2-1-2009

Methods to quantify variable importance: implications for theanalysis of noisy ecological data

Kim Murray
University of Montana

Mary M. Conner
Utah State University

Follow this and additional works at: https://digitalcommons.usu.edu/wild_facpub

Part of the Environmental Sciences Commons

Recommended Citation

UtahState University
MERRILL-CAZIER LIBRARY

Methods to quantify variable importance:
implications for the analysis of noisy ecological data

Kim Murray[1,2,3,4] and Mary M. Conner[3]

[1]*Snow Leopard Trust, Seattle, Washington 98103 USA*
[2]*Wildlife Biology Program, Division of Biological Sciences, University of Montana, Missoula, Montana 59812 USA*
[3]*Department of Wildland Resources, Utah State University, Logan, Utah 84322-5230 USA*

Abstract. Determining the importance of independent variables is of practical relevance to ecologists and managers concerned with allocating limited resources to the management of natural systems. Although techniques that identify explanatory variables having the largest influence on the response variable are needed to design management actions effectively, the use of various indices to evaluate variable importance is poorly understood. Using Monte Carlo simulations, we compared six different indices commonly used to evaluate variable importance; zero-order correlations, partial correlations, semipartial correlations, standardized regression coefficients, Akaike weights, and independent effects. We simulated four scenarios to evaluate the indices under progressively more complex circumstances that included correlation between explanatory variables, as well as a spurious variable that was correlated with other explanatory variables, but not with the dependent variable. No index performed perfectly under all circumstances, but partial correlations and Akaike weights performed poorly in all cases. Zero-order correlations was the only measure that detected the presence of a spurious variable, whereas only independent effects assigned overlap areas correctly once the spurious variable was removed. We therefore recommend using zero-order correlations to eliminate predictor variables with correlations near zero, followed by the use of independent effects to assign overlap areas and rank variable importance.

Key words: Akaike weights; beta coefficients; beta weights; dominance analysis; hierarchical partitioning; independent effects; partial correlation coefficients; relative weights; standardized regression coefficients.

Introduction

Ecologists, conservation biologists, and wildlife managers are often faced with the challenge of identifying factors associated with an ecological phenomenon of interest. For instance, Ottichilo et al. (2001) explored ecological and anthropogenic factors influencing population trends of resident wildebeest (*Connochaetes taurinus hecki*) in the Masai Mara ecosystem. Farmer et al. (2006) investigated habitat factors influencing the risk of mortality in black-tailed deer (*Odocoileus hemionus sitkensis*) in a managed forest landscape. And, Stokes and Cunningham (2006) used model selection techniques to determine which river flow regulation scheme and spatial biological factors would likely have the greatest impact on the control of invasive willows in riparian areas. In situations such as these, controlled experiments provide the optimal means to disentangle the contributions of various factors of interest and elucidate relationships among response and explanatory (independent, predictor) variables (Williams 1997). However, in many cases large-scale manipulations are not feasible for financial, logistical, or even ethical

reasons, necessitating the use of multivariate analysis methods to identify a "best" model, or suite of models, to be used for making valid inference. Recent developments in the use of information theory and multi-model inference have helped to improve the process of making valid inference by reducing uncertainty associated with model-selection (Burnham and Anderson 2002). However, these approaches focus on comparisons among models and not on the relative importance of the explanatory variables contained within the models.

Once the best model has been identified, ecologists and wildlife managers often desire to know which of the various independent variables included in the model has the strongest influence on the response variable, and should therefore be targeted, to achieve a desired management outcome. Explanatory variables are often only nominally independent, and correlation among the explanatory variables (i.e., multicollinearity) makes it difficult to tease apart the unique contributions of each explanatory variable to the response variable (Mac Nally 2000, Graham 2003). For example, given a response variable (y) and two explanatory variables (x_1, x_2; Fig. 1a), area A represents variance in y uniquely predicted by variable x_1, B represents variance in y uniquely predicted by variable x_2, C represents variance in y redundantly predicted by variables x_1 and x_2, D represents variance in y predicted by neither variable,

and E represents covariance between variables x_1 and x_2 that is not shared with the dependent variable. Considering the correlation between x_1 and x_2, how should overlap area C be assigned to determine the relative importance of each explanatory variable in terms of its contribution to the prediction of the dependent variable?

Although determining the relative importance of explanatory variables is of practical relevance to the allocation of limited resources to the management of natural systems, the use of various indices to evaluate variable importance is not well understood. In this paper, we focus on the process of variable selection, which can be used in conjunction with model selection to determine which of the explanatory variables contained in the top-ranked model has the strongest influence on the dependent variable. The results of variable selection can then be treated as multiple competing hypotheses for subsequent hypothesis testing within an adaptive management framework (Nichols 2001).

We begin by briefly reviewing four traditional measures (zero-order correlations, partial correlations, semipartial correlations, and standardized regression coefficients) and two relatively newer methods (Akaike weights and independent effects) used to assess variable importance in the ecological literature. Next, we report the results of Monte Carlo simulations comparing the performance of these indices in terms of their ability to (1) correctly weight the explanatory variables based on the proportion of the total variance in the dependent variable independently explained by each variable, and (2) detect the presence of a spurious variable that is correlated with other explanatory variables in the model, but not with the dependent variable. We conclude with a discussion of the results and recommendations for use.

A REVIEW OF INDICES OF VARIABLE IMPORTANCE

Zero-order correlations

Given a single dependent variable (y) and two explanatory variables (x_1, x_2), the squared zero-order (simple, bivariate) correlation between y and x_1, defined as

$$r^2_{yx_1} = \left(\frac{\sum yx_1}{\sqrt{\sum y^2 \sum x_1^2}} \right)^2$$

measures the direct effect of explanatory variable x_1 on dependent variable y, while ignoring the effect of variable x_2 (Fig. 1b, area A + C/area A + B + C + D [Cohen et al. 2003]). Because zero-order correlations measure only the direct effect of each predictor, they are unable to partition the variance shared by two or more correlated predictors into the variance attributable to each predictor. Thus, when the explanatory variables are uncorrelated, the interpretation of variable importance is straightforward. The squared zero-order correlations for all variables in the model sum to the multiple
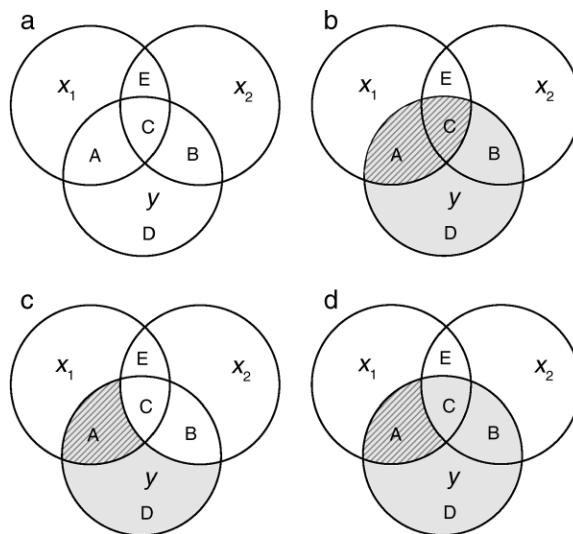

FIG. 1. Variance partitioning for (a) the general case of two correlated explanatory variables and a response variable, and for the specific cases of (b) zero-order correlations, (c) partial correlations, and (d) semipartial correlations. See *Introduction* for an explanation of variables and regions.

correlation coefficient (r^2, coefficient of determination), and relative importance is assessed by a rank ordering of the values of the observed squared correlations. However, when correlation between the explanatory variables exists, interpretation of variable importance is unclear. The squared zero-order correlations for all variables in the model no longer sum to the model r-squared, and the individual correlations will not accurately reflect the true contributions of the correlated variables.

Partial correlations

Given a single dependent variable (y) and two explanatory variables (x_1, x_2), the squared partial correlation (coefficient of partial determination) between y and x_1, defined as

$$r^2_{yx_1.x_2} = \left(\frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{1 - r^2_{yx_2}} \sqrt{1 - r^2_{x_1x_2}}} \right)^2$$

represents the correlation between y and x_1 after the influence of x_2 has been removed from both y and x_1 (Fig. 1c, area A/area A + D [Cohen et al. 2003]). Unlike zero-order correlations that ignore the effects of other explanatory variables, partial correlations measure the predictive efficacy of an explanatory variable in the presence of a specific subset of the remaining regressors. Importance is assessed via a rank ordering of the observed partial correlations, contingent upon the specific subset of explanatory variables included in the model. As with zero-order correlations, partial correlations are not designed to partition the variance shared between multiple correlated predictors and the depen-

Reports

dent variable. Thus, interpretation of variable importance is unclear when correlation between the explanatory variables exists. Additionally, because the overlap area is excluded from the partial correlations for both variables, the overall model may be significant (i.e., $F < 0.05$) but no single explanatory variable may account for a significant proportion of the variance in the dependent variable (Berry and Feldman 1985).

Semipartial correlations

Given a single dependent variable (y) and two explanatory variables (x_1, x_2), the squared semipartial correlation between y and x_1, defined as

$$r^2_{y(x_1 . x_2)} = \left(\frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{\sqrt{1 - r^2_{x_1 x_2}}} \right)^2$$

represents the correlation between y and x_1 after the influence of x_2 is removed from x_1, but not from y (Fig. 1d, area A/area A + B + C + D [Cohen et al. 2003]). Semipartial correlations measure the increase in r^2 associated with the addition of an explanatory variable, above and beyond all other explanatory variables included in the model. Consequently, they are generally considered a more appropriate measure for regression analysis than partial correlations (Cohen et al. 2003). However, semipartial correlations suffer from the same limitation as zero-order and partial correlations; they cannot partition the variance shared between multiple correlated predictors and the dependent variable.

Standardized regression coefficients

Given a single dependent variable (y) and two explanatory variables (x_1, x_2), the standardized regression coefficient (beta coefficient) for x_1 with respect to y is defined as

$$b'_1 = b_1 \sqrt{\frac{\sum x_1^2}{\sum y^2}}$$

where b_1 is the sample partial regression coefficient for predictor variable x_1. Thus, it represents the change in y that results from a change of one standard deviation in x_1 (Zar 1999). Because standardized regression coefficients are unitless, the importance of each predictor is indicated by the magnitude of its standardized coefficient. Although standardized regression coefficients are often used to assess variable importance, several limitations with this approach have been noted. First, there is no a priori reason to assume that a change of one standard deviation in one predictor should be equivalent to a change of one standard deviation in another predictor (Soofi et al. 2000). Second, collinearity among predictors makes interpretation of standardized regression coefficients difficult, as the value of the overlap area is disproportionately allocated to the explanatory variable with the larger zero-order correlation (LeBreton et al. 2004).

Akaike weights

Given the data and a set of R candidate models, the Akaike weight for the ith model is defined as

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum\limits_{r=1}^{R} \exp\left(-\frac{1}{2}\Delta_r\right)}$$

where Δ_i is the difference between the Akaike information criterion (AIC) values for the model with the lowest AIC value and model i ($\Delta_i = \text{AIC}_i - \text{AIC}_{\min}$ [Burnham and Anderson 2002]). Because the Akaike weights are normalized to one, they indicate the probability that the ith model is actually the best model, of those considered, to use for making valid inference.

Although the primary use of Akaike weights is in model selection, they have also been suggested as a useful technique for assessing the relative importance of variables (Burnham and Anderson 2002). For instance, the relative importance of explanatory variable x_1 can be estimated by summing the Akaike weights across all competing models in the set in which variable x_1 appears. Once the Akaike weights for all variables have been calculated, relative importance is assessed via a rank ordering of the observed values; larger sums indicate a variable is relatively more important that other variables. Note, however, that the weight of each variable is determined by the number of models in which the variable appears, in addition to the weight of those models. Thus, it is important to balance the number of competing models that contain each variable when making comparisons based on Akaike weights (Burnham and Anderson 2002).

Independent effects

Given a single dependent variable (y) and k explanatory variables (x_1, x_2, ..., x_k), the independent effect of predictor x_1 (I_{x_1}) represents the average contribution of variable x_1 to the variance in y over all 2^k possible models. The independent effect of each variable is calculated by comparing the fit of all models containing a particular variable to the fit of all nested models lacking that variable, through the process of hierarchical partitioning (Chevan and Sutherland [1991]; this method is similar to a technique called dominance analysis in the organizational research literature, cf. Budescu [1993]). Thus, for variable x_1,

$$I_{x_1} = \sum_{i=0}^{k-1} \frac{\sum \left(r^2_{y, x_1 x_h} - r^2_{y, x_h}\right) / \binom{k-1}{i}}{k}$$

where x_h is any subset of i predictors, x_1 excluded (Chevan and Sutherland 1991, Budescu 1993).

Hierarchical partitioning is not a substitute for other statistical methods, but rather a complement to any technique that yields a measure of model fit (e.g., linear, loglinear, or logistic regression, probit analysis, etc.). By

averaging each variable's contribution to the dependent variable over all possible combinations of explanatory variables, the variance shared by two or more correlated predictors can be partitioned into the variance attributable to each predictor. In addition, because this method utilizes an all possible models approach, it provides a more robust assessment of variable importance, relative to single-model approaches, by assuring that the contribution of a particular variable is neither enhanced nor masked through its correlation with other explanatory variables (Mac Nally 2000).

SIMULATION METHODS

We simulated four scenarios to evaluate the indices under progressively more complex circumstances. These scenarios corresponded to (1) four uncorrelated explanatory variables, (2) two uncorrelated explanatory variables and two explanatory variables correlated 0.4 with one another, (3) four uncorrelated explanatory variables and a spurious variable, and (4) two uncorrelated explanatory variables, two variables correlated 0.4 with one another, and a spurious variable.

For the simulation with $k = 4$ uncorrelated variables, we first randomly generated 20 000 cases from a normal distribution for each explanatory variable ($\bar{x} = 0$; $\sigma = 1$) plus the dependent variable ($\bar{y} = 500$; $\sigma = 1$). Selection of 500 as the mean for the dependent variable was arbitrary. We used a sample size of 20 000 to eliminate the positive bias in sample r-squared associated with random sampling fluctuations (Cohen et al. 2003). Next, we constructed matrices specifying the desired correlation structure between the dependent variable and each explanatory variable, as well as inter-correlations between the explanatory variables. The latter were set to zero for the simulation based on four uncorrelated explanatory variables. Two correlation matrices were constructed for each simulation, one in which the zero-order correlations summed to produce a total r-squared value of 0.70, and one in which the zero-order correlations summed to produce an r-squared value of 0.35. We selected these r^2 values because they represented a typical range for ecological studies. For each correlation matrix, the zero-order correlations between the dependent variable and explanatory variables were structured such that variable x_1 accounted for 10% of the explained variation in y, x_2 accounted for 20% of the explained variation in y, and x_3 and x_4 each accounted for 35% of the explained variation in y. Note that the specified r-squared values, weights, and correlations between the dependent and explanatory variables were based on direct (zero-order) correlations, only; they did not account for the confounding effects of collinearity or the inclusion of spurious variables, and thus served as the baseline against which to evaluate the actual performance of the various indices. Next, we multiplied the randomly generated variables by the Cholesky decomposition of the correlation matrix (ROOT procedure; SAS Institute 1999). This transformed the

uncorrelated variables into a data set drawn from a population with the desired correlations (Horn and Johnson 1985).

For each sample that consisted of 20 000 cases of the dependent and explanatory variables, we ran all 2^k possible models and calculated the Akaike weights and independent effects for each variable as the average over all 2^k models, and the zero-order correlations from each of the k bivariate models (Appendix A). The partial correlations, semipartial correlations, and standardized regressions coefficients were calculated from the model containing all k explanatory variables (Appendix A). We repeated this process 2500 times for each of the four simulations. At the end of the 2500 trials, we calculated the relative importance of each explanatory variable as the average for each measure over the 2500 trials. All analyses were performed in program SAS (REG procedure; SAS Institute 1999).

For the remaining simulations, we followed a procedure identical to that used for the first simulation with the following exceptions. For the second simulation, the 20 000 cases generated for explanatory variables x_2 and x_4 were correlated 0.40 with one another. For the third simulation, we randomly generated 20 000 cases of a fifth explanatory variable (x_5) that was not correlated with the dependent variable, but was correlated with explanatory variables x_2 and x_4 ($r_{x_2 x_5} = r_{x_4 x_5} = 0.45$). For the fourth simulation, the 20 000 cases generated for explanatory variables x_2 and x_4 were correlated 0.40 with one another, and we included a spurious variable (x_5) that was not correlated with the dependent variable, but was correlated with explanatory variables x_2 and x_4 ($r_{x_2 x_5} = r_{x_4 x_5} = 0.45$).

We evaluated the six indices based on their ability to (1) correctly rank the predictors and determine variable weights and (2) relate the values to the overall measure of model fit. For the simulations in which the explanatory variables were uncorrelated, the true values were equivalent to the zero-order correlations. This was also the case for the simulation involving uncorrelated variables and a spurious variable, since the spurious variable was not part of the generating model that produced the dependent variable. For the simulations based on correlated predictor variables, the true correlations for variables x_2 and x_4 were equivalent to the unique contribution of each variable (i.e., the squared semipartial correlation coefficient) plus half the value of the overlap area. To determine the value of the overlap area, we subtracted the squared semipartial correlation from the zero-order correlation for either variable x_2 or x_4, as using either variable produced the same outcome.

RESULTS

The performance of the indices did not differ materially between the simulations in which the correlation matrix was structured to produce a total r^2 of 0.70 vs. a total r^2 of 0.35. Only the weights of the partial

Reports

TABLE 1. Performance of six measures of variable importance ($r^2 = 0.70$).

Variable	Zero-order correlation		Partial correlation		Semipartial correlation		Standardized coefficient		Akaike weight		Independent effect		True correlation	
	Value†	Weight‡	Value	Weight	Value	Weight	Value	Weight	Value	Weight	Value	Weight	Value	Weight
Four uncorrelated explanatory variables														
x_1	0.070	10%	0.190	13%	0.070	10%	0.070	10%	1.000	25%	0.070	10%	0.070	10%
x_2	0.140	20%	0.318	23%	0.140	20%	0.140	20%	1.000	25%	0.140	20%	0.140	20%
x_3	0.245	35%	0.450	32%	0.245	35%	0.245	35%	1.000	25%	0.245	35%	0.245	35%
x_4	0.245	35%	0.450	32%	0.245	35%	0.245	35%	1.000	25%	0.245	35%	0.245	35%
Total	0.700	100%	1.408	100%	0.700	100%	0.700	100%	4.000	100%	0.700	100%	0.700	100%
Two independent and two correlated explanatory variables ($r_{x_2x_4} = 0.40$)														
x_1	0.070	10%	0.148	17%	0.070	14%	0.070	13%	1.000	25%	0.070	12%	0.070	12%
x_2	0.140	20%	0.084	10%	0.037	7%	0.044	8%	1.000	25%	0.088	15%	0.088	15%
x_3	0.245	35%	0.378	43%	0.245	50%	0.245	46%	1.000	25%	0.245	41%	0.245	41%
x_4	0.245	35%	0.261	30%	0.142	29%	0.169	32%	1.000	25%	0.194	32%	0.194	32%
Total	0.700	100%	0.871	100%	0.494	100%	0.528	99%	4.000	100%	0.597	100%	0.597	100%
Four uncorrelated explanatory variables and a spurious variable														
x_1	0.070	10%	0.621	15%	0.070	5%	0.070	4%	1.000	20%	0.070	7%	0.070	10%
x_2	0.140	20%	0.886	21%	0.335	24%	0.449	25%	1.000	20%	0.211	22%	0.140	20%
x_3	0.245	35%	0.851	21%	0.245	18%	0.245	13%	1.000	20%	0.245	26%	0.245	35%
x_4	0.245	35%	0.916	22%	0.467	34%	0.626	34%	1.000	20%	0.329	34%	0.245	35%
x_5	0.000	0%	0.857	21%	0.257	19%	0.432	24%	1.000	20%	0.102	11%	0.000	0%
Total	0.700	100%	4.131	100%	1.374	100%	1.822	100%	5.000	100%	0.957	100%	0.700	100%
Two independent variables, two correlated variables ($r_{x_2x_4} = 0.40$) and a spurious variable														
x_1	0.070	10%	0.193	12%	0.070	10%	0.070	8%	1.000	20%	0.070	10%	0.070	12%
x_2	0.140	20%	0.224	14%	0.084	12%	0.113	13%	1.000	20%	0.110	16%	0.088	15%
x_3	0.245	35%	0.455	29%	0.245	34%	0.245	28%	1.000	20%	0.245	35%	0.245	41%
x_4	0.245	35%	0.425	27%	0.216	30%	0.288	33%	1.000	20%	0.229	32%	0.194	32%
x_5	0.000	0%	0.272	17%	0.110	15%	0.154	18%	1.000	20%	0.053	7%	0.000	0%
Total	0.700	100%	1.569	99%	0.725	101%	0.870	100%	5.000	100%	0.707	100%	0.597	100%

† True r_{yx}^2.
‡ Scaled r_{yx}^2; scaled by total model r^2 for comparison to value used for variable weight.

correlations differed slightly (Appendix B). Thus, we present only the results of the simulations based on the higher r^2 value.

Uncorrelated explanatory variables

Zero-order correlations, semipartial correlations, standardized regression coefficients, and independent effects performed equally well when the explanatory variables were uncorrelated. All four indices correctly ranked the explanatory variables in the order of importance, and produced accurate estimates of the contribution of each explanatory variable to the variance in y (Table 1). In addition, each of these indices produced a total r^2 that was equal to the true value of 0.70 specified by the correlation matrix. Partial correlation coefficients ranked the variables correctly in order of importance; however, the weights were slightly overestimated for variables x_1 and x_2 and underestimated for variables x_3 and x_4. Furthermore, interpretation of variable importance was not straightforward using partial correlations, because the values did not equal the true correlations nor sum to the r-squared value for the full model (Table 1). Akaike weights performed poorly in all respects. Despite considerable differences in the true r^2 values, all four explanatory variables were assigned equal importance. Furthermore, because the unit of measure for Akaike weights differs from that of the other indices, the values are not interpretable with respect to the true correlations of the explanatory variables (Table 1).

Correlated explanatory variables

Hierarchical partitioning was the only method to correctly partition the shared variance between variables x_2, x_4, and the dependent variable (Table 1). It produced independent effects that correctly ranked the importance of the explanatory variables and reflected the true contributions of each independent and correlated explanatory variable to the variance in y. In addition, it was the only method to detect the effective decrease in model r-squared from 0.70 to ~0.60 as a result of the redundancy in variables x_2 and x_4. As expected, zero-order correlations overstated the contribution of variables x_2 and x_4, as well as the total r-squared, because the overlap area was attributed to both variables. Partial correlations also overestimated the contributions of variables x_2 and x_4 because of its inability to partition the shared variance; this index had the highest cumulative value, which overstated the true r^2 by 46%. Conversely, semipartial correlations understated both the total r^2 and the importance of variables x_2 and x_4, as the overlap area was omitted from the contributions of

both variables. Similarly, standardized regression coefficients underestimated the importance of variable x_2 by nearly 50%, as most of the overlap area was assigned to variable x_4. The cumulative value of the standardized regression coefficients also slightly understated model r^2. Akaike weights could not detect the overlap area and still assigned equal weights to all variables.

Uncorrelated explanatory variables and a spurious variable

Zero-order correlations was the only index that identified variable x_5 as a spurious variable (i.e., a zero-order correlation of zero), and produced values that correctly ranked the importance of the explanatory variables and reflected the true contributions of each explanatory variable to the variance in y (Table 1). Further, it was the only index that generated a cumulative value equal to the true r^2 of 0.70 specified by the correlation matrix; the cumulative values for all other indices overstated total r^2. All indices except zero-order correlations assigned a higher value to the spurious variable (x_5) than for variable x_1, and overstated the contributions of variables x_2 and x_4, as these latter variables picked up additional explanatory power from their correlation with the spurious variable. Partial correlations, semipartial correlations and standardized regression coefficients also produced cumulative values that exceeded 1.0, rendering them uninterpretable within the context of r^2. Independent effects had the same limitations as partial correlations, semipartial correlations, and standardized regression coefficients, but produced values that were closer to the true contribution of each explanatory variable. Akaike weights could not identify the presence of a spurious variable, and assigned equal weight to all variables.

Correlated explanatory variables and a spurious variable

Again, zero-order correlations was the only index to correctly identify variable x_5 as a spurious variable (i.e., zero-order correlation of zero; Table 1). However, the contributions of variable x_2 and x_4, and well as the cumulative value, were overstated as a result of the inability of zero-order correlations to partition the overlap area. Partial correlations, semipartial correlations, and standardized regression coefficients all assigned a value to the spurious variable (x_5) that exceeded the contributions of variables x_1 and x_2, which were included in the generating model. Interestingly, although semi-partial correlations and standardized regression coefficients incorrectly stated the importance of variables x_2 and x_4, the estimated values were closer to the true values than those produced by the third simulation; increases in the values resulting from correlations with the spurious variable were tempered by decreases associated with the overlap area. The cumulative values, though less than 1.0, were still overstated. Independent effects correctly ranked the variables in order of importance, but assigned some weight to the spurious

variable. Akaike weights could not identify the presence of a spurious variable, and assigned equal weight to all variables.

DISCUSSION

When predictor variables are uncorrelated, interpretation of variable importance is straight forward and the selection of an index simple; zero-order correlations, semipartial correlations, standardized regression coefficients, and independent effect all perform equally well. However, in ecological studies explanatory variables are often only nominally independent, and no single technique to assess variable importance performs satisfactorily under all circumstances. While zero-order correlations can identify spurious variables that have no relationship with the dependent variable but are correlated with other explanatory variables in the model, this index cannot assign overlap areas to determine variable importance. Conversely, independent effects cannot detect the presence of spurious variables, but it is the only index that can correctly partition shared variance and detect the effective reduction in total r^2 that results from redundancy in the correlated variables.

Akaike weights are primarily used to select among models with different combinations and numbers of variables, optimizing between model over-fit, with its corresponding high variance, and model under-fit, with its corresponding high bias. Beyond this, Akaike weights are useful for comparing the weight of evidence between models, evaluating the support for sets of candidate models containing predictor variables of interest, or calculating a model-averaged parameter estimate (Burnham and Anderson 2002). However, because weightings apply to each model as a whole, rather than to individual variables, we found this index was not sufficiently sensitive to correctly rank variable importance. In our simulations, the large sample size of 20 000, used to facilitate comparisons among the indices by eliminating the positive bias in r^2 that occurs in randomly generated samples (Cohen et al. 2003), resulted in no model-selection uncertainty. Consequently, the top-ranked model, and all corresponding variables contained within that model, received an Akaike weight of 1. Although the Akaike weights returned by our simulations were, to some extent, an artifact of the large sample size we used, the performance of this index was not materially improved when we repeated the simulations with a sample size of 100 to increase model selection uncertainty (Appendix B). Furthermore, although we expected AIC to consistently choose the true model, the five-variable model that included the spurious variable consistently received an Akaike weight of 1, while the true four-variable generating model received an Akaike weight of 0. Thus, this index was not appropriate for identifying spurious variables, or for ranking variable importance, and should not be used in these capacities; rather, Akaike

Reports

weights should be used for model selection prior to assessment of variable importance using alternative methods.

Correlation among predictor variables

Problems with multicollinearity, including inflated standard errors for the predictor coefficients with a concomitant increase in Type II errors, instability in coefficient estimates, and erroneous rankings of variable importance, are well-documented in the statistical (e.g., Zar 1999, Cohen et al. 2003, Belsley et al. 2004, Gotelli and Ellison 2004) and ecological (e.g., Mac Nally 2000, Graham 2003) literature. Although suggested cut-off values vary widely, remediation is typically not recommended unless pairwise correlations exceed ~0.80 (Katz 2006). However, we found that even the relatively low level of correlation ($r = 0.40$) incorporated in our simulations was sufficient to preclude an accurate assessment of variable importance. Similarly, Graham (2003) reported that correlations as low as 0.28 resulted in inaccurate model parameterization. Nearly all the indices we tested distorted the relative contributions of the predictor variables, and produced incorrect rank orderings, when correlation between the explanatory variables was included in the model. This was due to the inability of the indices to correctly assign the overlap area, which caused the contributions of the collinear variables (x_2 and x_4) to be understated. Only independent effects produced values for the predictor variables that reflected their true contributions towards the variance in y. Thus, failure to consider the influence of correlation on variable rankings, or reliance on traditional measures of variable importance such as semipartial correlations or standardized regression coefficients, will likely lead to a misidentification of priorities, and attendant inefficient allocation of resources, when selecting among competing management alternatives.

Identification of spurious variables

In recent years, scientists have raised concerns regarding the reporting of spurious effects in biological journals, especially for exploratory studies where relationships between the dependent variable and numerous factors are examined (Anderson et al. 2001). Of the indices tested, we found that only zero-order correlations could reliably detect the presence of a spurious variable (i.e., return a value/weight of zero). All other indices assigned at least some weight to the spurious variable, and frequently ranked the importance of the spurious variable ahead of predictors that were included in the true model. Furthermore, for all indices except zero-order correlations, inclusion of the spurious variable caused the model r^2 to be biased high. Thus, in the absence of some form of screening, the temptation will be to include the spurious variable in the final model due to the apparent improvement in explanatory power.

The indiscriminate use of statistical tests to screen variables for significance is discouraged because it inflates the risk of committing a Type I error (Gotelli and Ellison 2004). However, statistical significance is not an indication of variable importance. Rather, it reflects the strength of confidence in making inference about an unknown parameter based on a statistic (Soofi et al. 2000). We note that even with a reduced sample size of 100, the spurious variable remained highly significant ($P < 0.0001$) in our simulations. In contrast with significance testing, zero-order correlations rely on the existence of non-zero correlations between dependent and explanatory variables, rather than P values, to screen for spurious variables. Thus, the use of zero-order correlations to identify spurious variables does not carry an increased risk of committing a Type I error.

We caution that our simulations were designed to test for a specific type of spurious variable, one that was partially correlated with other explanatory variables in the generating model, but not with the dependent variable. Although zero-order correlations performed well under these circumstances, this index cannot be used to identify spurious effects in situations in which the dependent and explanatory variables are correlated but no biological relationship exists. In addition, none of the methods we tested will directly identify interaction effects, in which the relationship between a dependent and explanatory variable is moderated by the presence of a third interacting variable; these cases must be identified via other procedures such as the inclusion of an interaction term in the model or experimental testing using a factorial design (Cohen et al. 2003, Gotelli and Ellison 2004). Thus, the use of zero-order correlations should not be viewed as a substitute for careful a priori thinking about the biological basis for including variables in an analysis (Anderson et al. 2001).

Conclusion

Based on our simulation results, we recommend using zero-order correlations to eliminate predictor variables that have correlations with the response variable near zero, as the first step in variable importance analyses. Once spurious variables have been identified and eliminated, hierarchical partitioning can assign overlap areas and rank variable importance via independent effects. We stress that correlation does not imply causation and we make no claims about the ability of the indices tested to resolve the difficult issue of establishing causality among variables; validation of variable importance can only be determined by experimental means. Thus, this approach should be used to generate testable hypotheses regarding alternative management strategies that can be subsequently tested through application in an adaptive management framework. At a time when many resource management agencies are faced with shrinking budgets, we suggest that this approach can assist ecologists and wildlife managers in prioritizing management strategies and in allocating scarce resource to their most productive use.

Literature Cited

Anderson, D. R., K. P. Burnham, W. R. Gould, and S. Cherry. 2001. Concerns about finding effects that are actually spurious. Wildlife Society Bulletin 29:311–316.

Belsley, D. A., E. Kuh, and R. E. Welsch. 2004. Regression diagnostics: identifying influential data and sources of collinearity. John Wiley and Sons, Hoboken, New Jersey, USA.

Berry, W. D., and S. Feldman. 1985. Multiple regression in practice. Sage Publications, Beverly Hills, California, USA.

Budescu, D. V. 1993. Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression. Psychological Methods 8:542–551.

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

Chevan, A., and M. Sutherland. 1991. Hierarchical partitioning. American Statistician 45:90–96.

Cohen, J., S. G. West, L. Aiken, and P. Cohen. 2003. Applied multiple regression/correlation analysis for the behavioral sciences. Lawrence Erlbaum Associates, Mahwah, New Jersey, USA.

Farmer, C. J., D. K. Person, and R. T. Bowyer. 2006. Risk factors and mortality of black-tailed deer in a managed forest landscape. Journal of Wildlife Management 70:1403–1415.

Gotelli, N. J., and A. M. Ellison. 2004. A primer of ecological statistics. Sinauer Associates, Sunderland, Massachusetts, USA.

Graham, M. H. 2003. Confronting multicollinearity in ecological multiple regression. Ecology 84:2809–2815.

Horn, R. A., and C. R. Johnson. 1985. Matrix analysis. Cambridge University Press, New York, New York, USA.

Katz, M. H. 2006. Multivariable analysis: a practical guide for clinicians. Cambridge University Press, New York, New York, USA.

LeBreton, J. M., R. E. Ployhart, and R. T. Ladd. 2004. A Monte Carlo comparison of relative importance methodologies. Organizational Research Methods 7:258–282.

Mac Nally, R. 2000. Regression and model-building in conservation biology, biogeography and ecology: the distinction between—and reconciliation of—"predictive" and "explanatory" models. Biodiversity and Conservation 9:655–671.

Nichols, J. D. 2001. Using models in the management of natural resources. Pages 11–34 in T. M. Shenk and A. B. Franklin, editors. Modeling in natural resource management. Island Press, Washington, D.C., USA.

Ottichilo, W. K., J. de Leeuw, and H. H. T. Prins. 2001. Population trends of resident wildebeest [Connochaetes taurinus hecki (Neumann)] and factors influencing them in the Masai Mara ecosystem, Kenya. Biological Conservation 97:271–282.

SAS Institute. 1999. SAS OnlineDoc version 8. SAS Institute, Cary, North Carolina, USA.

Soofi, E. S., J. J. Retzer, and M. Yasai-Ardekani. 2000. A framework for measuring the importance of variables with applications to management research and decision models. Decision Sciences 31:595–625.

Stokes, K. E., and S. A. Cunningham. 2006. Predictors of recruitment for willows invading riparian environments in south-east Australia: implications for week management. Journal of Applied Ecology 43:909–921.

Williams, B. K. 1997. Logic and science in wildlife biology. Journal of Wildlife Management 61:1007–1015.

Zar, J. H. 1999. Biostatistical analysis. Prentice Hall, Upper Saddle River, New Jersey, USA.

APPENDIX A

Performance of six measures of variable importance (*Ecological Archives* E090-026-A1).

APPENDIX B

Performance of Akaike weights with a reduced sample size of 100 to increase model-selection uncertainty (*Ecological Archives* E090-026-A2).

SUPPLEMENT

Source code for variable importance situations (*Ecological Archives* E090-026-S1).

Reports