using telemetry immediately after the test plot was surveyed. If the radiocollared moose was still within the boundaries of the test plot, we collected the same suite of covariate information. VO in this case was measured from an oblique angle while circling the missed animal or group. Collared moose that had moved outside the test plot were not included in sightability trials.

Df Spending Approach: Determining Model Complexity

We collected data on 15 candidate predictor variables, which would require at least 24 regression parameters to model each predictor as a linear effect (with no interactions; Table 1). Given the small number of sightability trials ($n = 124$), we questioned whether we had sufficient information in our dataset to avoid model overfitting. In the case of a binary response variable, general guidelines for avoiding model overfitting is to limit the df associated with predictors (including complex terms such as interactions and nonlinear terms) to $m/10$ or $m/20$, where $m = \min[n_0 =$ number of observed zeros in the response, $n_1 =$ number of observed ones in the response] (Harrell 2001; also see Appendix A). We had $n_0 = 65$ missed and $n_1 = 59$ observed moose groups in 124 sightability trials; thus, $m = 59$, suggesting a maximum df = 3–6. Similarly, likelihood-based formulas for estimating the target number of total regression df (Harrell 2001:73; Appendix A) suggested a maximum df = 3–4. Some of our predictors had narrow distributions (e.g., group size) and some radio-marked moose were surveyed more than once. Therefore, we elected to take a conservative approach and limited total regression df to ≤ 3.

Df-Spending Approach: Choosing How to Spend Df

Intuitively, VO should have the strongest association with probability of detection (π), and our sample contained a wide distribution of VO values (median = 60%, range: 0–95%, interquartile range [IQR] = 30–80%). Thus, we felt VO should be included in any candidate model. We also considered spending another 1–2 df to model the effect of VO as nonlinear (e.g., via higher-order terms), but initially we reserved the remaining 2 df for other predictors. We graphically evaluated other predictors (e.g., via histograms, box plots, scatterplots) without looking at their relationships to Y, and considered eliminating predictors that had a narrow range of observed values, a large number of missing values (in either the sightability or operational surveys), potential measurement-error issues (e.g., imprecise measurements, subjective criteria), or were strongly correlated with VO. For example, cover type has been used or considered in several sightability-model studies (Anderson and Lindzey 1996, Anderson et al. 1998, Quayle et al. 2001). However, we suspect that cover type and VO often explain the same source of variation in Y (i.e., cover type would have little explanatory value once VO was in the model). In this study, high VO values were associated mostly with animals in conifer or mixed conifer-hardwood cover. Further, sample sizes were small ($n \leq 12$) in two cover classes (open and hardwoods). Finally, using cover type as a nominal predictor would require 3 df by itself, which would exceed our df guidelines when combined with VO. Given these limitations, one option would be to pool "open" and "hardwood" cover types, and reclassify cover as an ordinal factor (requiring 1 df) that

Table 1. Sightability predictors measured in aerial surveys of radio-collared moose in Minnesota, 2005–2007.

Predictor	Description	Data scale	Regression degrees of freedom[a]
Survey crew	Pilot, primary observer, secondary observer	Nominal (≥7 categories)	6
Snow depth	<20 cm, 20–41 cm, >41 cm	Ordinal (3 classes)	1–2
Wind speed	km/hr	Ratio	1
Wind direction	0 = calm, 1–360°	Interval	1
Temperature	°C	Interval	1
Barometric pressure	Millibars	Ratio	1
Barometric trend	Rising, falling, stable	Nominal (3 categories)	2
Cloud cover	0%, 1–25%, . . ., 76–100%	Ordinal (5 classes)	1
Survey conditions[b]	Good, marginal, poor	Ordinal (3 classes)	1–2
Sex-Age	Bull, cow, cow-calf	Nominal (3 categories)	2
Visual obstruction	Ocular estimate to nearest 5%	Ratio	1
Light intensity	Flat (no shadows), Bright (distinct shadows)	Nominal (2 categories)	1
Group size	Number of individuals	Ratio	1
Distance	Perpendicular distance (m) from transect line	Ratio	1
Cover type	Conifer, hardwood, mixed, open	Nominal (4 categories)	3

[a] Number of regression parameters (coefficients, excluding intercept) needed to model each predictor as a linear effect, without interactions.
[b] Subject assessment based on snow cover, light conditions, turbulence, wind speed, etc.

described the relative contribution of conifers (CONIF) to VO: 1 = little or no conifer cover; 2 = mixed conifers and hardwoods; 3 = mostly conifers. However, correlation with VO would remain a concern. Finally, moose often segregate by sex and age classes (Peek et al. 1974) with differential habitat use and group sizes, which can influence detectability (Gasaway et al. 1986, Samuel et al. 1992; but see McCorquodale 2001). Nonetheless, the effects of differential habitat use (on detectability) should already be captured by VO or, possibly, VO + CONIF.

Group size was another intuitive predictor (Samuel and Pollock 1981, Samuel et al. 1987, Cogan and Diefenbach 1998, McCorquodale 2001, Rice et al. 2009), and the effect of group size should interact with VO (i.e., group size should be more important at moderate to high VO values). However, in our study, group size had a narrow, positively skewed distribution (median = 2, range: 1–7, IQR: 1–2). Thus, our group-size predictor contained little information for modeling sightability, and even less information for evaluating an interaction with VO. Another candidate variable was survey conditions (subjective assessment of overall survey conditions based on snow cover, light conditions, turbulence, wind speed, etc.), which should supplement VO as a predictor. However, only 25% of our 124 sightability observations were collected under less than "good" conditions and all of those occurred in 2007. Furthermore, cross-tabulation summaries suggested that snow cover (depth classes) was the primary factor influencing the valuation of survey conditions. Hence, snow cover and survey conditions were positively correlated and both reflected mostly annual variation (i.e., within-year variation was absent, except in 2007). Thus, any potential variation explained by these variables could be due to an unmeasured confounder that also varied annually.

The remaining candidate predictors involved weather variables (e.g., temperature, wind speed, cloud cover, barometric pressure). Weather could theoretically influence moose behavior (e.g., habitat use, activity, group dispersion) that in turn could influence detectability. However, changes in habitat use should be reflected in average VO values, and linking activity and probability of detection should be done directly with an activity variable (e.g., activity = standing, bedded, walking, or running) rather than indirectly via a weather predictor. Unfortunately, describing the activity of moose during a sightability trial is not straightforward, especially for moose that are missed. Ideally, these types of data issues are identified during the design phase and all pertinent covariates are measured accurately, without missing data, and across the full range of potential values. Unfortunately, this is rarely the case in observational field studies because biological phenomena often involve complex and poorly understood relationships, important experimental-design components are lacking (e.g., the range and distribution of predictor values frequently is not controlled beforehand and final sample sizes often are much smaller than planned), and, possibly, not enough time was spent on a priori critical thinking (including the use of pilot studies). The bottom line is that despite collecting data on 15 potential predictors, we felt only VO had good distributional and theoretical properties. Consequently, our df-spending model included 1 predictor (VO) and we used the additional 2 df to model VO as a restricted cubic spline with four knots {VO(4k)}.

Alternatively, one could construct a model set containing all 3-df model choices that contain VO, and then use model averaging (Burnham and Anderson 2002, Lukacs et al. 2010) to compute estimates of regression parameters or detection probabilities that are not conditional on any given model. We constructed 28 candidate models that contained VO and 0–2 additional predictors (group size [GRP], conifer cover [CONIF], wind speed [WIND], temperature [TEMP], cloud cover [CLOUD], and either snow depth [SNOW] or survey conditions [SCOND]). We also included {VO(4k)} and {VO + GRP + VO × GRP} in the candidate set for a total of $R = 30$ models with predictor df = 1–3. We restricted our consideration of interactions to VO × GRP based on biological plausibility (discussed above; also see Appendix A—step 2), that is, we did not have a strong a priori basis for evaluating other two-way interactions involving VO.

Model Fit and Validation

As with other analytic approaches, model formulation is only one step in the overall strategy for developing reliable predictive models (Appendix A). Other steps are equally important for determining the usefulness of a particular model (e.g., checking model and data assumptions, interpreting the model, and validating the model for calibration and discrimination ability). Models that perform poorly may generate new questions that deserve further study (i.e., negative results can be informative). The model evaluation process may also lead one to fit alternative models with new or transformed variables. Since these latter models are influenced by the current data, they should be viewed more cautiously. Nonetheless, the impact of data-driven decisions can sometimes (and often should) be evaluated by replicating the entire model-fitting process using the bootstrap (Faraway 1992, Harrell 2001). We will give an example of how to do this in the next section, when evaluating the impact of using step-down selection.

We used functions in R libraries Design and Hmisc (R Version 2.9.2, http://www.R-project.org, accessed 21 Jun 2010) to fit our logistic-regression models, graphically assess linearity assumptions and goodness-of-fit, and check for overly influential observations, collinear predictors, and extra-binomial variation (due to repeated observations on some radio-tagged animals). We used the validate function in the Design library to compute several measures of a model's predictive ability as well as to estimate the degree of overfitting ($1 - \gamma_1$, where γ_1 = slope shrinkage factor). For reporting purposes, we focused on Somers' D_{xy}, which gives the rank correlation between predicted probabilities and observed responses. Somers' D_{xy} is closely related to the area under the receiver-operating-characteristic curve (c), a widely used measure of diagnostic discrimination in binary-logistic models (i.e., $D_{xy} = 2[c - 0.5]$). A model is making random predictions when $D_{xy} = 0$. Predictions are perfectly

discriminating when $D_{xy} = 1$. The validate function uses a bootstrap procedure to calculate a biased-corrected version of D_{xy} as follows:

1. D_{xy} is calculated for the original dataset.
2. For each bootstrap replicate ($i = 1$–500 in our case):
   a. A training dataset $(Y^*_{1,i}, X^*_{1,i})$ and a separate test dataset $(Y^*_{2,i}, X^*_{2,i})$ are created by resampling observations with replacement.
   b. A logistic regression model is fit to the training dataset, resulting in a vector of regression parameter estimates $\hat{\beta}_i$.
   c. D_{xy} is calculated as the rank correlation between responses in the training dataset $(Y^*_{1,i})$ and their predicted probabilities, $\exp(X^*_{1,i}\hat{\beta}_i)/\{1 + \exp(X^*_{1,i}\hat{\beta}_i)\}$;
   d. D_{xy} is calculated as the rank correlation between responses in the test dataset $(Y^*_{2,i})$ and predicted probabilities obtained using $\hat{\beta}_i$ from the model fit to the training data, $\exp(X^*_{2,i}\hat{\beta}_i)/\{1 + \exp(X^*_{2,i}\hat{\beta}_i)\}$.
3. The average difference between the statistics in steps 2c and 2d is used as an estimate of optimism and subtracted from D_{xy} calculated in step 1.

Test and training datasets are also required to calculate estimates of shrinkage factors. For linear regression models, the procedure is straightforward:

1. Regression parameters $(\hat{\beta})$ are estimated with the training dataset;
2. These parameters are used to make predictions for the test dataset, $\hat{Y}_2 = X_2\hat{\beta}$; and
3. Shrinkage factors are estimated using a linear regression of Y_2 on $\hat{Y}_2$.

Harrell (2001) suggests a slightly modified version for logistic regression in which responses in the test dataset, Y_2 (always 0 or 1), are replaced by the linear predictor $(X_2\hat{\beta}_2)$ estimated by fitting a logistic regression model to the test data only (note the subscripted 2 on the regression parameter vector). Thus, the shrinkage factors (γ_o, γ_1) in step 3 are calculated using a linear regression of $X_2\hat{\beta}_2$ on $X_2\hat{\beta}_1$. As with the biased-corrected version of Somers' D_{xy}, the validate function estimates (γ_o, γ_1) by taking the average of these statistics across a series of bootstrapped test and training datasets.

Impact of Step-Down Model Selection

We also evaluated sightability models constructed using a popular variable-selection method. We started with a full model {VO + GRP + VO × GRP + CONIF + WIND + CLOUD + TEMP + SNOW} and then used step-down variable selection with absolute change in AIC (ΔAIC) < 0 as a stopping rule (i.e., variables were removed in a backward stepwise algorithm as long as ΔAIC for the reduced model was <0). We used function stepAIC in R library MASS to perform step-down variable selection, which resulted in model {VO + CONIF + SNOW}. In addition, we performed step-down selection on our 3-df model {VO(4k)}, which resulted in model {VO}. We used the validate function to calculate bias-corrected versions of Somers' D_{xy} for reduced models {VO} and {VO + CONIF + SNOW}, accounting for the data-driven model selection process. To accomplish this goal, one must modify step 2a (in the bias-corrected Somers' D_{xy} calculation) and step 1 (of the shrinkage factor calculation) to include the step-down selection procedure (i.e., for each bootstrap replicate).

Impact on Population Estimates and Their Uncertainty

We used sightabilty-abundance estimators $(\hat{\tau})$ developed in Steinhorst and Samuel (1989), which we applied to operational-survey data from 2005 to 2007. Steinhorst and Samuel (1989) suggested an estimator for var$(\hat{\tau})$, but with no claim of unbiasedness (Thompson and Seber 1994, Wong 1996, Thompson 2002). Thompson and Seber (1994) derived an unbiased estimator for var$(\hat{\tau})$ when detection probabilities are assumed known, and Wong (1996) developed an estimator for the case where detection probabilities are estimated. Wong (1996) and Cogan and Diefenbach (1998) suggested bootstrapping as an alternative approach. We illustrate a hybrid approach. Specifically, we used Thompson and Seber's (1994) estimator for known detection probabilities to account for sampling variability associated with selecting a sample of aerial plots, and a non-parametric bootstrap to account for uncertain detection probabilities. The bootstrap approach also allowed us to incorporate uncertainty from model selection into $\hat{\text{var}}(\hat{\tau})$. Our bootstrap approach consisted of five steps:

1. We constructed 10,000 bootstrapped sightability datasets by resampling marked moose with replacement from surveys conducted during 2005–2007. Sample distributions for some predictors were highly skewed, which for ordinal predictors (e.g., CONIF, CLOUD, SNOW) meant that some bin values (factor levels) had small sample sizes ($n \leq 18$). Thus, some bootstrapped datasets did not contain a full range of values for ordinal predictors. This could create problems when attempting to predict on new datasets (e.g., operational-survey data) that contain factor levels that do not appear in the model matrix. For simplicity and to avoid prediction problems (in step 3, below), we modeled ordinal predictors as numeric data with linear effects on the logit scale. Alternatively, we could have restricted our bootstrap datasets to replicates that contained a full range of values for each ordinal or nominal predictor, but this would have conditioned our estimates on a subset of possible sightability datasets.
2. We fit our 3-df model {VO(4k)} and two reduced models ({VO}, {VO + CONIF + SNOW}; ignoring uncertainty from variable selection) to each bootstrapped dataset, which produced 10,000 estimates of β (vector of regression coefficients) for each model. We did not observe any cases where models failed to converge, but our models were relatively simple and we did not have problems of quasi- or complete-separation in our sightability datasets.
3. We used each $\hat{\beta}$ (vector of estimated regression coefficients) to predict $\hat{\pi}$ (estimated probability of detection) for each moose group observed in MNDNR operational surveys (2005–2007). We recognized the problem that $\hat{\pi}$ approaching 0 creates when using a Horvitz–Thompson

estimator ($\hat{\tau}_\pi = \sum y_i/\pi_i$; Thompson 2002) to adjust counts (y_i; moose group observations) for estimated detectability ($\hat{\pi}_i$). For example, define $\theta = 1/\pi$ as an expansion factor for detectability (applied to each moose group detected in an operational survey). Values of θ increase exponentially for $\pi < 0.15$ and approached biologically questionable values in our case study when $\pi < 0.02$ (implying $\geq$50 moose in a sample plot). Furthermore, for fixed sample sizes, the precision of $\hat{\tau}$ decreases as $\overline{\pi}$ decreases and may approach unacceptable levels when $\overline{\pi}$ is small (Wong 1996). In such cases, a sightability model would be rejected based on poor precision of population estimates. Consequently, we used an indicator variable to identify replicates where $\min(\hat{\pi}) < 0.02$ (for any of the three operational surveys), and computed estimates of τ and $\text{var}(\hat{\tau})$ after dropping these replicates (similar to the common approach for handling occasional non-convergence in bootstrapped model fits). In our case study, the proportion of replicates where $\min(\hat{\pi}) < 0.02$ was small (<1% of total replicates; range among models: 0–4%).

4. Using estimators from Thompson and Seber (1994) and Thompson (2002) for known detection probabilities, we computed 10,000 estimates of $\tau|\hat{\pi}$ and $\text{var}(\hat{\tau}|\hat{\pi})$ for each survey year (2005–2007) and sightability model. The MNDNR operational survey employed a stratified sampling design (Lenarz 2007). Thus, to avoid problems with correlated stratum-specific estimates (Fieberg and Giudice 2008), we estimated τ and its variance by applying each sightability model to the full dataset (vs. summing stratum-specific estimates).
5. We estimated $\text{var}(\hat{\tau})$ using a well-known conditional variance formula {$\text{Var}(y) = E_x[\text{var}(y|x)] + \text{var}_x[E(y|x)]$} (Casella and Berger 1990, Thompson and Seber 1994) and the bootstrap replicates (from step 4) to account for uncertainty in $\hat{\pi}$'s:

$$\hat{\text{var}}(\hat{\tau}) = \sum_{i=1}^{B} \frac{(\text{var}(\hat{\tau}|\hat{\pi})_i)}{B} + \sum_{i=1}^{B} \frac{\left[(\hat{\tau}|\hat{\pi})_i - \sum_{i=1}^{B} (\hat{\tau}|\hat{\pi})_i/B\right]^2}{(B-1)},$$

where B was the number of bootstrap replicates. Lastly, we used the estimated variances to calculate confidence intervals (under a normality assumption).

We evaluated model-selection uncertainty associated with the AIC step-down procedure by repeating steps 2–5 starting with the models {VO(4k)} and {VO + GRP + VO × GRP + CONIF + WIND + CLOUD + TEMP + SNOW}, but included step-down variable selection (with $\Delta\text{AIC} < 0$) in step 2. We also used the bootstrap to compute model-averaged estimates of τ ($\hat{\tau}_{MA}$) based on our $R = 30$ model set constructed using maximum df = 3 (see Df Spending Approach: Choosing How to Spend Df, above). A bootstrap usually is not needed to estimate $\text{var}(\hat{\tau}_{MA})$ in Burnham and Anderson's (2002) information-theoretic approach. However, we used a bootstrap in this case to account for the multiple sources of uncertainty in population estimates (sampling, model, sightability, and model-selection). For each bootstrap sightability dataset, we computed model-averaged estimates of $\hat{\tau}|\hat{\pi}$ and $\text{var}(\hat{\tau}|\hat{\pi})$ for the operational-survey data (2005–2007) using analytical formulas in Burnham and Anderson (2002: equations 4.1 and 6.12, respectively) and Akaike weights w_i estimated with AIC. We used AIC rather than AIC_c (for small n) to be consistent with the AIC step-down procedure and because in the case of a binary response it is not clear which sample size (n or $\min[n_0, n_1]$) should be used in the AIC_c formula (Harrell 2001:203; also see Burnham and Anderson 2002:332). Finally, we used the conditional variance formula (described in step 5, above) to compute $\hat{\text{var}}(\hat{\tau}_{MA})$ for each survey year.

RESULTS

We surveyed 78 test plots containing 1–5 radio-marked moose during January 2005–2007, which provided 124 moose-group observations for sightability modeling (Table 2). The annual proportion of marked moose detected during 2005–2007 averaged 0.476 and ranged from 0.375 to 0.564 (Table 2).

Sightability Models

Model {VO(4k)} fit the data substantially better (Δdeviance [absolute change in deviance] = 25.9, $\Delta\text{AIC} = 19.9$) than an intercept-only model and there was little evidence of overdispersion (binomial and robust sandwich variance estimates [treating animals as the experimental unit] differed by a factor of only 0.86 to 1.18) or general lack of fit. Furthermore, bootstrap estimates of expected predictive accuracy and degree of overfitting indicated that model {VO(4k)} had some utility for prediction ($D_{xy} = 0.47$) and estimated shrinkage due to model overfitting was small ($1 - \gamma_1 = 0.08$). One might have used AIC or a Wald statistic to simplify model {VO(4k)}. For example, applying step-down selection (with $\Delta\text{AIC} < 0$) resulted in model {VO}. However, there was some model-selection uncertainty (e.g., model {VO} was selected over {VO(4k)} in only 65% of

Table 2. Helicopter surveys of radio-marked moose in northeastern Minnesota, 2005–2007.

Year	Survey duration	Survey days	Test plots[a]	Radio-marked moose	Marked moose-group observations			
					n	Missed	Seen	Visibility
2005	5–27 Jan	6	25	29	39	17	22	0.564
2006	13–25 Jan	6	27	37	37	18	19	0.514
2007	9–24 Jan	7	26	35	48	30	18	0.375
2005–2007		19	78	55	124	65	59	0.476

[a] 4.0 km × 4.3-km plots containing at least one radio-marked moose.

the bootstrap datasets). Step-down selection on our full model (df = 8) resulted in {VO + CONIF + SNOW}, which had greater AIC support than model {VO(4k)} (ΔAIC = 6.1) or {VO} (ΔAIC = 3.7). However, expected predictive accuracy ($D_{xy} = 0.47$) was similar and degree of overfitting ($1 - \gamma_1 = 0.21$) was worse than for model {VO(4k)}, and model-selection uncertainty was much larger (e.g., model {VO + CONIF + SNOW} was selected in only 7% of the bootstrap datasets).

There was no clear best model in our set of $R = 30$ models used to calculate model-averaged predictions (e.g., max. Akaike weight = 0.11 for model {VO + CONIF + SNOW}). The model-averaged regression parameter for VO was slightly larger in absolute value than the coefficient in the {VO + CONIF + SNOW} model (−0.033 vs. −0.031). By contrast, model-averaged regression parameters for CONIF and SNOW were smaller in absolute value when compared to their values in the {VO + CONIF + SNOW} model (−0.228 vs. −0.560 for CONIF and 0.169 vs. 0.438 for SNOW).

The sightability models we fit described slightly different mean functions (expected detection probabilities) with respect to changes in VO while holding other predictors at median values (e.g., Fig. 1). The detection curve for the simplest model {VO} was below that of model {VO + CONIF + SNOW}, which was chosen via AIC step-down selection from the full model (df = 8) and also was the best model (Akaike weight = 0.11) in the $R = 30$ model set. The model-averaged detection curve was in between curves for {VO} and {VO + CONIF + SNOW} (Fig. 1). When applied to operational-survey data, mean $\hat{\pi}$ (averaged over bootstrap replicates and years) ranged from 0.546 for model {VO} to 0.643 for model {VO + CONIF + SNOW}.

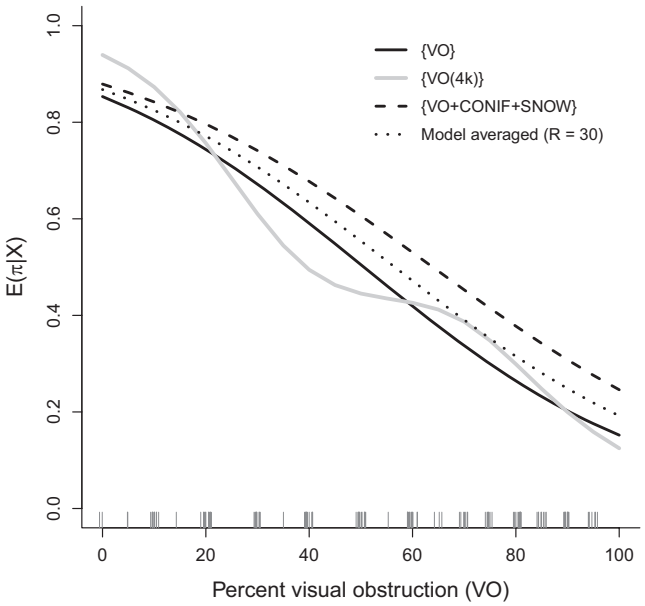

Figure 1. Mean prediction functions (detection curves) for sightability models developed from helicopter surveys of radio-marked moose in northeastern Minnesota, 2005–2007. Curves depict the change in expected detection probability as VO increases from 0% to 100%, while holding other predictors at median values observed in 124 sightability trials. VO(4k) = VO modeled as a restricted-cubic spline with four knots; CONIF = contribution of conifers to screening cover (1 = little or none, 2 = mixed conifers and hardwoods, 3 = mostly conifers); and SNOW = snow depth (<20 cm, 20–41 cm, >41 cm). Model-averaged predictions were based on a set of $R = 30$ sightability models constructed using VO in each model and 0–2 additional predictors. The *x*-axis rug depicts the distribution of VO values (jittered) in 124 sightability trials.

Population Estimates

The sightability models produced similar point estimates of population size when applied to operational-survey data from 2005 to 2007, whereas precision varied by year and model (e.g., Fig. 2). Precision of population estimates was similar for model {VO(4k)} (mean CV = 20.3%), model-averaged estimates (mean CV = 24.9%), and the two reduced models when model-selection uncertainty was ignored (mean CV = 19.3% and 19.6%). Step-down selection on model {VO(4k)} had little effect on precision of population estimates, whereas step-down selection on the full model (predictor df = 8) substantially increased the confidence intervals (Fig. 2).

DISCUSSION

Widespread availability of statistical software and advances in computing power allow researchers to easily fit a variety of

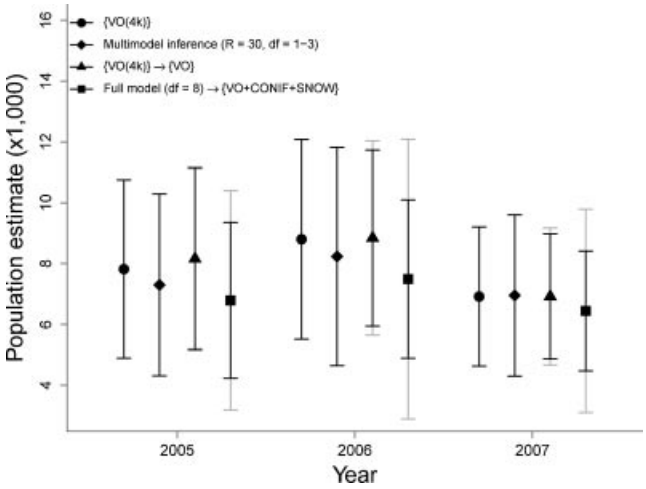

Figure 2. Population estimates ($\hat{\tau}$) and 90% confidence intervals based on sightability-model population estimators applied to operational surveys of moose in northeastern Minnesota, 2005–2007. Point estimates were derived from logistic-regression models fit to the original sightability dataset ($n = 124$ sightability trials), whereas $\widehat{\text{var}}(\hat{\tau})$ and associated confidence intervals were based on fitting logistic-regression models to 10,000 bootstrapped sightability datasets. Bootstrap replicates where the minimum estimated probability of detection was <0.02 for any of the 3-survey years were dropped prior to estimating τ (<1% of total replicates; range among models: 0–4%). Model {VO(4k)} was constructed using degrees-of-freedom (df) spending with max df = 3. Multi-model inference was based on $R = 30$ sightability models, all of which contained predictor VO and 0–2 additional predictor variables (regression df, excluding intercept = 1–3). Models {VO} and {VO + CONIF + SNOW} were chosen by applying Akaike's Information Criterion (AIC) step-down procedures to model {VO(4k)} and a full model containing 8 predictors, respectively. Gray error bars include model-selection uncertainty arising from step-down selection using absolute change in AIC = 0 as a stopping rule (black intervals do not). Model notation: VO = percent visual obstruction, VO(4k) = VO modeled as restricted-cubic spline with four knots, CONIF = relative contribution of conifers to screening cover, and SNOW = snow depth (three ordinal classes).

regression models, and it is not uncommon for researchers to consider a large number of possible predictor variables, transformations of variables, interactions, nonlinearities, etc. (Chatfield 2002). Increasing the degree of model flexibility, however, has important implications for the level of confidence we should ascribe to our analytic results (Babyak 2004). At one extreme is exploratory approaches that consider many predictors, with conclusions largely determined by iterative, data-driven processes. Inference from a prespecified model with a large enough sample size to allow adequate modeling of all important predictor variables represents the other extreme (Babyak 2004). Nonetheless, this latter ideal will be difficult to achieve with most observational studies due to the large number of confounding variables that can influence system responses. Thus, there will often be a tradeoff between looking at too many variables (which can result in overfitting) and not considering enough variables (i.e., key results could change if important confounding variables were included). We believe analysts are more likely to err on the side of looking at too many predictors rather than too few and it is important to recognize the potential cost of this choice (Babyak 2004).

Model-selection methods, in particular, come with a price; *P*-values, confidence intervals, and measures of model fit do not have their normal interpretation since the chosen model was not specified a priori (Altman and Andersen 1989, Grambsch and O'Brien 1991, Harrell 2001, Mundry and Nunn 2009, Dahlgren 2010). In addition, when the ratio of observations to predictors is low, it is easy to overfit the data (i.e., choose a model that fits the current dataset well but predicts poorly on new data; for example, see Babyak 2004). In our case study, estimated shrinkage ($1 - \gamma_1$) increased by 13% ($\Delta[1 - \gamma_1] \times 100$) with a relatively small increase in number of candidate predictors (from 3 to 8). This suggests a high degree of overfitting with 8 df. Likewise, confidence intervals that did not account for model selection were too small, especially when the number of candidate predictors exceeded df guidelines (Fig. 2).

In response to these problems, Burnham and Anderson (2002) argued for more a priori thinking when determining which models to fit, and they proposed using information-theoretic methods (e.g., AIC) and model averaging to overcome problems associated with model-selection uncertainty. We illustrated an alternative approach, namely to rely on the fit of a full model for inference, with the number of predictors limited by effective sample size. Consistent with the philosophy of Burnham and Anderson (2002), this approach forces the investigator to think carefully about their models and justify the inclusion of various predictors. Furthermore, the typical need for some data reduction focuses attention on exploring predictor variables (without looking at their relationship with *Y*) to determine their potential value; predictors with many missing values, limited ranges, large measurement error, or that are highly correlated with other predictors are likely to be excluded from consideration. It also emphasizes estimation of effect sizes and their uncertainty rather than using a rule to determine if variables are important or not, for example, based on their inclusion in the best model or set of models. Lastly, in our case study where we suspected analytical variance formulas were biased (Wong 1996, Cogan and Diefenbach 1998), df spending was computationally cheaper to implement (than multi-model averaging) with a bootstrap approach to inference. Similar advantages would apply to problems involving clustered data where bootstrapping provides a simple means of addressing correlated data issues (e.g., Zicus et al. 2006).

Yet, a possible benefit of using the multi-model inference procedures outlined in Burnham and Anderson (2002) is that they may allow one to examine more predictors in total while guarding against overfitting (e.g., our $R = 30$ model set included a total of eight predictor variables and one interaction). That is, because model averaging shrinks regression coefficients toward zero; it reduces the effective df spent (e.g., Harrell 2001:80). It is important to recognize, however, that the degree of coefficient shrinkage will depend on the candidate set of models. For example, each of our $R = 30$ candidate models included VO based on the a priori belief that it was the most important predictor, whereas other predictors appeared in some models but not others. As a result, model averaging did not shrink the coefficient for VO, but model averaged coefficients for SNOW and CONIF were 59–61% smaller than their corresponding values in the best approximating model. Frequently, multi-model inference techniques are employed as a means to deal with the "too many predictors not enough responses" problem (i.e., to perform variable selection) rather than to test truly competing mechanistic hypotheses. More work should be done to evaluate methods for determining appropriate model sets in these situations, but we suggest a reasonable strategy might be: 1) include in all models any variables that are of direct and primary scientific interest (so their coefficients will not be shrunk toward zero); 2) pair these primary predictors with other nuisance or confounding variables in a systematic fashion so the latter occur in roughly the same number of models; and 3) limit df in any one model using df-spending guidelines (Appendix A, Table A.1).

We often recommend a df-spending approach in our consulting work because it naturally forces investigators to think carefully about their models and predictors. However, model averaging (Burnham and Anderson 2002, O'Hara and Sillanpaa 2009, Lukacs et al. 2010) and shrinkage estimators (Harrell 2001, Dahlgren 2010) are also viable alternatives to aggressive model-selection procedures, especially when building predictive models from observational datasets with many candidate predictors. The initial steps involved in the df-spending approach (e.g., prescreening predictors for missing values, narrow data ranges, collinearities, etc.) are also important to apply with these latter approaches. In addition, sufficient time should be allocated to exploring fitted models (e.g., inspecting model assumptions and diagnostics), and the biological significance of predictor variables should be evaluated using estimated effect sizes with credible measures of uncertainty. Burnham and Anderson (2002) frequently emphasized these points in their book. Unfortunately, our experience (as statistical consultants and journal reviewers) has been that researchers appear to

spend less time inspecting models and interpreting effect sizes when using multi-model inferences procedures. Furthermore, multi-model inference procedures are still relatively new and they are not always straightforward to apply (e.g., estimating the covariance between model-averaged parameters can be complicated, and developing appropriate model-averaged parameter estimates can be challenging when models include nonlinear terms and interactions; Burnham and Anderson 2002:153, 344).

Ultimately, it seemed reasonable to use reduced model {VO} for prediction in future operational moose surveys because it was logical and simple (desirable properties of sightability models), model-selection uncertainty did not appreciably influence precision of population estimates (Fig. 2), and estimated detection probabilities reflected an "average" detection function with all bootstrap replicates resulting in minimum $\hat{\pi} > 0.02$. Another reasonable choice would be to use model-averaged $\hat{\pi}$'s (i.e., multi-model inference given a small, sensible set of candidate models based on df guidelines). However, in this particular application: 1) model-averaged population estimates and precision were similar to or larger than our 3-df model (Fig. 2); 2) it would require measuring eight predictor variables in operational surveys; and 3) estimating the unconditional variance of $\hat{\tau}$ from a Horvitz–Thompson estimator with multiple sources of variation (sampling, model, sightability, model selection) required a bootstrap approach that was computationally intensive (e.g., our bootstrap with B = 10,000 and a small set [$R = 30$] of relatively simple models required over 300 hr of dedicated time on a personal computer with a 3.0 gigahertz chip and 3.25 gigabytes of memory).

In contrast to our df-spending approach, Anderson and Lindzey (1996) and Quayle et al. (2001) used aggressive variable-selection methods to develop sightability models for helicopter surveys of moose in Wyoming, USA, and British Columbia, Canada, respectively. Despite the shortcomings of variable-selection methods (Dahlgren 2010), these researchers ended up with sightability models that were very similar to our model {VO}, except screening cover was an ordinal factor. Likewise, Drummer and Aho (1998) used aggressive variable-selection methods to choose a relatively simple sightability model (predictor df = 3, including screening cover) for fixed-wing surveys of moose in Michigan, USA. Thus, aggressive variable-selection methods led to reasonable models in these cases. This result may reflect the relative importance of screening cover as a predictor of detectability in many aerial big-game surveys. In general, predictors with considerable explanatory power will be included in best models regardless of the model selection approach (Murtaugh 2009). However, the danger of aggressive variable-selection methods is that chosen models may also include predictors that explain random noise in the data (overfitting), particularly when the number of candidate predictors is large relative to the available sample size.

Starting with smaller, less complex models likely introduced some bias into our estimates of π (inclusion probabilities in the Horvitz–Thompson estimator) since there was greater risk of excluding important predictors. On the other hand, more complex sightability models were more likely to result in $\hat{\pi}$ values near zero, which could lead to imprecise estimates when using the Horvitz–Thompson estimator (e.g., Fig. 2). Therefore, the application of sightability models for population estimation likely involves a strong bias-variance tradeoff in which less biased estimators (containing more predictor variables) may result in larger mean-squared errors. These concerns suggest that model-assisted survey estimators might be worth exploring (e.g., a Bayesian approach that allows one to place a prior on $\hat{\pi}$'s, effectively shrinking them away from zero). Lastly, survey costs are also an important consideration (Noyes et al. 2000, Rabe et al. 2002, Giudice et al. 2010) and may increase if a large suite of predictor variables must be recorded at each sighting. Thus, we concluded that a sightability model based solely on VO had many advantages.

MANAGEMENT IMPLICATIONS

Wildlife-management decisions should be based on sound scientific principles and informed by data whenever possible. The desire to accurately model complex ecological systems often results in datasets that include many predictors, but logistical and financial challenges may preclude collecting these data on many sample units. As a result, researchers are often faced with the challenge of drawing reliable conclusions from small observational datasets with many potential predictors and relatively few observations per predictor. Clearly, choosing an appropriate analysis strategy will depend on the context of the problem (e.g., effect estimation, prediction, hypothesis testing, exploratory data analysis). However, a generic strategy for developing reliable predictive models (e.g., Appendix A) will also generally apply to model development for effect estimation and hypothesis testing (Harrell 2001:82). Thus, for many prediction and estimation problems, we argue that inference from a single pre-specified model or model averaging over a small set of predictive models, after first limiting the analysis a priori to include only the most promising variables (e.g., those with few missing values, low correlations with other predictors, and strongest support based on previous studies or underlying biological principles), provides a sensible strategy that overcomes many limitations associated with aggressive model-selection strategies (e.g., optimistic measures of uncertainty and overfit models that predict new data poorly). Obviously, exploratory approaches can also contribute to scientific evidence and often form the basis for hypothesis generation and future research. However, we need to keep in mind the limitations of exploratory analyses and models, especially with respect to producing replicable results (e.g., Babyak 2004).

ACKNOWLEDGMENTS

The MNDNR, the Fond du Lac Band of Lake Superior Chippewa, and the 1854 Treaty Authority provided funding and field support for the sightability and operational moose surveys. The United States Geological Survey, Northern Prairie Wildlife Research Center provided in-kind support. The United States Fish and Wildlife Service's Tribal

Wildlife Grants Program provided additional funding. We thank K. Carlisle, A. Edwards, D. Litchfield, M. Nelson, T. Rusch, B. Sampson, M. Schrage, and MNDNR pilots A. Buchert, J. Heineman, M. Trenholm, and B. Maas for aerial-telemetry and survey contributions. We are grateful to D. Johnson, S. McCorquodale, and 2 anonymous reviewers for providing constructive comments on earlier drafts. However, opinions expressed in the manuscript are the responsibility of the authors and may not reflect the opinions of reviewers.

LITERATURE CITED

Altman D. G., and P. K. Andersen. 1989. Bootstrap investigation of the stability of a Cox regression model. Statistics in Medicine 8:771–783.

Anderson C. R., and F. G. Lindzey. 1996. Moose sightability model developed for helicopter surveys. Wildlife Society Bulletin 24:247–259.

Anderson, C. R., D. S. Moody, B. L. Smith, F. G. Lindzey, and R. P. Lanka. 1998. Development and evaluation of sightability models for summer elk surveys. Journal of Wildlife Management 62:1055–1066.

Babyak, M. A. 2004. What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. Psychosomatic Medicine 66:411–421.

Bartmann, R. M., G. C. White, L. H. Carpenter, and R. A. Garrott. 1987. Aerial mark-recapture estimates of confined mule deer in pinyon-juniper woodland. Journal of Wildlife Management 51:41–46.

Buckland, S. T., D. R. Anderson, K. P. Burnham, J. L. Laake, D. L. Borchers, and L. Thomas. 2001. Introduction to distance sampling: Estimating abundance of biological populations. Oxford University Press, New York, New York, USA.

Burnham K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: A practical information-theoretic approach, Second edition. Springer-Verlag, New York, New York, USA.

Casella G., and R. G. Berger. 1990. Statistical inference. Wadsworth, Pacific Grove, California, USA.

Caughley, G. 1974. Bias in aerial survey. Journal of Wildlife Management 38:921–933.

Chatfield, C. 2002. Confessions of a pragmatic statistician. The Statistician 51:1–20.

Cogan R. D., and D. R. Diefenbach. 1998. Effect of undercounting and model selection on a sightability-adjustment estimator for elk. Journal of Wildlife Management 62:269–279.

Copas, J. B. 1997. Using regression models for prediction: Shrinkage and regression to the mean. Statistical Methods in Medical Research 6:167–183.

Copas J., and T. Long. 1991. Estimating the residual variance in orthogonal regression with variable selection. The Statistician 40:51–59.

Dahlgren, J. P. 2010. Alternative regression methods are not considered in Murtaugh (2009) or by ecologists in general. Ecology Letters 13:E7–E9.

Derksen S., and H. J. Keselman. 1992. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. British Journal of Mathematical and Statistical Psychology 45:265–282.

Drummer T. D., and R. W. Aho. 1998. A sightability model for moose in upper Michigan. Alces 34:15–19.

Faraway, J. J. 1992. On the cost of data analysis. Journal of Computational and Statistical Graphics 1:213–229.

Fieberg J., and J. Giudice. 2008. Variance of stratified survey estimators with probability of detection adjustments. Journal of Wildlife Management 72:837–844.

Gasaway, W. C., S. D. DuBois, D. J. Reed, and S. J. Harbo. 1986. Estimating moose population parameters from aerial surveys. Biological Papers of the University of Alaska 22.

Giudice, J. H., J. R. Fieberg, M. C. Zicus, D. P. Rave, and R. G. Wright. 2010. Cost and precision functions for aerial quadrat surveys: A case study of ring-necked ducks in Minnesota. Journal of Wildlife Management 74:342–349.

Grambsch P. M., and P. C. O'Brien. 1991. The effects of transformations and preliminary tests for non-linearity in regression. Statistics in Medicine 10:697–709.

Harrell, F. E. Jr. 2001. Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis. Spring-Verlag, New York, New York, USA.

Heinselman, M. 1996. The boundary waters wilderness ecosystem. University of Minnesota Press, Minneapolis, USA.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. Bayesian model averaging: A tutorial. Statistical Science 14:382–417.

Lenarz, M. S. 1998. Precision and bias of aerial moose surveys in northeastern Minnesota. Alces 34:117–124.

Lenarz, M. S. 2007. 2007 Aerial moose survey. Minnesota Department of Natural Resources, St. Paul, USA. <http://files.dnr.state.mn.us/recreation/hunting/moose/moose_survey_2007.pdf> Accessed 30 Jun 2010.

Lenarz, M. S., M. E. Nelson, M. W. Schrage, and A. J. Edwards. 2009. Temperature mediated moose survival in northeastern Minnesota. Journal of Wildlife Management 73:503–510.

Lukacs, P. M., K. P. Burnham, and D. R. Anderson. 2010. Model selection bias and Freedman's paradox. Annals of the Institute of Statistical Mathematics 62:117–125.

McCorquodale, S. M. 2001. Sex-specific bias in helicopter surveys of elk: Sightability and dispersion effects. Journal of Wildlife Management 65: 216–225.

Minnesota Department of Natural Resources [MNDNR]. 2007. Ecological classification system. Minnesota Department of Natural Resources, St. Paul, USA. <http://www.dnr.state.mn.us/ecs/index.html> Accessed 18 Jun 2010.

Mundry R., and C. L. Nunn. 2009. Stepwise model fitting and statistical inference: Turning noise into signal pollution. American Naturalist 173:119–123.

Murtaugh, P. A. 2009. Performance of several variable-selection methods applied to real ecological data. Ecology Letters 12:1061–1068.

Noyes, J. H., B. K. Johnson, R. A. Riggs, M. W. Schlegal, and V. L. Coggins. 2000. Assessing aerial survey methods to estimate elk populations: A case study. Wildlife Society Bulletin 28:636–642.

O'Hara R. B., and M. J. Sillanpaa. 2009. A review of Bayesian variable selection methods: What, how, and which? Bayesian Analysis 4:85–118.

Peek, J. M., R. E. LeResche, and D. R. Stevens. 1974. Dynamics of moose aggregations in Alaska, Minnesota, and Montana. Journal of Mammalogy 55:126–137.

Pollock K. H., and W. L. Kendall. 1987. Visibility bias in aerial surveys: A review of estimation procedures. Journal of Wildlife Management 51:502–510.

Potvin, F., L. Breton, L. P. Rivest, and A. Gingras. 1992. Application of a double-count aerial survey technique for white-tailed deer, *Odocoileus virginianus*, on Anticosti Island, Quebec. Canadian Field-Naturalist 106:435–442.

Quayle, J. F., A. G. Machutchon, and D. J. Jury. 2001. Modeling moose sightability in southcentral British Columbia. Alces 37:43–54.

Rabe, M. J., S. S. Rosenstock, and J. C. deVox. Jr., 2002. Review of big-game survey methods used by wildlife agencies of the western United States. Wildlife Society Bulletin 30:46–52.

Rice W. R., and J. D. Harder. 1977. Application of multiple aerial sampling to a mark-recapture census of white-tailed deer. Journal of Wildlife Management 41:197–206.

Rice, C. G., K. J. Jenkins, and W. Chang. 2009. A sightability model for mountain goats. Journal of Wildlife Management 73:468–478.

Samuel M. D., and K. H. Pollock. 1981. Correction of visibility bias in aerial surveys where animals occur in groups. Journal of Wildlife Management 45:993–997.

Samuel, M. D., E. O. Garten, M. W. Schlegel, and R. G. Carson. 1987. Visibility bias during aerial surveys of elk in northcentral Idaho. Journal of Wildlife Management 51:622–630.

Samuel, M. D., R. K. Steinhorst, E. O. Garton, and J. W. Unsworth. 1992. Estimation of wildlife population ratios incorporating survey design and visibility bias. Journal of Wildlife Management 56:718–725.

Steinhorst R. K., and M. D. Samuel. 1989. Sightability adjustment methods for aerial surveys of wildlife populations. Biometrics 45:415–425.

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

Sun, G. W., T. L. Shook, and G. L. Kay. 1996. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. Journal of Clinical Epidemiology 49:907–916.
Thompson, S. K. 2002. Sampling. Second edition. John Wiley & Sons, New York, New York, USA.
Thompson S. K., and G. A. F. Seber. 1994. Detectability in conventional and adaptive sampling. Biometrics 50:712–724.
van Houwelingen, J. C. 2001. Shrinkage and penalized likelihood as methods to improve predictive accuracy. Statistica Neerlandica 55:17–34.
van Houwelingen J. C., and S. le Cessie. 1990. Predictive value of statistical models. Statistics in Medicine 8:1303–1325.
Walsh, D. P., C. F. Page, H. Campa, III, S. R. Winterstein, and D. E. Beyer. Jr., 2009. Incorporating estimates of group size in sightability models for wildlife. Journal of Wildlife Management 73:136–143.
Whittingham, M. J., P. A. Stephens, R. B. Bradbury, and R. P. Freckleton. 2006. Why do we still use stepwise modelling in ecology and behaviour? Journal of Animal Ecology 75:1182–1189.
Wong, C. 1996. Population size estimation using the modified Horvitz-Thompson estimator with estimated sighting probability. Dissertation, Colorado State University, Fort Collins, USA.
Zicus, M. C., D. P. Rave, and J. R. Fieberg. 2006. Cost-effectiveness of single- versus double-cylinder over-water nest structures. Wildlife Society Bulletin 34:647–655.

Appendix A. Degrees-of-freedom (df) spending approach

The df-spending approach we advocate is largely adopted from Harrell (2001). Although Harrell (2001) outlined slightly different strategies for the purposes of prediction, effect estimation, and hypothesis testing, the core principles (i.e., related to the idea of df spending) remain the same regardless of the intended application. The primary difference among strategies is that Harrell (2001) suggested some limited (and structured) model selection may be worthy of consideration when the goal is prediction, provided that uncertainty arising from any data driven decisions are accounted for using the bootstrap. Thus, Harrell (2001:79) stated, "it is only a mild oversimplification to say that a good overall strategy is to decide how many degrees can be 'spent,' where they should be spent, and then to spend them. If statistical tests or confidence limits are required, later reconsideration of how d.f. are spent is not usually recommended." We attempt to highlight the core steps of the df-spending approach below.

1. Determine an appropriate level of model complexity (or flexibility). Use a likelihood ratio approach (Box A.1) or effective sample-size guidelines to determine df to spend. General guidelines for avoiding model overfitting is to limit the df associated with predictors (including complex terms such as interactions and nonlinear terms) to $m/10$ or $m/20$, where m is the limiting sample size (Table A.1).
2. Allocate df to available predictors.
   a. Develop (mechanistically based) a priori hypotheses, exploiting previous knowledge and data, to suggest which predictors are likely to influence system responses and whether there are any plausible interactions that should be considered.
   b. Prescreen predictors (without looking at their relationship to the response, Y) to potentially eliminate non-informative variables. Consider:
      i. Dropping variables that have many missing observations (but also note that imputation strategies may be important for variables that have moderate levels of missing data).
      ii. Dropping variables that are highly correlated; combining similar variables into a single index (e.g., using principle components analysis).
      iii. Dropping variables that exhibit little variability.
      iv. If sample size permits, allocating multiple df to the most important variables, thus allowing for nonlinear effects. Harrell (2001:52–56) also discussed a rank-correlation approach that could be used to allocate

Table A.1. Limiting sample sizes for various response variables[a].

Type of response variable	Limiting sample size, m[b]
Continuous	n (total sample size)
Binary	$\min(n_0, n_1)$
Ordinal (k categories)	$n - 1/n^2 \sum_{i=1}^{k} n_i^3$
Failure (survival) time	Number of failures

[a] © 2001 Springer-Verlag New York, Inc. Reproduced with permission from Springer Science + Business Media: Regression modeling strategies, Chapter 4: multivariable modeling strategies, 2001, page 61, Frank E. Harrell Jr., Table 4.1.
[b] See Harrell (2001:61) for justifications and more detailed explanations.

Box A.1: Determining available df using the likelihood ratio statistic for the full model.

Harrell (2001:73) describes an approach to determining degrees of freedom (df) available to spend based on a heuristic shrinkage estimate ($\hat{\gamma}$) developed by van Houwelingen and le Cessie (1990): $\hat{\gamma} = (\text{model } \chi^2 - p)/(\text{model } \chi^2)$, where p is the total predictor df and model χ^2 is the likelihood ratio χ^2 statistic for testing the null hypothesis that coefficients associated with predictors are all equal to 0. To apply the approach:

1. Fit a full model with all predictor variables under consideration, along with potential interactions and nonlinear terms. Let p indicate the total predictor df in this model and let LR be the likelihood ratio χ^2 statistic for this model. Do *not* inspect the model to determine which candidate predictors appear to have merit in explaining the response.
2. Using $\hat{\gamma}$ calculate the shrinkage $(1-\hat{\gamma})$ expected when applying this model to new data. If the expected shrinkage is deemed acceptable (e.g., ≤ 0.10), proceed with the full set of predictors. If not, go to step 3.
3. Reduce the candidate df to (LR-p)/9 or, more conservatively, (LR-2p)/8. The former represents the degree of data reduction necessary for shrinkage not to exceed 10% under a best-case scenario where the variables eliminated from further consideration are completely unrelated to the response (note again, variables should be eliminated *without* looking at the relationship between predictors and the response). The latter, more conservative bound assumes eliminated variables reduce the LR statistic only to the extent that the reduced model has the same AIC as the full model. Lastly, note that if $\text{LR} < p + 9$, then there is little hope that a reduced model will calibrate well on new data.

multiple df once a list of model predictors has been fully specified.

3. Fit the model. Assess validity of model assumptions, predictive ability, and extent of overfitting (and shrinkage) using bootstrap methods.
   a. Use residual plots and standard regression diagnostics to look for poorly met assumptions, outliers, and influential data points.
   b. Use the bootstrap approach described in the text (e.g., implemented using the validate function) to estimate shrinkage coefficients and to obtain bias-corrected measures of model performance. If changes to the model were made as a result of looking at diagnostics in step 3a, account for uncertainty due to these data-driven decisions by repeating the decision process for each bootstrap replicate.

Following step 3, one should interpret the fitted model graphically. For example, by plotting predicted responses as a function of each covariate (while holding other covariates at mean or median values—see Figure 1 in the main text). Additional models may also be considered, particularly if the df-spending model performs poorly in step 3. However, predictions of these latter models should be viewed more cautiously than pre-specified models.

Associate Editor: Scott M. McCorquodale.