Ecology Letters, (2009) **12**: 1061–1068 doi: 10.1111/j.1461-0248.2009.01361.x

LETTER

Performance of several variable-selection methods applied to real ecological data

Paul A. Murtaugh*
Department of Statistics, Oregon State University, Corvallis, OR 97331, USA
**Correspondence: E-mail: murtaugh@science.oregonstate.edu*

Abstract

I evaluated the predictive ability of statistical models obtained by applying seven methods of variable selection to 12 ecological and environmental data sets. Cross-validation, involving repeated splits of each data set into training and validation subsets, was used to obtain honest estimates of predictive ability that could be fairly compared among methods. There was surprisingly little difference in predictive ability among five methods based on multiple linear regression. Stepwise methods performed similarly to exhaustive algorithms for subset selection, and the choice of criterion for comparing models (Akaike's information criterion, Schwarz's Bayesian information criterion or F statistics) had little effect on predictive ability. For most of the data sets, two methods based on regression trees yielded models with substantially lower predictive ability. I argue that there is no 'best' method of variable selection and that any of the regression-based approaches discussed here is capable of yielding useful predictive models.

Keywords

AIC, all subsets, BIC, regression tree, statistical model building, stepwise, subset selection, variable selection.

Ecology Letters (2009) 12: 1061–1068

INTRODUCTION

Building statistical models of a response as a function of multiple explanatory variables is a common exercise in ecology. Such models serve a variety of purposes, including prediction of responses for new cases (Leigh *et al.* 2008); risk estimation (Gerritsen *et al.* 1996); understanding, or at least forming hypotheses about, cause-and-effect relationships (Knick & Rotenberry 1995); and constructing parsimonious models of spatial or temporal correlation of response values (Hoeting *et al.* 2006; Lee & Ghosh 2009).

Many criteria are available for the statistical comparison of multiple-variable models. Recently, information-theoretic criteria such as Akaike's information criterion (AIC) and Schwarz's Bayesian information criterion (BIC) have gained favour among ecologists (Burnham & Anderson 2002; Hobbs & Hilborn 2006; Ward 2008). Some of the appeal of these approaches seems to be based on their independence from the hypothesis-testing framework of frequentist statistics. Mazerolle (2006), for example, writes that the AIC "is remarkably superior in model selection (i.e. variable selection) than hypothesis-based approaches", and Lukacs *et al.* (2007) comment that "exploratory data analysis based on null hypothesis testing methods such as stepwise selection simply removes thought from data analysis".

The oldest algorithms for selecting explanatory variables are stepwise procedures, in which candidate predictors are screened for possible inclusion and variables already in the model are considered for possible removal in a sequential fashion. More modern algorithms involve exhaustive searches of many more subsets of predictors than are usually evaluated in stepwise procedures, leading some ecologists to doubt the usefulness of stepwise variable selection. Whittingham *et al.* (2006), for example, bemoan the widespread use of stepwise procedures in ecological and behavioural journals, given the well-established 'biases and shortcomings of stepwise multiple regression', and Mundry & Nunn (2009) 'follow others in recommending that biologists refrain from applying these methods'.

Many authors have discussed and evaluated different methods of variable selection (e.g. see Olden & Jackson 2000; Sauerbrei *et al.* 2007; Ward 2008; Lee & Ghosh 2009). Raffalovich *et al.* (2008) provide an excellent review of work done in this area. Simulation is the only approach in which the 'true' model can be known, but the conclusions from simulation studies are very dependent on the ways that data are generated in the simulations and the measures that

are chosen for comparison of different model-building techniques. Murtaugh (1998), for example, found that approaches based on F tests, Mallows' C_p, and the BIC had similar frequencies of correct decisions about inclusion or exclusion of explanatory variables. Regression-tree approaches were markedly inferior to the regression-based methods, but Murtaugh (1998) questioned the validity of the comparison, given that the data were simulated according to a linear regression model.

Here I examine the performance of seven methods of variable selection applied to 12 data sets that were obtained, opportunistically, from the literature and from colleagues. As there are no known 'true' models of responses as functions of explanatory variables, I used cross-validation to obtain honest estimates of predictive ability that can be fairly compared among different methods of variable selection. This approach avoids the arbitrariness of choosing an algorithm to generate data in simulation studies, but it produces results of unknown generality, as the data sets used are a select subset of the enormous variety of data sets to which variable-selection techniques are applied in practice.

Approaches to variable selection

Multiple linear regression is a familiar way of modelling a quantitative response variable as a function of multiple explanatory variables. Approaches to selecting variables for inclusion in the model from a pool of candidate predictors include:

(1) Stepwise procedures, in which some quantitative criterion is used to compare regression models with and without a particular predictor, and sequential addition and/or deletion of explanatory variables continues until a stopping point based on the value of the criterion is reached; and
(2) All subsets, or exhaustive, variable selection, in which the set of all possible groupings of explanatory variables is searched and subsets of predictors giving the most favourable values of the quantitative criterion are identified.

The criteria that are used to compare regression models include:

1. P-values from extra-sum-of-squares F tests. To compare models with and without a particular predictor, we compute

$$F^* = \frac{\text{SSE}_{\text{without}} - \text{SSE}_{\text{with}}}{\text{d.f.}_{\text{without}} - \text{d.f.}_{\text{with}}} \div \frac{\text{SSE}_{\text{with}}}{\text{d.f.}_{\text{with}}}, \quad (1)$$

where SSE is the error, or residual, sum of squares, and d.f. is the number of residual degrees of freedom. The P-value is then obtained by comparing F^* to an F distribution with the corresponding numerator and denominator degrees of freedom. In the case of a quantitative predictor, F^* is the square of the t statistic for that predictor in the regression output.

(2) Akaike's information criterion. For a regression model with Gaussian errors, this statistic can be written (Ramsey & Schafer 2002, p. 356):

$$\text{AIC} = n\log(\text{MSE}) + 2p, \quad (2)$$

where n is the number of observations, p is the number of regression coefficients and MSE is the mean square error, equal to the residual sum of squares divided by its degrees of freedom ($n-p$).

(3) Schwarz's BIC. For regression with Gaussian errors, this can be written (Ramsey & Schafer 2002, p. 356):

$$\text{BIC} = n\log(\text{MSE}) + p\log n. \quad (3)$$

Notice how all three of these statistics balance explained variation against the number of predictors in a model: as the number of explanatory variables increases, the residual sum of squares decreases, but a penalty for model complexity increases (reflected in the values of p and the residual degrees of freedom).

Classification and regression trees provide a method of model building that is very different from the regression approaches discussed above (Breiman *et al.* 1984; De'ath & Fabricius 2000). Starting with a 'root' corresponding to the whole data set, the method produces successive splits of the data set based on values of the explanatory variables. For quantitative predictors, at each level of the tree, the approach considers all possible binary splits of the predictors, with each split leading to a pair of predicted responses equal to the response means in the two groups created by the split. A bifurcation is created for the predictor and cutpoint that lead to the smallest deviance, or sum of the squared differences between observed and predicted responses. This procedure is repeated recursively to produce a branching tree of binary splits based on one or more of the explanatory variables.

Unlike the linear regression approaches, the regression tree is not based on a statistical model that can be used to quantify the trade-off between model complexity and explained variation. Instead, model selection is accomplished by 'pruning' overly complicated trees back to simpler trees that will presumably have greater generality. A commonly used approach involves cross-validation, in which, over multiple iterations, a tree based on all but a small subset of the data is used to predict responses for the 'validation' subset (e.g. see Therneau & Atkinson 1997).

METHODS

The data sets

Twelve data sets were obtained by searching literature and on-line sources, and by making enquiries among colleagues. The goal was to find ecological or environmental data sets that included quantitative variables, at least one of which could be considered as a 'response' predictable by the others. Table 1 presents some features of the data sets, which are described in more detail in Appendix S1 and are available from the author upon request.

Variable selection

The variables in each data set were examined individually prior to model fitting. Response variables and predictors with extreme skewness were log-transformed to reduce the chance that variable selection would be strongly influenced by a few extreme observations. I then applied seven methods of variable selection and tree building that I automated in R (R Development Core Team 2007). In the regression modelling, interactions between predictors were not included in the pool of explanatory variables.

It should be emphasized that automatic selection of variables is not good statistical practice; an iterative and interactive approach is much more likely to yield useful models (Chatfield 2002). Among other things, automating the process makes it difficult to consider the proper functional forms of predictors, possible spatial and temporal correlation of observations, and interactions and collinearity among explanatory variables.

The methods of model fitting are as follows.

(1) *Stepwise variable selection with F tests (Efroymson's algorithm).* *P*-values are obtained by comparing models with and without a particular explanatory variable with an extra-sum-of-squares *F* test (eqn 1). In this paper, I use $P = 0.05$ as the threshold for inclusion or exclusion of predictors; different values of *P* would lead to different levels of model complexity. Starting with a model having no predictors, we repeat the following steps: (i) for the predictors not in the model, add the one having the smallest, statistically significant '*P* to enter', and (ii) if any of the predictors in the model now have a nonsignificant '*P* to stay', drop the one with the largest *P*-value. The procedure ends when all of the variables in the model, and none of those outside the model, have $P<0.05$.

(2) *Stepwise variable selection using AIC.* Starting with a model having no predictors, we repeat the following steps: (i) for the predictors not in the model, add the one leading to the largest reduction in the AIC (eqn 2), and (ii) if any of the predictors in the model lead to a reduction in the AIC when dropped, remove the one producing the largest reduction. The procedure ends when there are no further variable additions or deletions that can lower the AIC.

Table 1 Description of the data sets. R^2 is the coefficient of determination for a model containing all of the candidate predictors. The condition index is a measure of collinearity of the predictors (Belsley *et al.* 1980). Appendix S1 has more details about the data sets

Label	Description	No. of observations	No. of predictors	R^2	Cond. index
A	Chlorophyll *a* in north-eastern U.S. lakes, predicted from water chemistry	348	10	0.67	102
B	Bird species richness around north-eastern U.S. lakes, predicted from lake and watershed characteristics	185	8	0.22	32
C	Species richness of native fish in north-eastern U.S. lakes, predicted from watershed characteristics and lake biota	194	8	0.48	26
D	Seed production by weedy rice, predicted from plant morphology	356	7	0.96	84
E	Wave height in the Pacific Ocean, predicted from weather variables	335	7	0.94	585
F	Faecal coliform bacteria in Oregon rivers, predicted from water chemistry	77	8	0.32	121
G	Biochemical oxygen demand in the Deschutes River, predicted from flow, water chemistry and temperature	26	8	0.58	126
H	Sleep duration for mammal species, predicted from life-history characteristics, weight and exposure	51	7	0.70	29
J	Human population density in stream watersheds, predicted from watershed characteristics	310	7	0.33	44
K	Secchi depth in mid-Atlantic estuaries, predicted from water temperature and chemistry	870	8	0.25	180
L	Abundance of caddisfly larvae in an Oregon stream, predicted from local stream and substrate characteristics	311	5	0.27	23
M	Zooplankton species richness in North American lakes, predicted from area, depth, elevation and proximity to other lakes	66	8	0.64	29

(3) *Stepwise variable selection using Schwarz's BIC*. As in (2), but with the BIC (eqn 3).
(4) *All subsets using the AIC*. The R function 'regsubsets' identifies the best subsets of predictors using a branch-and-bound algorithm (Miller 2002). I ranked the subsets according to their values of the AIC, and chose the subset having the minimum value.
(5) *All subsets using the BIC*. As in (4), except that rankings are based on the BIC.
(6) *Regression trees pruned by the 1-SE rule* (Breiman *et al.* 1984). I used functions in R's 'rpart' package (Therneau & Atkinson 1997) to fit regression trees, which were then pruned back to avoid overfitting. The pruning algorithm uses cross-validation to identify trees with small values of a risk measure that balances explained variation against tree complexity. The 1-SE rule chooses the simplest tree having risk that is within 1 standard error of the achieved minimum.
(7) *Regression trees pruned to the minimum risk*. As in (6), except that the pruned tree with the minimum risk is chosen. This introduces some randomness into tree selection that the 1-SE rule, above, seeks to avoid, but it also provides a less aggressive pruning method for comparison with the previous method.

Cross-validation

The predictive ability of each method applied to a particular data set was estimated using cross-validation (e.g. see Harrell 2001, p. 93). For each data set, the following steps were repeated 2000 times.

(1) Randomly divide the data set into a training subset consisting of about 75% of the observations and a validation subset consisting of the remaining 25% of the observations.
(2) For each variable-selection method, use the method to build a predictive model based on the training data.
(3) Apply the model obtained in step 2 (regression coefficients or regression tree) to the explanatory variables for observations in the validation subset to predict responses for the validation subset. Compute the cross-validation mean squared prediction error:

$$\text{MSPE} = \frac{\sum_{i=1}^{n^*}(Y_i - \hat{Y}_i)^2}{n^*},$$

where Y_i is the observed response for item i, $\hat{Y}_i$ is the response predicted by the model based on the training data, and the sum is over the n^* observations in the validation subset. In cases where no variables were selected for a particular set of training data, the mean of the responses in the training data was used as the predicted response for all items in the validation subset.

For each data set and variable-selection method, the 2000 values of the cross-validation MSPE were averaged to give an overall summary of predictive ability.

RESULTS

An example

Table 2 shows modelling results for data set C, in which the response is the species richness of native fish in north-eastern U.S. lakes. I chose this data set because it yields the largest number of different models, out of the 12 data sets studied here. The regression-based methods resulted in three different models, having two, three and four explanatory variables. Two more distinct 'models' resulted from the regression-tree approaches.

Figure 1 shows the regression tree obtained with method 7. The first split, which is the only split produced by method 6, indicates that the lakes should first be discriminated on the basis of area: those with log-transformed area less than 2.88 have a mean log-transformed species richness of 1.41, and the remaining lakes have a mean response of 2.07. These subgroups of lakes are then further split according to their values of the zooplankton, elevation, depth and area variables. Notice in Fig. 1 that the association of depth with species richness is negative in the split on the left-hand side of the tree and positive in the right-hand split. This is an example of the very flexible 'modelling' of the varying association of a predictor and response over subgroups that is available with regression trees.

Number of explanatory variables included

Table 3 shows, for each combination of data set and method, the average number of explanatory variables included in models fit to the training data sets. A conspicuous pattern is the generally smaller number of variables included by the regression-tree methods (6 and 7), compared to the other methods. However, because a single predictor can occur at more than one node of a regression tree (e.g. see lake area in Fig. 1), the comparison of number of predictors in trees vs. ordinary regression models provides an imperfect contrast of model complexity.

Figure 2a is a graphical summary of the results presented in Table 3. On average, methods using the AIC (2 and 4) yielded the largest models, which is consistent with the AIC's relatively small penalty for model complexity. Models of intermediate size were produced by the methods based on F-tests (1) and the BIC (3 and 5).

Interestingly, models fit using the all-subsets algorithm were about the same size as those fit using stepwise procedures, for both the AIC (4 vs. 2 in Fig. 2a) and BIC (5 vs. 3).

Predictive ability

Table 4 and Fig. 2b summarize the patterns of MSPE found in the different combinations of data set and variable-selection method.

Tree-based vs. regression-based approaches

The two tree-based methods (6 and 7) generally have higher squared prediction errors than those of the regression-based methods (1–5), although the pattern is reversed for two of the 12 data sets (F and G; see Table 4).

Taking a closer look at these two data sets, we see that both have relatively high condition indices – a measure of the collinearity of the explanatory variables – and small sample sizes. High collinearity usually inflates the variability of regression coefficients and predicted responses, which would tend to increase the magnitude of squared prediction errors, especially when the training data sets are small. In such cases, the more parsimonious models produced by the tree-based methods may outperform the regression-based approaches, at least with respect to this measure of predictive ability.

Data set G is an extreme example: in *none* of the cross-validations did the tree-based methods select *any* explanatory variables, yet the associated MSPEs (equal to the population variances of the responses in the training data sets) were usually smaller than those for the regression-based approaches, which, on average, made use of 1.3 to 3.3 predictors (Table 3).

This is not to say that regression trees without branches are useful predictive tools. But the pruning algorithms that resulted in these 'root-only' trees are in a sense protection against the overfitting and variance inflation that can occur when one fits regression models to small data sets having relatively large numbers of collinear predictors.

Comparisons among the regression-based approaches

Maybe the most interesting feature of Fig. 2b is the close similarity of the MSPEs associated with the regression-based approaches (1–5). In spite of the larger number of predictors included by the AIC-based methods (2 and 4), compared to the *F*-test and BIC-based methods (1, 3 and 5; see Fig. 2a), the mean predictive abilities of all five methods were nearly identical.

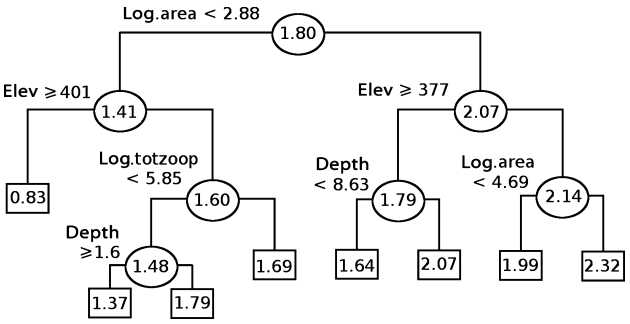

Figure 1 Regression tree for predicting species richness of native fish in north-eastern U.S. lakes (data set C), using the minimum-risk criterion (method 7). The numbers in the ovals ('nodes') and rectangles ('leaves') give the means of the log (number of species) for lakes with different combinations of predictor values, determined by splits higher in the tree. The tree chosen by the 1-SE rule (method 6) has just the single, initial split based on log (area).

DISCUSSION

Applied to the 12 data sets in this paper, the five regression-based methods of variable selection produced models with very similar predictive ability, while the performances of the two tree-based methods were usually inferior (Fig. 2b). There was surprisingly little difference between stepwise and all-subsets approaches (methods 4 vs. 2, and 5 vs. 3) with respect to either model size (Fig. 2a) or predictive ability (Fig. 2b). These results contrast with some ecologists' assertions about the shortcomings of stepwise procedures (e.g. see Whittingham *et al.* 2006; Mundry & Nunn 2009).

As expected, the models with the largest number of predictors were obtained with the two AIC-based methods

Table 2 Results of applying the seven variable-selection techniques to data set C, predicting species richness of native fish in north-eastern U.S. lakes. Lake area was log transformed; chlorophyll *a*, the number of non-native species, and the response received the $\log(y + 1)$ transformation. Entries for the predictors are regression coefficients for the variables included in the model, or, for regression trees (methods 6 and 7), the direction of the 'effect' of the predictor on the response (see Fig. 1, and discussion in text). The value of the cross-validation mean squared prediction error (MSPE) reported for models 2, 4 and 5 is the average over the three methods

Method	Lake area (ha)	Elev. (m)	Mean depth (m)	Chl. *a* ($\mu g\ L^{-1}$)	No. of non-native fish sp.	Total zoopl. indiv.	Cross-val. MSPE
1	0.180	−0.00125		0.0811			0.187
2, 4, 5	0.198	−0.00133		0.0884	−0.0996		0.185
3	0.168	−0.00131					0.185
6	+						0.226
7	+		+/−			+	0.214

Table 3 Mean numbers of variables included in models produced by the seven methods applied to the 12 data sets (see Table 1 for the correspondence between letters and data sets). Each entry is the average number of predictors in models fit to the training data in 2000 random splits. This is an imperfect summary of the complexity of regression trees, in which a single explanatory variable may be used at more than one level of the tree. Methods: 1, stepwise F tests; 2, stepwise Akaike's information criterion (AIC); 3, stepwise Bayesian information criterion (BIC); 4, all subsets AIC; 5, all subsets BIC; 6, tree with 1-SE rule; 7, tree with minimum risk

	Method						
Data set	1	2	3	4	5	6	7
A	3.2	4.1	3.0	4.9	3.1	1.1	2.2
B	2.8	3.4	2.5	3.4	2.6	0.4	1.6
C	2.5	3.5	2.3	3.5	2.3	1.5	2.5
D	2.1	2.9	1.6	3.1	1.6	1.0	1.0
E	4.8	5.6	4.4	5.5	4.4	2.2	2.3
F	1.6	2.1	1.5	2.0	1.5	0.8	1.5
G	1.3	3.3	2.2	3.3	3.2	0.0	0.0
H	2.4	3.4	2.5	3.3	2.6	1.0	1.4
J	4.8	6.1	4.1	6.1	4.2	1.1	2.8
K	3.0	3.1	2.8	3.2	2.8	1.8	4.0
L	3.5	3.9	2.9	3.9	2.9	2.2	3.3
M	1.9	2.2	1.9	2.2	1.9	0.9	1.7

(Fig. 2a). As the predictive ability of these models was not markedly better than that of models produced by approaches based on F-tests and the BIC (Fig. 2b), the choice among variable-selection techniques must be guided by other considerations. For example, one can weigh the relative costs of the different kinds of mistakes that can be made in model building: including a predictor that is truly uninformative, or excluding one that is in fact informative (Murtaugh 1998). Depending on which kind of mistake is more costly or consequential, one might prefer a more conservative (F-test or BIC-based) or more liberal (AIC-based) method of variable selection.

As many statisticians have pointed out, it is difficult or impossible to interpret the P-values for the explanatory variables in a regression model that was obtained by winnowing a multitude of other possible models, as those P-values do not account for the so-called model selection uncertainty (e.g. see Miller 2002; Ramsey & Schafer 2002). Some authors have consequently denigrated the use of F tests in variable selection (e.g. see Burnham & Anderson 2002; Mundry & Nunn 2009). But the F statistic can nevertheless be a useful currency for expressing the trade-off between explained variation and model complexity. In fact, the F-to-enter, the adjusted R^2 and the AIC can all be viewed as special cases of a generalized form of Mallow's C_p statistic (Miller 2002, p. 205).

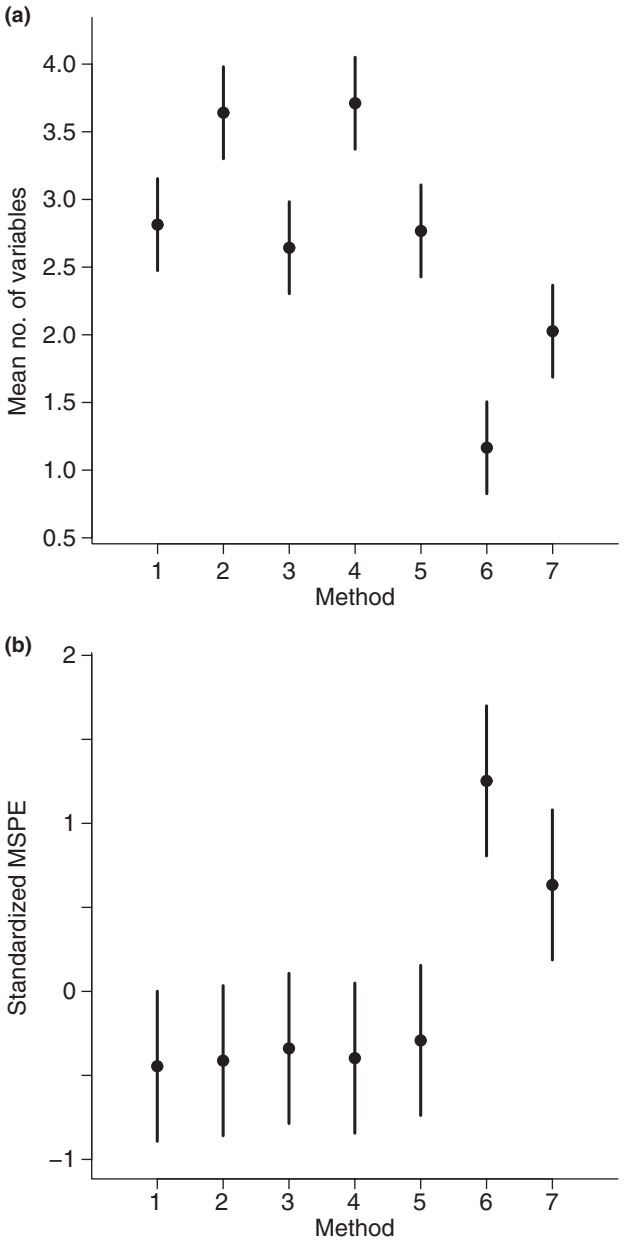

Figure 2 Summary statistics for the seven methods, averaged over data sets: (a) mean number of predictors per model (see the comment about regression trees in the legend of Table 3); and (b) mean value of the cross-validation mean squared prediction error (MSPE), standardized as described in the legend of Table 4. The vertical lines are 95% confidence intervals based on the mean square error (MSE) from linear models of the responses as a function of method and data set: $\bar{y} \pm t_{0.975,66} \cdot \sqrt{\text{MSE}/12}$.

As an example, consider the various models obtained for data set C, predicting species richness of native fish in north-eastern U.S. lakes (Table 2). Applying stepwise selection (method 1) with a P-to-enter and P-to-stay of 0.05, we obtain the three-variable model shown in the first

Table 4 Standardized mean values of mean squared prediction error (MSPE) based on cross-validation for the seven methods applied to the 12 data sets (see Table 1 for the correspondence between letters and data sets). In each row, the mean MSPEs were standardized by subtracting the mean (column 2) and dividing by the standard deviation (column 3). Methods: 1, stepwise *F* tests; 2, stepwise Akaike's information criterion (AIC); 3, stepwise Bayesian information criterion (BIC); 4, all subsets AIC; 5, all subsets BIC; 6, tree with 1-SE rule; 7, tree with minimum risk

			Standardized MSPE by method						
Data set	Mean MSPE	SD of MSPE	1	2	3	4	5	6	7
A	2.44	0.186	−0.67	−0.44	−0.67	−0.44	−0.67	1.57	1.34
B	0.420	0.0428	−0.62	−0.73	−0.30	−0.73	−0.51	1.47	1.42
C	0.357	0.0597	−0.55	−0.61	−0.55	−0.61	−0.55	1.78	1.08
D	0.291	0.0764	−0.60	−0.58	−0.57	−0.57	−0.59	1.60	1.31
E	12.3	2.67	−0.57	−0.59	−0.58	−0.60	−0.58	1.56	1.36
F	1.18	0.123	0.25	0.58	0.39	0.49	0.82	−0.46	−2.08
G	0.348	0.0146	0.21	0.68	0.49	0.79	0.71	−1.42	−1.45
H	0.240	0.139	−0.45	−0.60	−0.60	−0.60	−0.67	1.53	1.38
J	0.0761	0.0243	−0.56	−0.80	−0.40	−0.80	−0.31	1.48	1.40
K	0.283	0.0143	−0.59	−0.52	−0.45	−0.31	−0.38	2.22	0.03
L	37270	2105	−0.58	−0.79	−0.23	−0.86	−0.16	1.94	0.68
M	0.195	0.0171	−0.60	−0.55	−0.60	−0.53	−0.59	1.75	1.12

line of Table 2. If we instead use a threshold of 0.01, we obtain the same two-variable model that was produced by method 3. Finally, using a threshold of 0.10, we obtain the four-variable model produced by methods 2, 4 and 5. Rather than a liability from the rigid framework of hypothesis testing, the *F*-test significance level can be thought of as a tuning parameter that adjusts the penalty for model complexity in a way that is not possible using the AIC or BIC alone (e.g. see Sauerbrei *et al.* 2007). In this context, it seems unfair to dismiss stepwise variable selection with *F*-tests as a viable tool for model selection, just because it uses the machinery of hypothesis testing (Whittingham *et al.* 2006; Lukacs *et al.* 2007).

It is important to remember that the comparisons among variable-selection methods that are summarized here are based on predictive ability; if the goal of model building is explanation or identification of possible causal relationships, the criteria for comparing approaches could be different from those considered here (Sauerbrei *et al.* 2007).

Another caveat is that the use of real, rather than simulated, data makes it difficult to identify the relevant scope of inference for this work. The 12 data sets were identified in a decidedly non-random way, and it is possible that special or unusual features of these data had a strong influence on the results. Additional effort could be directed to finding more data, which could be helpful in identifying features of data sets that make them more or less amenable to the different model-building techniques. But it is hard to envision a method of sampling data sets that would permit generalization of results to an identifiable larger population.

Simulation, in which the 'true' model is known, would seem the only definitive way to compare model-building techniques. Investigators have simulated data in so many different ways and used such a variety of metrics for comparing methods that it is difficult to synthesize their results, although Raffalovich *et al.* (2008) make a determined attempt. In their own research, Raffalovich *et al.* (2008) evaluated the ability of several procedures to include important variables and exclude irrelevant ones. They found that stepwise regression and BIC-based approaches performed best, while AIC-based methods 'are clearly inferior and should be avoided'. Murtaugh (1998), on the other hand, found little difference in the discriminating ability of methods based on *F*-tests, the BIC and Mallows' C_p (similar to the AIC), consistent with the empirical results reported here.

The variety of models that can be obtained for individual data sets (e.g. see Table 2) and the similar predictive ability achieved by some fairly different methods of variable selection (Fig. 2b) suggest that there is no 'best' method of selecting statistical models. This conclusion is consistent with the frequently quoted assertion that 'all models are wrong but some are useful' (Box 1979). If there is no 'correct' model, there can be no best method of model building.

Single-minded promotion of one method of variable selection over another places undue emphasis on purely statistical considerations, a practice that some authors have grown weary of (Guthery *et al.* 2005; Murtaugh 2007; Chamberlain 2008). There is a wide array of approaches to variable selection, any of which can generate models worthy of consideration in a particular application. Which models

are most useful is determined not by the method by which they were obtained, but rather by their appropriateness for the task at hand.

ACKNOWLEDGEMENTS

I thank Jeannie Sifneos for guidance on the use of regression trees and for many constructive comments on an earlier version of the manuscript. I also thank three anonymous referees for detailed suggestions for clarifying and strengthening the paper, and Steve Stehman for helpful discussions.

REFERENCES

Belsley, D., Kuh, E. & Welsch, R. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York.

Box, G.E.P. (1979). Robustness in the strategy of scientific model building. In: *Robustness in Statistics* (eds Launer, R.L. & Wilkinson, G.N.). Academic Press, New York, pp. 201–236.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth & Brooks/Cole, Monterey.

Burnham, K.P. & Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edn. Springer, New York.

Chamberlain, M.J. (2008). Are we sacrificing biology for statistics? *J. Wildl. Manage.*, 72, 1057–1058.

Chatfield, C. (2002). Confessions of a pragmatic statistician. *Statistician*, 51, 1–20.

De'ath, G. & Fabricius, K.E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81, 3178–3192.

Gerritsen, J., Dietz, J.M. & Wilson, H.T., (1996). Episodic acidification of coastal plain streams: an estimation of risk to fish. *Ecol. Appl.*, 6, 438–448.

Guthery, F.S., Brennan, L.A., Peterson, M.J. & Lusk, J.J. (2005). Information theory in wildlife science: critique and viewpoint. *J. Wildl. Manage.*, 69, 457–465.

Harrell, F.E., (2001). *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer, New York.

Hobbs, N.T. & Hilborn, R. (2006). Alternatives to statistical hypothesis testing in ecology: A guide to self teaching. *Ecol. Appl.*, 16, 5–19.

Hoeting, J.A., Davis, R.A., Merton, A.A. & Thompson, S.E. (2006). Model selection for geostatistical models. *Ecol. Appl.*, 16, 87–98.

Knick, S.T. & Rotenberry, J.T. (1995). Landscape characteristics of fragmented shrubsteppe habitats and breeding passerine birds. *Conserv. Biol.*, 9, 1059–1071.

Lee, H. & Ghosh, S. (2009). Performance of information criteria for spatial models. *J. Stat. Comput. Simul.*, 79, 93–106.

Leigh, G.T., Read, A.J. & Halpin, P. (2008). Fine-scale habitat modeling of a top marine predator: do prey data improve predictive capacity. *Ecol. Appl.*, 18, 1702–1717.

Lukacs, P.M., Thompson, W.L., Kendall, W.L, Gould, W.R., Dougherty, P.F., Jr, Burnham, K.P. *et al.* (2007). Concerns regarding a call for pluralism of information theory and hypothesis testing. *J. Appl. Ecol.*, 44, 456–460.

Mazerolle, M.J. (2006). Improving data analysis in herpetology: using Akaike's information criterion (AIC) to assess the strength of biological hypotheses. *Amphib-reptil.*, 27, 169–180.

Miller, A. (2002). *Subset Selection in Regression*, 2nd edn. Chapman & Hall/CRC, Boca Raton, FL.

Mundry, R. & Nunn, C.L. (2009). Stepwise model fitting and statistical inference: turning noise into signal pollution. *Am. Nat.*, 173, 119–123.

Murtaugh, P.A. (1998). Methods of variable selection in regression modeling. *Commun. Stat. Simul. Comput.*, 27, 711–734.

Murtaugh, P.A. (2007). Simplicity and complexity in ecological data analysis. *Ecology*, 88, 56–62.

Olden, J.D. & Jackson, D.A. (2000). Torturing data for the sake of generality: How valid are our regression models? *Ecoscience*, 7, 501–510.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. Available at: http://www.R-project.org R Foundation for Statistical Computing, Vienna.

Raffalovich, L.E., Deane, G.D., Armstrong D. & Tsao, H.S. (2008). Model selection procedures in social research: Monte-Carlo simulation results. *J. Appl. Stat.*, 35, 1093–1114.

Ramsey, F. & Schafer, D. (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis*, 2nd edn. Duxbury Press, Belmont, CA.

Sauerbrei, W., Royston, P. & Binder, H. (2007). Selection of important variables and determination of functional form for continuous predictors in multivariable modeling. *Stat. Med.*, 26, 5512–5528.

Therneau, T.M. & Atkinson, E.J. (1997). An introduction to recursive partitioning using the RPART routines. *Technical Report Series No. 61*. Department of Health Science Research, Mayo Clinic, Rochester, MN.

Ward, E.J. (2008). A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. *Ecol. Modell.*, 211, 1–10.

Whittingham, M.J., Stephens, P.A., Bradbury, R.B. & Freckleton, R.P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *J. Anim. Ecol.*, 75, 1182–1189.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Description of the data sets.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

Editor, Jessica Gurevitch
Manuscript received 13 July 2009
Manuscript accepted 14 July 2009