The Journal of Wildlife Management 79(5):708–718; 2015; DOI: 10.1002/jwmg.894

*Multimodel Inference Special Section*

# MMI: Multimodel Inference or Models With Management Implications?

**JOHN FIEBERG,**[1] *Department of Fisheries, Wildlife and Conservation Biology, University of Minnesota, 2003 Upper Buford Circle, Suite 135, Saint Paul, MN 55108, USA*

**DOUGLAS H. JOHNSON,** *U.S. Geological Survey, Northern Prairie Wildlife Research Center, 2003 Upper Buford Circle, Suite 135, Saint Paul, MN 55108, USA*

**ABSTRACT** We consider a variety of regression modeling strategies for analyzing observational data associated with typical wildlife studies, including all subsets and stepwise regression, a single full model, and Akaike's Information Criterion (AIC)-based multimodel inference. Although there are advantages and disadvantages to each approach, we suggest that there is no unique best way to analyze data. Further, we argue that, although multimodel inference can be useful in natural resource management, the importance of considering causality and accurately estimating effect sizes is greater than simply considering a variety of models. Determining causation is far more valuable than simply indicating how the response variable and explanatory variables covaried within a data set, especially when the data set did not arise from a controlled experiment. Understanding the causal mechanism will provide much better predictions beyond the range of data observed. Published 2015. This article is a U.S. Government work and is in the public domain in the USA.

**KEY WORDS** causation, effect size, mechanism, model selection, multicollinearity, multimodel inference, observational data, overfitting, prediction, statistics.

Consider the following—very realistic—scenario. Because of the concern about declining populations of many grassland birds, investigators decide to study a number of grassland fields in the Midwest. Their objective is to identify features conducive to supporting high densities of grassland birds. They select, as randomly as feasible, a number of fields to study and record the number of each species present in each field during the birds' breeding season. They also record a number of potential explanatory variables. Some of these they obtain from remote sensing, such as the area of contiguous grassland embedding each field; proximity to trees; and the percentages of tree cover, grassland cover, wetland cover, and cropland cover within 400 m, 800 m, and 1,200 m of a field. Some explanatory variables are measured in the field, including vegetation height-density measures (Robel readings; Robel et al. 1970) and their coefficient of variation (to measure heterogeneity), and percentage cover by grasses, forbs, woody species, litter, and bare ground along with their coefficients of variation. They also record the latitude and longitude of each study field, which, along with their interaction, could account for variation in bird density in relation to the breeding range of a particular species.

The investigators were able to collect complete information of 25 fields each year for 3 years. We assume that different fields were available each year so that we do not have to deal with repeated-measures issues. These 75 observations are analyzed in relation to the 29 potential explanatory variables. Note that the 29 does not include any possible quadratic or other nonlinear effects, or interactions among the explanatory variables (except latitude and longitude). The rather vague objective of the investigators' research, stated as "identify important features," typifies many wildlife studies. What are the challenges involved with meeting this objective? What role should multimodel inference play in addressing this objective? More generally, how should the investigators use these data to inform management, and are there other things the investigators should have considered when designing the study?

Received: 23 September 2014; Accepted: 16 March 2015
Published: 25 May 2015

[1]*E-mail: jfieberg@umn.edu*

## THE CHALLENGES

The investigators' objective may be translated into questions about what is the most appropriate model for the data, which in turn requires knowledge of how a model would be used. Among the many purposes of models, 2 are paramount: predicting and understanding. To distinguish these purposes, consider the following 2 questions:

1. Given knowledge of site-specific explanatory variables associated with a number of grassland sites, which is most likely to host a particular species of grassland bird?
2. What would happen to the abundance of a particular species of grassland bird if I apply some management action, such as removing trees or conducting a prescribed burn?

The first question involves prediction. We are interested in determining the output of a model (here, species presence), given specified inputs (values of explanatory variables). The

second question is much more challenging to address and requires a deeper understanding of how the explanatory variables will respond to the management action and how the response variable will be affected by those changes. For prediction, a black box will suffice; for understanding, we need to peer inside and look at the causal mechanisms. Explanation is more valuable to wildlife managers because it provides much more general guidance than prediction, which is safely applied only to sites with conditions similar to those used to build the model. Further, a model that explains a process and incorporates causal mechanisms can serve well for prediction, but the converse is not true.

A good example of contrasting modeling objectives involves predicting weather versus understanding climate. Deciding whether or not to bring an umbrella to work can be based on weather patterns to the west, the rates and directions fronts are moving, etc. Or the probability of rain can be estimated from calculating, from all days with recorded weather conditions very similar to the present, the portion of them with rain. Predicting climate, in contrast, requires an understanding of solar insulation and sunspot activity, the orbital variation of Earth, circulation of air and water masses, plate tectonics, volcanic activity, atmospheric composition, and, evidently, political affiliation. Models that represent a solid understanding of the system can in fact be used successfully to predict, but the converse is not true. Models for prediction are especially unreliable when extrapolating beyond the range of values used to fit the model. Models based on understanding the mechanisms involved are much more trustworthy over a wider range of explanatory variables.

For explanatory purposes, the main difficulty with the grassland bird analysis is that the study was observational, rather than experimental. Investigators did not have a large number of grasslands from which to randomly select some to be large, others small, some with no trees in their landscape, others with many trees, and so on. They had to select from available grassland fields. Although they could select some large and some small fields, and some with no nearby trees and others with many, they could not randomly assign these features to the fields. Because of this constraint, the possibility always exists that the fields selected varied in numerous other ways; observational studies typically involve a large number of possibly relevant variables that cannot be controlled by the investigators. The best investigators can do is to identify as many of these variables as they think might be influential, measure them, and attempt to account for the truly important ones in the course of the analysis. This poses a quandary. Barry Commoner's first law of ecology is that everything is connected to everything else (Commoner 1971). A corollary to that law is that any relationship posited between 2 ecological variables will become statistically significant with a sufficiently large sample.

Although ecological processes may be affected by everything else, we hope in an analysis to find the big chunks, those few variables that have the major effect on the response variable and—importantly—estimate the size of those effects. Estimates of effect sizes and their standard errors are more important products of an analysis than are *P*-values or changes in information criterion values associated with adding or dropping 1 or more explanatory variables from a model. An effect size ideally will indicate the magnitude and direction of the explanatory variable's influence on the response variable. Its standard error should indicate how much confidence to place in the estimate of that effect. And, most critically for managers, the effect size should inform us about the response anticipated if a particular action is taken. That's what wildlife management is about.

How should one go about trying to identify the big chunks? A variety of approaches are available for analyzing observational data involving numerous potential explanatory variables. Several forms of stepwise selection of explanatory variables allow the data to speak for themselves and let the computer select some optimal subset of variables. One could examine correlation coefficients between response variables and explanatory variables, or plots relating the 2 kinds of variables; these techniques fall under the category of exploratory data analysis. One might simply choose explanatory variables that similar studies already have found important. Or one could follow the guidance of Burnham and Anderson (2002:440) and simply think hard about the problem before beginning analysis.

Important to this discussion is that too much analysis can actually be a bad thing— "If you torture your data long enough, it will confess to anything" (a quote usually attributed to economist Ronald Coase). When models are too finely tuned to specific aspects of the data in hand, we say the model is overfit, in the sense that the model is accommodating, not only the actual effects of the explanatory variables, but also the noise inherent in any real data set (Mundry and Nunn 2009). Especially prone to overfitting are 1) overly complex models, those with too many explanatory variables relative to the available sample size; and 2) models that allow for very flexible relationships between explanatory and response variables (Harrell 2001). An extreme example involves fitting a polynomial of degree $n-1$ to a set of $n$ data points—the model will fit the data exactly, but predictions of future data likely will be poor.

One guideline for avoiding overfitting is to limit model degrees of freedom to 5% to 10% of the effective sample size (Harrell 2001, Burnham and Anderson 2002, Babyak 2004, Giudice et al. 2012). For our simple motivating example with 75 observations, this rule would suggest that we can afford to consider at most 7 candidate explanatory variables in a linear model with no interactions. Unfortunately, both model degrees of freedom and effective sample size can be difficult to quantify precisely. One generally thinks of model degrees of freedom as the number of explanatory variables considered. Certain modeling decisions, however, also should be considered as consuming degrees of freedom; among these are including non-linear terms or interactions, pre-screening potential explanatory variables with univariate statistics or scatterplots, deleting outliers, and transforming explanatory or response variables (Faraway 1992, Harrell 2001, Babyak 2004).

Effective sample size can be difficult to quantify because it depends not only on the number of observations but also on the type of data (Harrell 2001:60–61). For example, binary data tend to contain less information than continuous response data, with the effective sample size driven by the minimum number of 0s or 1s. We can discover next to nothing about the importance of explanatory variables if our response variable assumes all 0s (or all 1s). Determining effective sample size is even more daunting if observational units are sampled repeatedly. With such repeated measures, we have more information about factors that vary within observational units but not for factors that vary among observational units. A telemetry study resulting in 1,000 locations per year for each of 10 animals (8 male, 2 female) may provide useful insight into movement patterns of those 10 animals, but it will tell us little about sex-specific differences in habitat use. Information content is also influenced by the variation of the explanatory variables; in the grassland bird example, if all study sites are close to one another, little will be learned about effects of latitude and longitude.

In addition to problems associated with overfitting, regression models can also be unstable when some explanatory variables are highly correlated (multicollinear). What do we mean by unstable? If we collected the same type of data again and then determined the best model, we would likely end up with a very different model. The problem is that correlated variables will compete to explain the same variation in the response; minor differences in the data can easily result in a different set of variables being chosen. With experiments, randomization is used to ensure that treatment groups have, at least in expectation, balanced distributions of explanatory variables. Thus, there should be no relationship between the treatment indicator variable (1 if treated, 0 otherwise) and other explanatory variables. With observational data, explanatory variables nearly always exhibit some degree of correlation. Including multiple correlated variables in the same model rather than selecting among them can be problematic. A regression coefficient in a multivariable regression model describes the change in the response variable per unit change in an explanatory variable while holding all other explanatory variables constant. This partial effect will be difficult to interpret if variables are highly correlated; it is not possible to change 1 variable while holding other closely correlated variables constant. In turn, standard errors of the fitted regression coefficients will be inflated compared to models without collinear variables (Kutner et al. 2005). Variables that sum to a constant or nearly so are especially problematic. Compositional variables, such as the percentage cover variables, are a perfect example because values for the categories have to sum to 100%. When explanatory variables are multicollinear, it may be possible to infer only the importance of the entire group of them. In contrast, including multiple correlated explanatory variables in the same model may have little effect on prediction error, provided that the target population is similar to the sampled population (Harrell 2001, Graham 2003); predictions for spatially or temporally separate populations will likely be poorer because relationships among the explanatory variables will usually differ from that in the modeled population (Dorman et al. 2013).

Related to these challenges of overfitting and multicollinearity is the difficulty in quantifying uncertainty after having chosen a particular model. Assume that the investigators have somehow used their data to decide on the most-appropriate model; they then proceed to fit the model to the data, that is, to estimate the parameters in the selected model, which usually are regression coefficients associated with explanatory variables. Investigators typically then proceed to interpret confidence intervals and *P*-values as though this model had been pre-specified. This process exemplifies what Breiman (1992) labeled the "quiet scandal" in the statistical community: data-driven model selection leading to overly optimistic results and models unlikely to fare well when applied to new data. Statisticians have come to recognize that model selection needs to be acknowledged as a source of uncertainty in statistical inference. Chatfield (1995:419), for example, noted that "model uncertainty is a fact of life and likely to be more serious that other sources of uncertainty which have received far more attention from statisticians." Draper (1995:1) cautioned that accepting a particular model as correct and proceeding with estimation may lead to "inaccurate scientific summaries and overconfident decisions."

In frequentist modeling applications, we think of uncertainty in terms of the different possible realizations that could result from repeating the process of collecting data, analyzing the data, and predicting outcomes from fitted models. Traditional inference techniques ignore uncertainty associated with analyzing data, assuming the analysis approach was pre-specified. Significant progress has been made in the past 15–20 years with respect to quantifying the importance of model uncertainty in the case of a set of pre-specified models. Buckland et al. (1997), arguing that model uncertainty should be fully incorporated into statistical inference, suggested that predictions or estimates be based on the full set of models considered, weighted appropriately by the support each model receives from the data available. Buckland et al. (1997) further proposed that these model weights be based on information criteria, either Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC). This paradigm shift was reflected in the change of title of Burnham and Anderson's influential book, from *Model Selection and Inference* for the first (1998) edition to *Model Selection and Multimodel Inference* for the second (2002) edition. It is now common for authors of papers in *The Journal of Wildlife Management* to consider and summarize results from multiple models—these models are frequently compared using AIC. This transition also presents new challenges and opportunities when it comes to reporting effect sizes. Rather than report effect sizes from the best model, one may report effect sizes for multiple competitive models or, alternatively, some form of model-averaged effect size.

## HOW SHOULD THE INVESTIGATOR USE THE DATA TO INFORM MANAGEMENT?

Given the challenges noted above with overfitting and multicollinearity, how should the investigator proceed? We begin by considering 3 different approaches: 1) stepwise selection or all-subsets regression, 2) some form of multimodel inference using AIC (or another information criterion), and 3) a full pre-specified model that includes a limited number of explanatory variables selected without ever looking at the response variable. We then discuss how these methods, and potential alternatives, might fit in within Chamberlin's (1890) framework of entertaining multiple working hypotheses about causal mechanisms.

### All-Subsets or Stepwise Regression

Assume the investigators use all-subsets regression to identify the most important explanatory variables; that is, a computer program will perform regressions on all possible combinations of explanatory variables. This approach or stepwise regression will lead us to find the best-fitting model for the data in hand, as judged by whatever criterion was used to choose the model (e.g., $R^2$, adjusted $R^2$, AIC, BIC). In this best-fitting model, it is unlikely that 2 highly correlated variables will both appear. Thus, for example, the model probably will not include the same landscape variable measured at more than 1 spatial scale (400 m, 800 m, 1,200 m). Although multiple models are likely to fit the data equally well, someone reading a description of the investigators' work may not be aware of this fact. For each variable retained in the final model, it will be easy to estimate its effect size (the estimated regression coefficient) along with a measure of its uncertainty (its standard error). Yet, if the same model were fit to another data set (collected at the same or different sites), we would expect the regression coefficients to be different (in fact closer to 0 in absolute value) and the standard errors to be larger than measured by the fit to the original data (Altman and Andersen 1989, Harrell 2001:56–57, Foster and Stine 2006). The model almost certainly would explain much less of the variation in the new data. Further, if the model selection procedure were repeated with new data, different variables might be selected. Many analysts, including the authors, have been flabbergasted by how poorly a model that fit the original data well performed with different data.

There are several reasons that these and other problems arise with stepwise or all-subsets regression. For one, choosing among many different potential explanatory variables is analogous to conducting multiple hypothesis tests or making multiple comparisons from an analysis of variance model; the probability of making a Type I error increases with the number of tests performed or, in the present case, with the number of potential explanatory variables considered (Murtaugh 1998, Whittingham et al. 2006). These concerns hold true regardless of the selection criterion used (Murtaugh 2009, 2014). Thus, it is likely that some of the variables included in the final model represent spurious relationships. Second, variable selection results in biased parameter estimators (Whittingham et al. 2006). Why? Inclusion in the final model depends not on the true relationship between the explanatory and response variables but rather on the estimated relationship determined from the original sample. A variable is more likely to be included if by chance its importance in the original sample was overestimated than if it was underestimated (Copas and Long 1991). These and additional problems associated with all-subsets and stepwise selection are well known to statisticians, who agree that these methods should be abandoned (Burnham and Anderson 2002, Whittingham et al. 2006, Hegyi and Garamszegi 2011, Guidice et al. 2012).

### Multimodel Inference

Buckland et al. (1997) and Burnham and Anderson (2002) suggested developing a limited number of a priori hypotheses that could be translated into competing models and compared using AIC or BIC. Investigators might use these methods in the hypothetical example by considering a variety of candidate model sets, ranging from the set of all possible models with linear and additive effects, to a set with only a few models containing carefully chosen variables.

A range of possibilities exists for summarizing results from multiple models. Commonly, investigators report the explanatory variables contained in several best-fitting models (e.g., all models with $\Delta\text{AIC}_i < 5$, where $\Delta\text{AIC}_i$ is the difference in AIC values between model $i$ and the model with the smallest AIC), along with AIC-derived model weights,

$$w_i = \exp\left(-\frac{1}{2}\Delta\text{AIC}_i\right) / \sum_{j \in M} \exp\left(-\frac{1}{2}\Delta\text{AIC}_j\right).$$

One may also determine a, say, 95% confidence set of models by including models in order of largest to smallest $w_i$ until their summed weights equals 0.95. Lastly, one may report estimated effect sizes in a variety of manners, including 1) effect sizes for variables included in the best-fitting model, 2) effect sizes for variables in a set of best-fitting models (either those in a 95% confidence set or those with $\Delta$AIC values less than some cut-off value), or 3) model-averaged effect sizes. In the last case, one may model-average regression coefficients or model-average predictions from the models. Buckland et al. (1997) and Burnham and Anderson (2002) provided formulas for estimating model weights and calculating model-averaged regression parameters and predictions. When model averaging regression coefficients, one should include a zero coefficient for explanatory variables that are absent from a given model (Lukacs et al. 2010). With generalized linear models (e.g., logistic regression) or generalized linear mixed effects models, model-averaged predictions will differ from the predictions formed using model-averaged regression coefficients.

Because model weights are associated with models, not individual explanatory variables in the models, effect sizes are best described using model-averaged predictions rather than model-averaged coefficients, particularly for non-linear models (Cade 2015). A reasonable strategy for reporting a model-averaged effect size for a variable, $Z$, is to 1) fix all

other variables, $X$, to specific values (e.g., sample means for quantitative variables, most frequent category for nominal variables); 2) consider 20–30 values for $Z$ equally spaced between the minimum and the maximum $Z$ in the observed data; 3) calculate model-averaged predictions for the mean response at $X = x$ and $Z = z$ for each value of $z$; and 4) plot the predictions as a function of $z$. Multiple lines (e.g., using different values of $x$) may be overlaid on the same plot to explore predictions for different levels of a categorical variable (in $X$) or the importance of interactions. This type of effect plot is often used to summarize generalized linear models on the natural scale of the response variable (Fox 2003).

One should also incorporate a measure of uncertainty with the associated effect size. Burnham and Anderson (2002, 2004) provided 2 different formulas for confidence intervals; both rely on a normality assumption for the model-averaged estimator. Several other methods for producing model-averaged confidence intervals have been proposed (Hjort and Claeskens 2003, Turek and Fletcher 2012, Yu et al. 2014, Jensen and Ritz 2015) but have yet to be adopted in the wildlife literature. The performance of different interval estimators likely will depend on the characteristics of the data and the models, and so far, no clear winner has emerged (e.g., Turek and Fletcher 2012, Kabaila et al. 2014, Yu et al. 2014). Simulation and analytical results suggest that some model-averaged interval estimators have similar width and coverage as those based on fitting a full model (Claeskens and Carroll 2007, Wang and Zhou 2013). Thus, for interval estimation, model averaging may not provide much of an advantage relative to fitting a full model (Wang and Zhou 2013). Nonetheless, we expect interval estimation for model-averaged estimators to remain an active area of statistical research in the coming years.

Assume the investigators decide to look at all possible models with linear and additive effects and report all models within 5 AIC units of the top model, along with their associated model weights, $w_i$. This approach has advantages over stepwise modeling approaches, which tend to emphasize a single best-fitting model. For one, the investigators likely would discover that multiple models fit the data about equally well. Thus, they may be able to see clearer the effects of multicollinearity (different sets of variables explaining variation in the response variable equally well) than if they had selected a single best-fitting model.

One concern with looking at many possible models is the possibility that the models that rise to the top will be too finely tuned to the data (i.e., they overfit the data). For this reason, Burnham and Anderson (2002), among others, have argued against looking at all possible models: "An investigator with, say, 10 explanatory variables cannot expect to learn much from his data and a multiple linear regression analysis unless there is some substantial supporting science that can be used to help narrow the number of models to consider. In this example, there would be 1,024 models (many more if transformations or interaction terms were allowed), and over-fitting would surely be a serious risk. The analysis, by whatever method, should probably be considered exploratory and the results used to design further data gathering leading to a more confirmatory analysis, based on some a priori considerations" (Burnham and Anderson 2002:85).

The investigators could start with a more limited set of models constructed from the set of explanatory variables. What might be gained from multimodel inference in this case, relative to using a single, full model for inference? By considering more parsimonious models, and averaging predictions across models, it is likely that the investigator will be able to make better predictions. In particular, model averaging has been shown in many cases to decrease prediction error, relative to any single model (e.g., Hoeting et al. 1999, Hjort and Claeskens 2003, Raftery and Zheng 2003). Yet, the investigator would need to continue measuring all of the explanatory variables included in the analysis. Consideration of multiple models also adds complexity to both the analysis and reporting of effect sizes and results. An additional danger is that investigators may focus too much on describing model uncertainty and too little on effect sizes and their uncertainty. Also, investigators often rely on summed model weights associated with each variable to determine variable importance. Yet, this simple summary is likely to be much less informative than estimates of regression coefficients and their standard errors (Murray and Connor 2009, Galipaud et al. 2014, Cade 2015).

### Using a Single Model for Inference

One way to avoid some of the problems associated with model selection is to "just say no" to data-driven model selection—in other words, fit a single pre-specified model, one in which all explanatory variables are chosen prior to looking at their relationship with the response variable, and then use this model for inference (Harrell 2001, Babyak 2004, Whittingham et al. 2006, Guidice et al. 2012). With a pre-specified model, $P$-values and confidence intervals have their intended interpretation, as long as the assumptions of the model (e.g., normality of error terms, constant error variance in linear regression) are approximately met. Although the problems associated with multiple testing are still pertinent when making inference with respect to several explanatory variables, the issue is much less severe than when using stepwise selection, assuming the investigators successfully whittled the pool of explanatory variables down to a much smaller set prior to model fitting. Given the small effective sample sizes associated with many wildlife studies, some variable reduction will typically be necessary to avoid problems associated with overfitting—that is, the level of model complexity should be in line with the information content of the data. Conversely, studies should be designed to collect enough data to allow consideration of all variables of interest. Thus, the usual first step in the process of fitting a single model will be thinking hard about which explanatory variables to consider. Importantly, explanatory variables should be chosen without examining the strength of their relationship with the response variable.

Several guidelines exist for helping decide which explanatory variables to include. To begin, one might consider the extent of missing data for each variable, its range, and its relationship to other potential explanatory variables (Harrell 2001, Guidice et al. 2012). Dorman et al. (2013) suggested

selecting variables that are 1) ecologically relevant, 2) feasible to measure, and 3) related to the causal mechanism. For management-focused research, choosing variables that actually can be manipulated is desirable.

Analyses of observational studies frequently include categorical explanatory variables such as year or study area. Inclusion of variables such as these helps to increase precision of estimates of treatment effects in designed experiments. This strategy may also be reasonable in observational studies when the goal is to estimate the effect of a management action, provided that the action is not confounded with year or study area. Prediction, however, requires that we replace year and study area effects, to the extent possible, with other variables that actually cause the effects that were observed. An added benefit to this approach is that it may lead to better understanding (i.e., we may gain insights into why year or study area has an effect). Year effects might involve weather conditions or phenology, for example. Study areas may vary because of location relative to the breeding range of a particular species, soil type, general habitat conditions, and the like. Hierarchical models are well suited for incorporating higher-level effects into variables such as year and study area (e.g., Gelman and Hill 2007).

For our hypothetical example, we might begin by choosing, for each explanatory variable, what we consider to be the most appropriate spatial scale. We might decide to include percentage cover associated with grasslands and trees (but not both wetland and crop cover because these 4 variables sum to 100%). We might add variables that capture year-to-year or spatial variability (e.g., mean spring temperature or precipitation levels, latitude, or longitude). We might then explore pairwise scatterplots and correlations among the Robel measurements (mean, coefficient of variation), the percentage cover variables already chosen, and the remaining within-field measurements (e.g., percent cover of forbs, litter, bare ground), selecting an additional 2 or 3 variables using the considerations outlined above (e.g., extent of missingness, range of each variable, feasibility of data collection, degree of collinearity with already chosen variables, relevance to management). In making our decisions, we should record the variables that were eliminated from consideration and why. In particular, if variables are eliminated because they are highly correlated with other variables, this should be stated so the reader understands that a different suite of variables might capture similar patterns in the response data. Alternatively, one might choose to combine similar variables using an index or a principal components analysis (Harrell 2001).

Most relationships in the real world are nonlinear. Taking 20 aspirins for a headache will not cure it 10 times as fast or as well as would 2 aspirins. Most models of relationships, in contrast, are linear in the explanatory variables (or in the link function for generalized linear models). This approach usually works because well-behaved relationships are not too different from linear for a reasonably small range of values of the explanatory variable(s). This approximation may fail, however, when applied to values beyond the range of values used to estimate parameters in the model. Accordingly, one might also want to consider nonlinear effects for the most important variables. Interactions with a strong theoretical basis (i.e., prior to data collection) may also be worthy of consideration.

One risk of using a single model is that it may be too simplistic, that is, important explanatory variables are not included. Then, estimated regression coefficients again will be biased, but for a different reason. Estimated coefficients will account for not only the relationship between the response variable and explanatory variables included in the model but also the influence of omitted variables that happen to be correlated with explanatory variables. To avoid being misled, one may fit more complex models than the data can afford to support, recognizing that the model may not predict new data well and that measures of fit (e.g., $R^2$) are likely to be overly optimistic. In many cases, the degree of optimism can be measured with cross-validation procedures or bootstrapping (e.g., Harrell 2001, Guidice et al. 2012; Fig. 1A). The bootstrap steps, as outlined in Harrell (2001) and Guidice et al (2012), are to 1) fit a model to the original data set and calculate a measure of model fit or performance; we use $R^2$; 2) form bootstrap training and bootstrap test data sets, each by resampling the original data set with replacement; 3) fit the model to the bootstrap training data set; 4) calculate the percent of the variation in the training data explained by the model from step 3, $R^2_{train}$; 5) calculate the percent of the variation in the test data explained by the model from step 3, $R^2_{test}$. The difference in $R^2$ values, $R^2_{train} - R^2_{test}$, can be used to estimate the degree of optimism arising from using the same data set for both fitting and testing the model. Steps 2–5 are repeated many times and then the average optimism can be subtracted from the $R^2$ from step 1.

If one were to plot new response data against predicted values from a previously overfit model, one would typically find that low predictions are too low and high predictions are too high (i.e., there is a "regression towards the mean" effect; Copas 1997). A similar bootstrapping process can be used to calibrate the model so it predicts new data well (Harrell 2001:94–97; Fig. 1B): 1) estimate regression parameters using the full data set, yielding $\widehat{\beta}_{full}$; 2) form bootstrap training ($x_{train}$, $y_{train}$) and bootstrap test ($x_{test}$, $y_{test}$) data sets, each by resampling the original data set with replacement; 3) fit the model to the bootstrap training data, yielding $\widehat{\beta}_{train}$; 4) use the model from step 3 to form predicted responses for the bootstrap test data, $\widehat{y}_{test} = x_{test}\widehat{\beta}_{train}$; 5) fit a linear regression model using the test data responses (as the response variable) and the predicted values from step 4 as the only explanatory variable, $y_{test} = \alpha_0 + \alpha_1\widehat{y}_{test} + \varepsilon$; repeat steps 2–5 many times. The average intercept and slope from step 5, $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$ respectively, can be used to re-calibrate the model: $\widehat{y}_{calibrated} = \widehat{\alpha}_0 + \widehat{\alpha}_1 x\widehat{\beta}_{full}$. Typically, $\widehat{\alpha}_1 < 1$ if the data have been overfit, in which case, the model is calibrated by shrinking regression parameter estimates towards 0.

A variety of other shrinkage methods (e.g., penalized likelihood, ridge regression, LASSO) have been developed to improve model performance (e.g., reviews by van Houwelingen 2001, Hooten and Hobbs 2015). Shrinkage

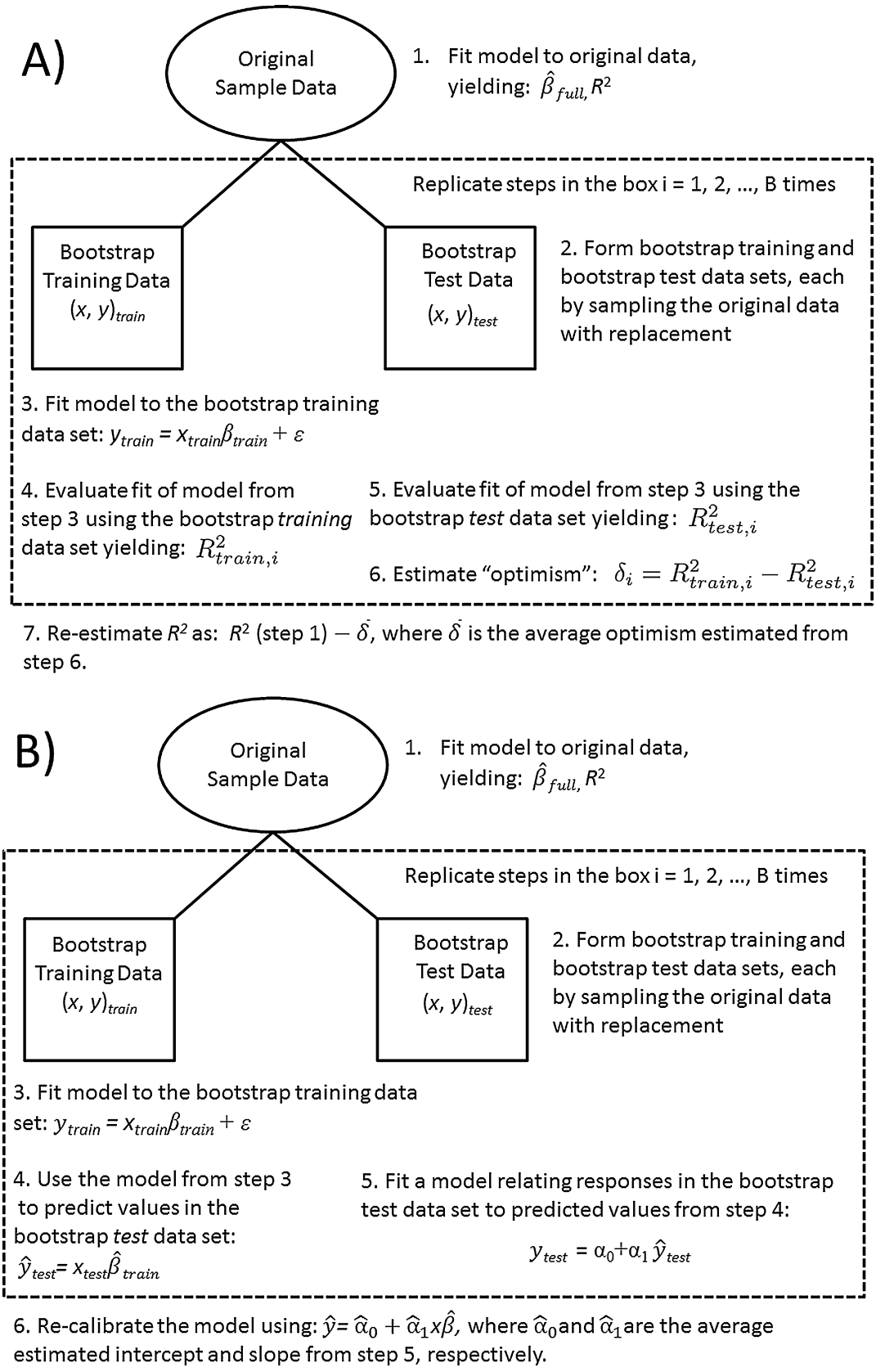


**Figure 1.** Bootstrapping can be used to evaluate how well a model will perform when applied to a new data set. Bootstrap training and bootstrap test data are formed, each by resampling the original data with replacement. The model is then fit to the bootstrap training data. Differences in model fit and predictive ability, measured using the bootstrap training and bootstrap test data, are then used to calculate an adjusted $R^2$ (panel A) or to re-calibrate the model so that it predicts new data well (panel B).

methods reduce model complexity, and hence the effective model degrees of freedom, by decreasing regression coefficients in absolute value. This process also stabilizes regression parameter estimators in the presence of multicollinearity (Hooten and Hobbs 2015). Thus, a single model can be used for inference, without the need to drop 1 or more collinear variables (e.g., based on inspection of pairwise correlations or variance inflation factors) prior to model fitting. Although a thorough review of alternative shrinkage methods is outside the scope of this paper, we suggest these approaches deserve more attention (see also Dahlgren 2010). In particular, shrinkage methods provide an attractive alternative to AIC-based multimodel inference in the case of variable selection, particularly when explanatory variables are correlated (Hooten and Hobbs in 2015).

Another downside to the single-model approach is that one is again forced to continue measuring all of the variables included in that model, even if a more parsimonious model may suffice. This is a cost of avoiding model selection. Another potential concern is that one might fail to discover other important relationships in the data. If we looked at explanatory variables not included in the single model, we might find that we can improve model fit. This is a reason to conduct further exploratory analyses after making inferences from the single model. Results of these exploratory analyses should be carefully distinguished from those that were specified a priori (Burnham and Anderson 2002) and treated with more skepticism (Babyak 2004). If we find a posteriori that a different explanatory variable appears to be more appropriate, then we can develop models with this variable and fit them to data collected in subsequent studies. Such data-informed choices may be improvements over our original choice, but typically improvements are more modest than expected based on the original exploratory analyses.

## GRAPHICAL MODELS FOR PORTRAYING CAUSAL RELATIONSHIPS

In many instances, we need to acknowledge the exploratory nature of much of our work and recognize the limitations associated with using regression models to analyze observational data. Such models are suited for describing associations among a set of variables, associations that may also depend on how the data were collected. Understanding cause and effect is much more challenging. As Box (1966:629) stated, "To find out what happens to a system when you interfere with it you have to interfere with it (not just passively observe it)." When we interfere with (manipulate, or manage) a system by changing 1 variable, we may influence the response through both direct effects and indirect effects; the latter may be mediated by other variables along the causal path that connects the intervention and the response variable. For example, burning a field may result in reduced litter depth and reduced abundance of standing grass and forbs, both of which may in turn lead to decreased numbers of grassland birds. Burns might also reduce woody vegetation, which would have the opposite effect, increasing the number of such birds. The net effect of burning a field thus may be positive or negative, depending on the strength of these individual relationships.

Recent research has been aimed at determining causal effects from observational data (Pearl 2000, Shipley 2002, Pugesek and Tomer 2003). One theme involves graphical modeling, including path analysis and structural equation modeling, which require that one posit a graph connecting variables in a system to each other, with directional arrows indicating cause and effect. If we are willing to make assumptions about causal connections among a set of variables, based on our current scientific understanding of the study system, we can then determine the logical implications of these assumptions. Specifically, we can ask, "what would happen to $Y$ if we manipulate $X$?" We can actually estimate causal effects, not just correlations, under the assumption that our graphical model represents the salient characteristics of our study system.

The caveat in the last sentence highlights a potential downside to graphical models: they are dependent on assumptions, only some of which may be testable. Goodness-of-fit tests are commonly applied to graphical models. If we find that our graphical model could have plausibly generated our data, can we then say that we have established causality? No, graphical models do not magically turn an observational study into an experimental one. Beyond the usual problems of accepting the null hypothesis, different models are often capable of producing the same set of statistical relationships among variables. "Researchers should keep in mind therefore that only a tiny portion of the assumptions behind each SEM [structural equation model] study lends itself to scrutiny by the data; the bulk of it must remain untestable, at the mercy of scientific judgment" (Pearl 2012:72).

Although a rich and elegant theory is available for understanding graphical models (Pearl 2000), simulation offers a simple tool for model exploration (e.g., Kaplan 2009). To illustrate, consider an example (Fig. 2A) involving assumed causal links among a treatment applied to a grassland (burn or not), mediating habitat variables, and grassland bird numbers as a response variable. By simulating data consistent with the model, we see that the regression coefficient associated with the treatment indicator variable, burn (=1 if burned, 0 if control), on average, will equal 0 if all mediating variables (woody vegetation, grass and forbs, and litter depth) are included in the model (Fig. S3A in Supplementary Appendix A, available online at www.onlinelibrary.wiley.com). As a corollary, we should not include any of the mediating variables in a model if we want to determine the net effect of burning on grassland bird abundance (Fig. S4 in Supplementary Appendix A). Further, when we omit important explanatory variables, the regression parameter estimators for those covariates that remain in the model will be biased. For example, if we fit a model with only woody vegetation, the coefficient for this variable will capture the combined effects of woody vegetation, grass and forbs, and litter depth because all 3 of these variables are correlated (because of their responses to burn intensity). As a result, the estimator for the regression coefficient associated with woody vegetation will not represent the effect solely of

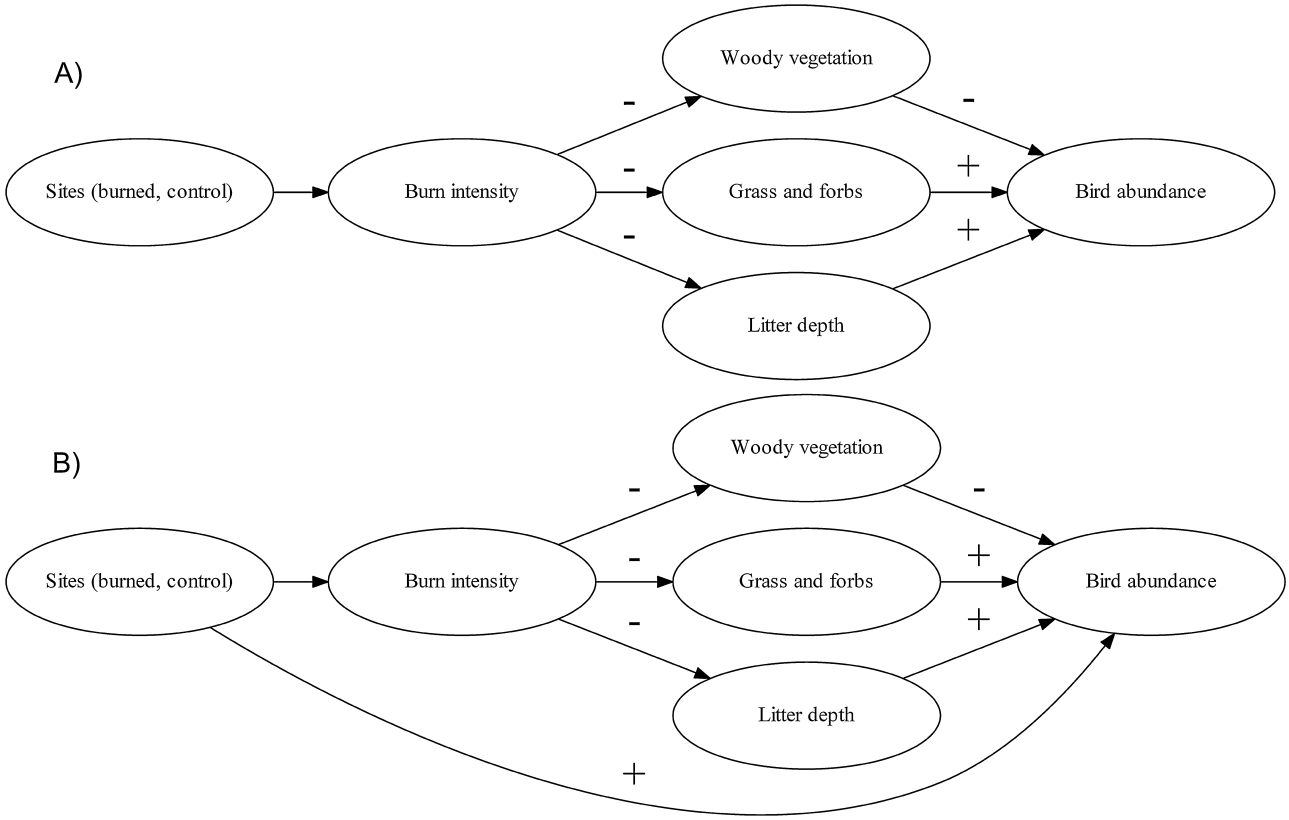


**Figure 2.** Graphical models relating a treatment (burn = yes or no), intensity of the burn (if it occurs), measures of standing grass and forb abundance, litter depth, and woody vegetation, and bird abundance. Arrows connecting variables represent causal effects, with the type of association indicated by a + (positive association) or – (negative association). In panel B, the line connecting sites (burned, control) to abundance represents pre-treatment differences between burn and control sites (burns having higher pre-treatment abundances, on average).

woody vegetation. In general, the direction and extent of bias will depend on the correlations between the included and omitted variables as well as the relationships between the omitted variables and the response variable (Schildcrout et al. 2011). We may even find, as we occasionally did in the simulated data example, that the coefficient is positive despite the negative association between woody vegetation and bird abundance (Fig. S5 in Supplementary Appendix A).

The interpretation of regression parameters will change if we change the structure of our graphical model. For example, if we thought that burn sites and control sites differ in their initial abundances (prior to the treatment), we could include a direct link between burn (yes or no) and bird abundance, in addition to the indirect links already present (Fig. 2B). Simulating data consistent with the model (Fig. 2B), we see that the regression parameter estimator for burn will measure this initial difference in bird abundance, provided we also include woody vegetation, grass and forbs, and litter depth in the model (Fig. S7 in Supplementary Appendix A). In summary, a graph encodes a set of statistical dependencies among both explanatory and response variables, and these dependencies determine the most appropriate model(s) for addressing a particular research question (Hernán 2002, 2011). When fitting multiple models, it is not uncommon to find that coefficients have opposite signs in different models. Rather than averaging over these different models, it may be more fruitful to explore the mechanisms responsible for any unanticipated changes in model coefficients by considering alternative, plausible graphical models (e.g., Fieberg and Ditmer 2012).

## CONCLUSIONS

There is no unique best way to analyze most data sets. Each of the analysis approaches we outlined has its advantages and disadvantages. Different analysts may well settle upon different methods. Although taking multiple approaches sounds like it could be confusing, it actually is a good thing, even for a single analyst, to do. If the major results from different analyses are similar, that robustness will induce greater confidence in those results. Substantial differences in results suggest that the assumptions underlying the analyses are speaking louder than the data themselves, and their veracity should be investigated. In most situations, fortunately, we expect different methods, properly conducted, to identify the same big chunks, those explanatory variables that have the strongest relationships with the response variable. Less importantly, the models likely will differ in terms of the rest of the story.

Identifying important explanatory variables from observational data is challenging, particularly when the number of potential explanatory variables is large and the sample size is small. When analyzing small- to moderate-sized

observational data sets, it is not surprising to find that regression models are unstable, with different sets of variables showing up as most important in different years or in different places, or with multiple models fitting the data equally well. For prediction, we may increase precision by averaging among competing models. We should, however, be cautious of conclusions drawn from individual studies, keep an open mind to the possibility of different explanations for observed phenomenon, and begin to accept a theory as useful only if results are replicated under a range of conditions (Johnson 2002).

What else can be done to advance wildlife science? Opportunities to conduct more experimental research are scarce but should be sought out (Eberhardt 2003). When experimentation is not possible, graphical models may offer an attractive alternative to traditional regression approaches because of their ability to represent direct and indirect causal effects in observational data (Pearl 2000, Shipley 2002). In addition, we can attempt to ask more focused research questions—rather than try to identify the most important factors associated with grassland bird abundance, for example, we might try to focus on the effects of 1 or 2 important factors that managers can control. When possible, we can select study sites to maximize variability associated with these factors and minimize variation in variables not of interest to us (nuisance variables) that may serve as potential confounders. With focused research questions, we can explore multiple models, not to find the optimal set of explanatory variables, but rather as a form of sensitivity analysis to determine if the estimated effect size associated with the variable of interest is robust to the inclusion of other variables (Zicus et al. 2003 offers an example).

Multimodel inference has sometimes been conflated with Chamberlin's (Chamberlin 1890) concept of multiple working hypotheses but, as practiced, it differs substantially from his notion. Chamberlin argued for considering a full suite of tenable mechanisms for a phenomenon, rather than deciding early on a favorite. Multimodel inference typically involves considering a single, possibly large, set of potential explanatory variables and how they relate to a response variable without thinking about causal mechanisms or how the different explanatory variables themselves may be related. Because graphical models can clearly represent different working hypotheses regarding how an ecological system functions, they can facilitate Chamberlin's approach and the hard thinking advocated by Burnham and Anderson (2002) when developing a priori models. Further, it is exactly this type of framework that lets us ask questions about what might happen if we intervene in the system, acknowledging that manipulation of 1 variable may result in both direct and indirect effects on the intended response variable (Pearl 2000). These are the questions wildlife managers care about.

## ACKNOWLEDGMENTS

We are grateful to P. Dixon, S. Lele, G. Sargeant, and an anonymous referee for comments on an early draft of this paper.

## 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.

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.

Box, G. E. P. 1966. Use and abuse of regression. Technometrics 8:625–629.

Breiman, L. 1992. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. Journal of the American Statistical Association 87:738–754.

Buckland, S. T., K. P. Burnham, and N. H. Augustin. 1997. Model selection: an integral part of inference. Biometrics 53:603–618.

Burnham, K. P., and D. R. Anderson. 1998. Model selection and inference: a practical information-theoretic approach. First edition. Springer-Verlag, 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.

Burnham, K. P., and D. R. Anderson. 2004. Multimodel inference understanding AIC and BIC in model selection. Sociological Methods and Research 33:261–304.

Cade, B. S. 2015. Model averaging and muddled multimodel inferences. Ecology 96: in press.

Chamberlin, T. C. 1890. The method of multiple working hypotheses. Science 15:92–96, Reprinted 1965, Science 148:754–759.

Chatfield, C. 1995. Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society, Series A 158:419–466.v

Claeskens, G., and R. J. Carroll. 2007. An asymptotic theory for model selection inference in general semiparametric problems. Biometrika 94:249–265.

Commoner, B. 1971. The closing circle. Alfred Knopf, New York, New York, USA.

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. Statistician 40:51–59.

Dahlgren, J. P. 2010. Alternative regression methods are not considered in Murtaugh (2009) or by ecologists in general. Ecology Letters 13.5: E7– E9.

Dormann, C. F., J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. G. Marquéz, B. Gruber, B. Lafourcade, P. J. Leitao, T. Munkemuller, C. McClean, P. E. Osborne, B. Reineking, B. Schroder, A. K. Skidmore, D. Zurell, and S. Lautenbach. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36:027–046.

Draper, D. 1995. Assessment and propagation of model uncertainty (with discussion). Journal of the Royal Statistical Society, Series B 57:45–97.

Eberhardt, L. L. 2003. What should we do about hypothesis testing? Journal of Wildlife Management 67:241–247.

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

Fieberg, J., and M. Ditmer. 2012. Understanding the causes and consequences of animal movement: a cautionary note on fitting and interpreting regression models with time-dependent covariates. Methods in Ecology and Evolution 3:983–991.

Foster, D. P., and R. A. Stine. 2006. Honest confidence intervals for the error variance in stepwise regression. Journal of Economic and Social Measurement 31:89–102.

Fox, J. 2003. Effect displays in R for generalized linear models. Journal of Statistical Software 8:1–27.

Galipaud, M., M. A. Gillingham, M. David, and F. X. Dechaume-Moncharmont. 2014. Ecologists overestimate the importance of predictor variables in model averaging: a plea for cautious interpretations. Methods in Ecology and Evolution 5:983–991.

Gelman, A., and J. Hill. 2007. Data analysis using regression and multilevel/ hierarchical models. Cambridge University Press, New York, New York, USA.

Giudice, J., J. Fieberg, and M. Lenarz. 2012. Spending degrees of freedom in a poor economy: a case study of building a sightability model for moose in northeastern Minnesota. Journal of Wildlife Management 76:75–87.

Graham, M. H. 2003. Confronting multicollinearity in ecological multiple regression. Ecology 84:2809–2815.
Harrell, F. E., Jr. 2001. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer-Verlag, New York, New York, USA.
Hegyi, G., and L. Z. Garamszegi. 2011. Using information theory as a substitute for stepwise regression in ecology and behavior. Behavioral Ecology and Sociobiology 65:69–76.
Hernán, M. A., D. Clayton, and N. Keiding. 2011. The Simpson's paradox unraveled. International Journal of Epidemiology 40:780–785.
Hernán, M. A., S. Hernández-Díaz, M. M. Werler, and A. A. Mitchell. 2002. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. American Journal of Epidemiology 155:176–184.
Hjort, N. L., and G. Claeskens. 2003. Frequentist model average estimators. Journal of the American Statistical Association 98:879–899.
Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. Bayesian model averaging: a tutorial. Statistical Science 14:382–401.
Hooten, M. B., and N. T. Hobbs. 2015. A guide to Bayesian model selection for ecologists. Ecological Monographs 85:3–28.
Jensen, S. M., and C. Ritz. 2015. Simultaneous inference for model averaging of derived parameters. Risk Analysis 35:68–76.
Johnson, D. H. 2002. The importance of replication in wildlife research. Journal of Wildlife Management 66:919–932.
Kabaila, P., A. H. Welsh, and W. Abeysekera. 2014. Fletcher-Turek model averaged profile likelihood confidence intervals. arXiv preprint arXiv:1404.6855.
Kaplan, D. T. 2009. Statistical modeling: a fresh approach. CreateSpace, Seattle, Washington, USA.
Kutner, M. H., C. J. Nachtsheim, J. Netter, and W. Li. 2005. Applied linear models, 5th edition. McGraw-Hill, New York, New York, USA.
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.
Mundry, R., and C. L. Nunn. 2009. Stepwise model fitting and statistical inference: turning noise into signal pollution. American Naturalist 173:119–123.
Murray, K., and M. M. Conner. 2009. Methods to quantify variable importance: implications for the analysis of noisy ecological data. Ecology 90:348–355.
Murtaugh, P. A. 1998. Methods of variable selection in regression modeling. Communications in Statistics - Simulation and Computation 27:711–734.
Murtaugh, P. A. 2009. Performance of several variable-selection methods applied to real ecological data. Ecology Letters 12:1061–1068.
Murtaugh, P. A. 2014. In defense of *P* values. Ecology 95:611–617.
Pearl, J. 2000. Causality: models, reasoning, and inference. Cambridge University Press, New York, New York, USA.
Pearl, J. 2012. The causal foundations of structural equation modeling. Pages 68–91 *in* R. H. Hoyle, editor. Handbook of structural equation modeling. Guilford Press, New York, New York, USA.
Pugesek, B. H., and A. Tomer. 2003. Structural equation modeling: applications in ecological and evolutionary biology. Cambridge University Press, New York, New York, USA.
Raftery, A. E., and Y. Zheng. 2003. Discussion: performance of Bayesian model averaging. Journal of the American Statistical Association 98: 931–938.
Robel, R. J., J. N. Briggs, A. D. Dayton, and L. C. Hulbert. 1970. Relationships between visual obstruction measurements and weight of grassland vegetation. Journal of Range Management 23:295–297.
Schildcrout, J. S., S. Haneuse, J. F. Peterson, J. C. Denny, M. E. Matheny, L. R. Waitman, and R. A. Miller. 2011. Analyses of longitudinal, hospital clinical laboratory data with application to blood glucose concentrations. Statistics in Medicine 30:3208–3220.
Shipley, B. 2002. Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference. Cambridge University Press, Cambridge, United Kingdom.
Turek, D., and D. Fletcher. 2012. Model-averaged Wald confidence intervals. Computational Statistics and Data Analysis 56:2809–2815.
van Houwelingen, J. C. 2001. Shrinkage and penalized likelihood as methods to improve predictive accuracy. Statistica Neerlandica 55:17–34.
Wang, H., and S. Z. Zhou. 2013. Interval estimation by frequentist model averaging. Communications in Statistics-Theory and Methods 42: 4342–4356.
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.
Yu, W., W. Xu, and L. Zhu. 2014. Transformation-based model averaged tail area inference. Computational Statistics 29:1713–1726.
Zicus, M. C., J. Fieberg, and D. P. Rave. 2003. Does mallard clutch size vary with landscape composition: a different view. Wilson Bulletin 114: 409–413.

*Associate Editor: Evelyn Merrill.*

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site.