Freshwater Biology (2006) **51,** 359–372 doi:10.1111/j.1365-2427.2005.01487.x

APPLIED ISSUES

Selecting discriminant function models for predicting the expected richness of aquatic macroinvertebrates

JOHN VAN SICKLE,* DAVID D. HUFF† AND CHARLES P. HAWKINS‡
*US Environmental Protection Agency, National Health and Environmental Effects Laboratory, Western Ecology Division, Corvallis, OR, U.S.A.
†Oregon Department of Environmental Quality, Watershed Assessment Section, Portland, OR, U.S.A.
‡Western Center for Monitoring and Assessment of Freshwater Ecosystems, Department of Aquatic, Watershed and Earth Resources, Utah State University, Logan, UT, U.S.A.

SUMMARY

1. The predictive modelling approach to bioassessment estimates the macroinvertebrate assemblage expected at a stream site if it were in a minimally disturbed reference condition. The difference between expected and observed assemblages then measures the departure of the site from reference condition.
2. Most predictive models employ site classification, followed by discriminant function (DF) modelling, to predict the expected assemblage from a suite of environmental variables. Stepwise DF analysis is normally used to choose a single subset of DF predictor variables with a high accuracy for classifying sites. An alternative is to screen all possible combinations of predictor variables, in order to identify several 'best' subsets that yield good overall performance of the predictive model.
3. We applied best-subsets DF analysis to assemblage and environmental data from 199 reference sites in Oregon, U.S.A. Two sets of 66 best DF models containing between one and 14 predictor variables (that is, having model orders from one to 14) were developed, for five-group and 11-group site classifications.
4. Resubstitution classification accuracy of the DF models increased consistently with model order, but cross-validated classification accuracy did not improve beyond seventh or eighth-order models, suggesting that the larger models were overfitted.
5. Overall predictive model performance at model training sites, measured by the root-mean-squared error of the observed/expected species richness ratio, also improved steadily with DF model order. But high-order DF models usually performed poorly at an independent set of validation sites, another sign of model overfitting.
6. Models selected by stepwise DF analysis showed evidence of overfitting and were outperformed by several of the best-subsets models.
7. The group separation strength of a DF model, as measured by Wilks' Λ, was more strongly correlated with overall predictive model performance at training sites than was DF classification accuracy.
8. Our results suggest improved strategies for developing reliable, parsimonious predictive models. We emphasise the value of independent validation data for obtaining a

Correspondence: John Van Sickle, US Environmental Protection Agency, National Health and Environmental Effects Laboratory, Western Ecology Division, 200 SW 35th St., Corvallis, OR 97333, U.S.A. E-mail: vansickle.john@epa.gov
Present address: David D. Huff, Department of Fisheries, Wildlife and Conservation Biology, 200 Hodson Hall, University of Minnesota, St Paul, MN 55108, U.S.A.

© 2005 Blackwell Publishing Ltd. No claim to original US government works

realistic picture of model performance. We also recommend assessing not just one or two, but several, candidate models based on their overall performance as well as the performance of their DF component.
9. We provide links to our free software for stepwise and best-subsets DF analysis.

Keywords: discriminant function, expected richness, model selection, model validation, *O*/*E* model, RIVPACS

Introduction

The predictive modelling approach is widely used for assessing the biological integrity of streams based on their sampled macroinvertebrate assemblages (Wright, Furse & Armitage, 1993; Parsons & Norris, 1996; Hawkins *et al*., 2000). The approach compares the assemblage observed at a site with an assemblage that would be expected if that site was in a minimally disturbed 'reference' condition. The expected assemblage is predicted by a statistical model that relates assemblages at reference sites throughout a large region to natural environmental factors. A significant difference between expected and observed assemblages then indicates that the assessed site is not in reference condition and has likely been impacted by human activities.

A variety of statistical model structures have been suggested for predicting the expected assemblage (Reynoldson *et al*., 1995; Chessman, 1999; Moss *et al*., 1999; Linke *et al*., 2005). However, the structure pioneered by the United Kingdom's River Invertebrate Prediction and Classification System (RIVPACS) model remains the most widely implemented (Moss *et al*., 1987; Clarke, Wright & Furse, 2003). In the RIVPACS structure, a numerical clustering algorithm is used to identify groups of reference sites within which sampled macroinvertebrate assemblages are most similar. A discriminant function (DF) model is then developed to predict the probability that any stream site is a member of each reference site group, based on a suite of natural environmental variables. The expected occurrence probability of each species at any site can then be computed as a weighted average of that species' occurrence frequencies in the reference site groups. The weights used for this averaging are the group membership probabilities for the site, as predicted by the DF model.

Although expected occurrence probabilities can be compared with observed occurrences on a species-by-species basis, such a comparison may involve from tens to over one hundred species. As a result, various summary indices of the two assemblages have been formulated and used in comparisons (Clarke *et al*., 1996, 2003). In this paper, we employ the widely used ratio of observed to expected species richness, *O*/*E*, where *E* is the sum of model-predicted occurrence probabilities at a site, across a specified list of species, and *O* is the total number of those species observed at the site (Clarke *et al*., 1996; Parsons & Norris, 1996; Hawkins *et al*., 2000). Hawkins & Carlisle (2001) give a step-by-step example of developing a RIVPACS-type predictive model, and Clarke *et al*. (1996, 2003) supply statistical details.

In this paper, we focus on strategies for selecting the 'best' environmental predictor variables for inclusion in the DF component of a predictive model. Variable selection strategies have received little attention in the predictive modelling literature, in contrast to the study of alternative clustering strategies to achieve optimal site classifications (Moss *et al*., 1999; Johnson, 2000). However, some general guidelines for choosing predictor variables are available. First of all, candidate predictors clearly should be known to have the potential for influencing macroinvertebrate communities, either directly or as surrogates. Secondly, predictive models attempt to explain the variability of assemblages in minimally impacted (i.e. reference) streams because of 'natural' factors. Thus, candidate predictors should represent environmental gradients that are relatively unaltered by pollution or other anthropogenic stress (Hawkins & Carlisle, 2001; Clarke *et al*., 2003). Finally, predictive model users prefer predictor variables that are simple and accurate to measure for any stream sampling location (Clarke *et al*., 2003). Having satisfied these requirements, the model developer is still likely to be left with a large number of equally viable candidate predictors. For example, viable candidates often include variables representing geographic location (latitude, longitude, ecoregion), stream topography (altitude, slope, sinuosity, valley landform, stream network structure),

catchment or stream size and catchment or streambed geology. Here, we focus on the problem of choosing 'best' DF predictors from a relatively long list of viable candidates.

It is possible to include all viable candidates in a DF model. However, this strategy can lead to an overfitted DF model, that is, a model so closely tailored to the particular data set used in model building that it performs poorly when applied to new cases (Harrell, 2001; Burnham & Anderson, 2002). Predictive model developers have often tried to avoid overfitting by seeking a small subset of variables that nevertheless provide reliable and accurate DF predictions of site group membership. Users of a parsimonious predictive model also benefit from the fact that the model's need for input data has been minimised (Moss *et al.*, 1987).

Developers of predictive models have had few tools available to help simplify their DF models. Early versions of RIVPACS retained all candidate predictors, numbering as many as 28, but those DFs (linear combinations of the predictors) that were deemed statistically non-significant were then eliminated (Wright *et al.*, 1984). This approach may have reduced the risk of overfitting, but model users would still have to measure a long list of predictor variables to assess each new stream site. Later and current versions of RIVPACS offer users several optional DF models, each of which has a somewhat different reduced set of predictors, to allow for maximum user flexibility in data acquisition (Moss *et al.*, 1987; Wright, 2000). However, the method used by RIVPACS developers to find these reduced subsets has not been articulated.

Developers of most other predictive models have relied heavily on the stepwise method to search for a single best DF model employing a minimal set of predictors (Parsons & Norris, 1996; Marchant *et al.*, 1997; Reynoldson *et al.*, 1997; Hawkins *et al.*, 2000; Hawkins & Carlisle, 2001; Joy & Death, 2002). However, stepwise DF analysis is vulnerable to the same problems as stepwise regression analysis (McKay & Campbell, 1982; Harrell, 2001; Burnham & Anderson, 2002). First of all, the stepwise method explores only a small subset of all possible DF models that could be built from a set of candidate predictors. Secondly, variable choices and stopping rules in the stepwise method are based on partial *F*-tests for inclusion of individual predictors. But because the stepwise method performs an unpredictable and large number of such tests, its nominal statements about predictor significance are seriously compromised.

Another challenge to DF variable selection is the lack of any clear linkage between the overall performance of a predictive model and the performance of its DF component. Because the DF component is used to predict group memberships, predictive classification accuracy of reference sites has, until now, been the primary basis for selecting best DF models (Moss *et al.*, 1987; Reynoldson *et al.*, 1997; Hawkins *et al.*, 2000; Joy & Death, 2002). In contrast, the overall performance of a predictive model can be directly assessed by how closely its predictions of expected richness match the observed richness at reference sites, that is, how close the *O*/*E* ratio is to unity (Moss *et al.*, 1999). If the mean value of *O*/*E* at reference sites is equal to 1.0 then a predictive model is unbiased, and a small standard deviation of *O*/*E* indicates a model with high precision (Ostermiller & Hawkins, 2004; Van Sickle *et al.*, 2005). We are unaware of any research showing that the best DF model (i.e. highest classification accuracy) will yield the best predictive model (minimum bias and maximum precision of *O*/*E*) for a given set of site groups.

In this paper we describe and implement a 'best-subsets' approach to DF variable selection. Our approach screens all possible candidate DF models for a given site classification and then calculates statistics of *O*/*E* for a large subset of the best candidates. We apply this approach to assemblage and predictor data collected throughout the state of Oregon, U.S.A. Our goal is to illustrate how the number of predictors included in DF models affects DF model performance as well as the overall performance of the predictive model. The resulting patterns help identify overfitted DF models, they show whether stepwise DF analysis can reliably identify best models, and they show the linkage between DF model performance and overall predictive model performance. Based on these results, we suggest some guidelines for developing the DF component of a predictive model. Our software for the best-subsets approach is made available free to readers.

Methods

Study area and data sources

Oregon is a large state in north-western U.S.A. that includes land area from three major ecological regions

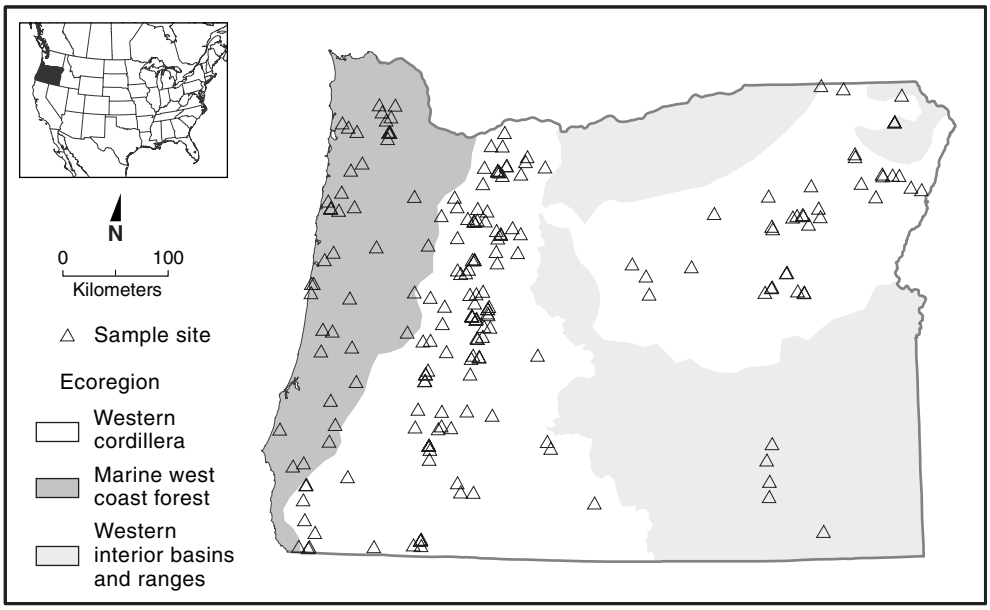

Fig. 1 Map of Oregon U.S.A., showing locations of 199 sampled reference sites.

(Fig. 1; Omernik, 1987; Commission for Environmental Cooperation, 1997). Proceeding from west to east in Fig. 1, the Marine ecoregion includes low coastal mountains dominated by conifer forests and receiving high rainfall, as well as the agricultural and urban-dominated Willamette Valley. The mountains of the Western Cordillera ecoregion are also conifer-dominated but have higher altitudes and a drier climate. The Western Interior Basin and Range ecoregion is dominated by semi-arid grasslands and high-altitude deserts.

The Oregon State Department of Environmental Quality sampled macroinvertebrate assemblages during the summer months (from June until September), 1998–2003, from first to third order streams, with stream order identified on 1 : 100 000 scale maps. Sampling locations were chosen randomly with a spatially systematic component from throughout Oregon's stream networks (Herlihy *et al.*, 2000). Additional sites were handpicked for possible use in predictive model development because they were determined to be minimally affected by human disturbance. The catchments of all random and hand-picked sites were then screened using digital maps for human disturbance factors in their drainage areas such as high road density, urban and agricultural use, active or recent logging and presence of cattle grazing. Site visits were also made to identify reach-level disturbance factors such as channel modification and land use activities within 10 m of the stream bank (roads, mining activity, buildings, etc.). Reach level and catchment assessments were combined into an overall human disturbance score for each sampling site, and all sampled sites were assigned a disturbance grade based on a five-level scale. Disturbance grades were verified, and altered if necessary, after follow-up site sampling of water chemistry and instream habitat variables (see Drake, 2004, for further details). We employed 199 sites from the two least-disturbed grade levels and distributed around Oregon (Fig. 1) as reference sites for predictive model development. We randomly selected 133 of the 199 sites to develop predictive models. We fitted clustering and DF models to these 133 sites and refer to them as model 'training' sites. We set aside the remaining 66 sites for model validation.

Macroinvertebrates were sampled in riffle habitats with kick-nets and then composited to give a single assemblage for each site. Full details of macroinvertebrate sampling methods are given by Waite *et al.* (2000) and Klemm *et al.* (2003). Composite samples were sorted and subsampled to a maximum of 500 individuals and then identified to the lowest possible taxonomic level (usually genus). A subsequent random subsampling yielded a fix-count of 300 individuals for each site in the reference data set.

Table 1 Candidate predictors for DF models

Description	Name	Source
Julian day of macroinvertebrate sampling	SAMPDAY	Calendar
Site longitude	LONGITUDE	1:24000 scale map
Site latitude	LATITUDE	1:24000 scale map
Catchment area (log-transformed)	WS_AREA	Catchments were hand-delineated from 1:24000 topographic maps
Stream gradient (square-root transformed)	GRADIENT	Estimated for sample reach from 1:24000 scale topographic maps
Indicator for Marine ecoregion	MAR_ECO	Commission for Environmental Cooperation (1997)
Indicator for Western Cordillera ecoregion	WEST_ECO	Commission for Environmental Cooperation (1997)
Mean annual precipitation at site	PRECIP	PRISM model: Daly, Neilson & Phillips (1994). See also http://www.ncgc.nrcs.usda.gov/products/datasets/climate/
Mean of long-term annual maximum temperature at site	TEMP	PRISM model: Daly *et al.* (1994). See also http://www.ncgc.nrcs.usda.gov/products/datasets/climate/
Indicator for site lithology type: sand, loose sediment, and unstable alluvium	ALLUVIUM	Walker & MacLeod (1991)
Indicator for site lithology type: igneous rock	IGN_ROCK	Walker & MacLeod (1991)
Indicator for site lithology type: sedimentary and metasedimentary rock	SED_ROCK	Walker & MacLeod (1991)
Site altitude (square-root transformed)	ALTITUDE	Digital elevation model at 30 m resolution
Stream power index (log-transformed)	STRM_POW	Product of stream gradient and square root of catchment area

We employed 14 candidate predictors for our DF models (Table 1). Except for 'Julian day', all candidate predictors were estimated from geographical information system (GIS) operations on digital maps. Our reliance on map-based predictors was dictated by our desire to build models that could estimate *O*/*E* from a macroinvertebrate sample alone, without the need for other field data. We recognised at the outset that our models might show suboptimal performance because instream site-level attributes such as stream chemistry and local habitat were not included in our list of candidates. Transformations of selected variables were employed to induce approximate normality (Table 1).

Predictive model construction

We calculated Bray–Curtis dissimilarities between all pairs of the 133 training sites, based on arcsine square-root transformed relative abundances of all 164 macroinvertebrate taxa found at one or more of these sites. Training sites were then clustered using the flexible-β algorithm with $\beta = -0.6$ (Legendre & Legendre, 1998; McCune & Grace, 2002). We pruned the cluster dendrogram to yield two different site classifications, comprised of five and 11 groups, that each contained a minimum of five sites per group with the exception of one group that had four sites (Ostermiller & Hawkins, 2004). Separate follow-up analyses were then conducted for each of the two classifications.

We fitted DF models to training site data to predict site membership in the groups of a single classification, using the predictors in Table 1. We explored only linear DF models that assume multivariate normality of predictors with a common covariance matrix for all groups (Legendre & Legendre, 1998; Clarke *et al.*, 2003). We retained the full set of canonical DFs in each candidate model and employed group sizes as prior weights when predicting posterior group membership probabilities for each site (Clarke *et al.*, 2003).

Choice and evaluation of candidate DF models

We wrote a computer program that implements a 'best subsets' algorithm to explore multiple candidate DF models for a given site classification (McKay & Campbell, 1982). The program first evaluates how strongly the site groups are separated by each of the (2^p-1) possible models that can be constructed from a set of p candidate predictors. Group separation is measured by Wilks' Λ, defined as the ratio between the within-class and total generalised variances of the predictors in a model (Johnson & Wichern, 1988; Rencher, 2002). A small value of Wilk's Λ denotes strong group separation. We chose Wilks' Λ because

this statistic is the basis for the entry and removal of predictors in stepwise DF analysis (Rencher, 2002).

Our program ranks the set of all possible models by their values of Wilks' Λ, separately within each model order $k = 1, 2, \ldots p$, where a model's order is defined as the number of predictor variables it includes. Program users may specify a number of best models of each order to be retained for further evaluation. For the example in this paper, we retained the five best DF models each from orders 1, 2, …13, and also the single model containing all 14 predictors, for a total of 66 candidate models. For each of these best models, the program then calculates statistics of *O*/*E* separately for training and validation sites. DF classification accuracies of training sites, based on leave-one-out cross-validation as well as resubstitution, are also calculated for each of the best models. In this paper we illustrate patterns of model performance by plotting different performance measures against model order, and against each other, with each of the 66 models contributing one point on each plot. Separate sets of 66 best models were developed for predictive models based on the five-group and the 11-group site classifications.

We also used both the stepwise forward and stepwise backward DF analyses to select best models, for comparison with the larger subsets of best models provided by the best-subsets approach. The stepwise algorithms were terminated when all predictors in the model had nominal *P*-values <0.05, and all excluded variables had $P > 0.05$, based on the partial Wilks' Λ statistic (Rencher, 2002).

Our programs for best-subsets and stepwise DF analysis are written as function scripts for the R language (Ihaka & Gentleman, 1996), and they can be obtained free from the senior author upon request. The R language itself is also available free at http://www.r-project.org/.

Calculating O/E and evaluating overall model performance

The RIVPACS model calculates *O*/*E* for a site based on the full list of reference species, as well as for a subset of 'common' species, defined as those having model-predicted occurrence probabilities ≥0.5 for that site (Wright, 2000). Here, we base *O*/*E* on only the common species, as advocated by Parsons & Norris (1996) and Hawkins *et al.* (2000).

The bias of a predictive model is given by $B = [\text{Mean}(O/E) - 1.0]$, where Mean (*O*/*E*) denotes the average *O*/*E* over a set of reference sites. As predictive models are fitted to training sites, their bias is close to zero for those sites. In this case, the standard deviation of *O*/*E*, by itself, characterises the magnitude of prediction errors and hence overall model performance. However, model predictions at validation sites may have a bias that is not negligible. As a result, we used the root mean squared error (RMSE) of *O*/*E* to combine the bias and variability of prediction errors into a single overall measure of model performance. The RMSE is given by:

$$\text{RMSE}(O/E) = \sqrt{\frac{1}{n}\sum\left(\frac{O}{E} - 1\right)^2} = \sqrt{B^2 + SD^2}, \quad (1)$$

where the sum is across a collection of n reference sites and *SD* is the standard deviation of *O*/*E* for those sites (Rice, 1988).

Upper and lower baselines for model performance

The RMSE measures the relative overall performance of alternative models. However, as Guthery *et al.* (2005) point out, a sole reliance on relative statistics allows model selection algorithms to '… identify ''best'' or ''plausible'' models in a set of outrageously bad models'. For this reason, we used null-model predictions of *O*/*E* and the replicate sampling standard deviation to provide upper and lower baselines, respectively, for the model performance measured by RMSE(*O*/*E*) (Van Sickle *et al.*, 2005).

The null-model occurrence probability for any species is estimated by the proportion of training sites at which that species was observed to occur (Van Sickle *et al.*, 2005). The null model specifies its own fixed set of common species to be those species having null-model occurrence probabilities ≥0.5. Null-model expected richness (E_N) is the sum, over this set of common species, of their null-model occurrence probabilities. This single value E_N is the expected richness predicted by the null model for all sites, whether they be a training, validation or 'test' sites. In the null-model richness ratio, denoted by O_N/E_N, observed richness (O_N) is then defined as the number of null-model common species observed at the assessed site.

A null model does not explain any of the variability in species richness across sites. As a result, the standard deviation of O_N/E_N over a set of training

sites serves as an upper baseline on the standard deviation of O/E from any predictive model, for that same set of sites (Van Sickle *et al.*, 2005). The null model and any predictive model have exactly zero and nearly zero bias, respectively, for training sites. Thus, for the set of training sites, eqn 1 shows that RMSE is either exactly or nearly equal to the standard deviation of O/E, so that RMSE(O_N/E_N) also serves as an upper baseline for RMSE(O/E). In short, the overall performance of any predictive model at training sites can be judged by how far it reduces RMSE(O/E) below the upper baseline of RMSE(O_N/E_N).

When a predictive model is applied to set-aside validation sites, one expects to see somewhat reduced model precision as well as increased bias, relative to its performance at training sites. However, we have frequently observed a decrease in the standard deviation of O/E at validation sites [see for example, Table 1 of Hawkins (in press)]. If taken at face value, such a decrease indicates greater model precision at new sites, which is highly unlikely. This suggests that sampling variation or other as-yet unidentified factors contribute substantially to differences in O/E variability between training and validation sites. As a result, a direct comparison of standard deviations or RMSE's between training and validation sites does not appear to give a reliable measure of how a model's performance alters when it is applied to new sites. For this reason, we use RMSE(O_N/E_N), with O_N values from validation sites, as a separate baseline for assessing predictive model performance at validation sites.

In addition to the null-model upper baseline, we used the replicate sampling standard deviation (SD_R) to provide a lower baseline for model performance at training sites (Van Sickle *et al.*, 2005). Given the 'true' occurrence probabilities for all taxa at a site, SD_R measures the irreducible level of variability in O/E across replicate observed assemblages that is due solely to presence/absence realisations of those probabilities. Because SD_R assumes that true occurrence probabilities are known exactly, Van Sickle *et al.* (2005) argued that no predictive model can be expected to yield an SD for O/E at training sites that is less than SD_R, hence the utility of SD_R as a lower baseline. The derivation of SD_R assumes zero bias in O/E at training sites, so that SD_R also serves as a lower baseline for RMSE(O/E). As Van Sickle *et al.* (2005) show, although SD_R is based on true probabilities, it can be reasonably well estimated for training sites from their model-predicted taxon occurrence probabilities. An analogous estimate of SD_R could be made for validation sites, but the accuracy of such estimates are not yet known and we do not report them here.

Relative importance of predictor variables

Although DF models are intended purely for prediction, assessing the relative importance of candidate predictors in the model(s) helps identify the dominant natural factors that control macroinvertebrate assemblages at reference sites. Hawkins *et al.* (2000) used the partial F-statistics of predictors in their final, stepwise-selected DF model to rank predictor importance. Under our best-subsets approach, it is more appropriate to define the importance of a predictor in terms of its 'average' role in multiple, nearly equal models (Kruskal, 1987; Budescu, 1993; Burnham & Anderson, 2002). Specifically, we measure a variable's importance by the proportion of our 66 best models that include that variable.

Results

Classification accuracy of DF models

The leave-one-out cross-validation approach gives a more realistic estimate of DF model accuracy in classifying new sites than does the resubstitution method (Clarke *et al.*, 1996; Reynoldson *et al.*, 1997; Rencher, 2002). For our best models, cross-validation accuracies stopped increasing with model order, for five-group models at about seventh order, and for 11-group models at about eighth order (Fig. 2). Models larger than these showed no further improvement in classification accuracy. In contrast, resubstitution accuracy increased steadily for models up to order 13, as DF models became increasingly better 'tuned' to the training data. Thus, models larger than the order (seventh or eighth) at which cross-validation and resubstitution trajectories diverge are likely being overfitted (Fig. 2).

Our sets of 66 best models included the best models found by the forward and backward stepwise methods. For the five-group case the two algorithms selected different eighth-order models, but they selected the same 10th order model for the 11-group case. Fig. 2 suggests that all of our stepwise-selected models are probably somewhat overfitted. In addi-

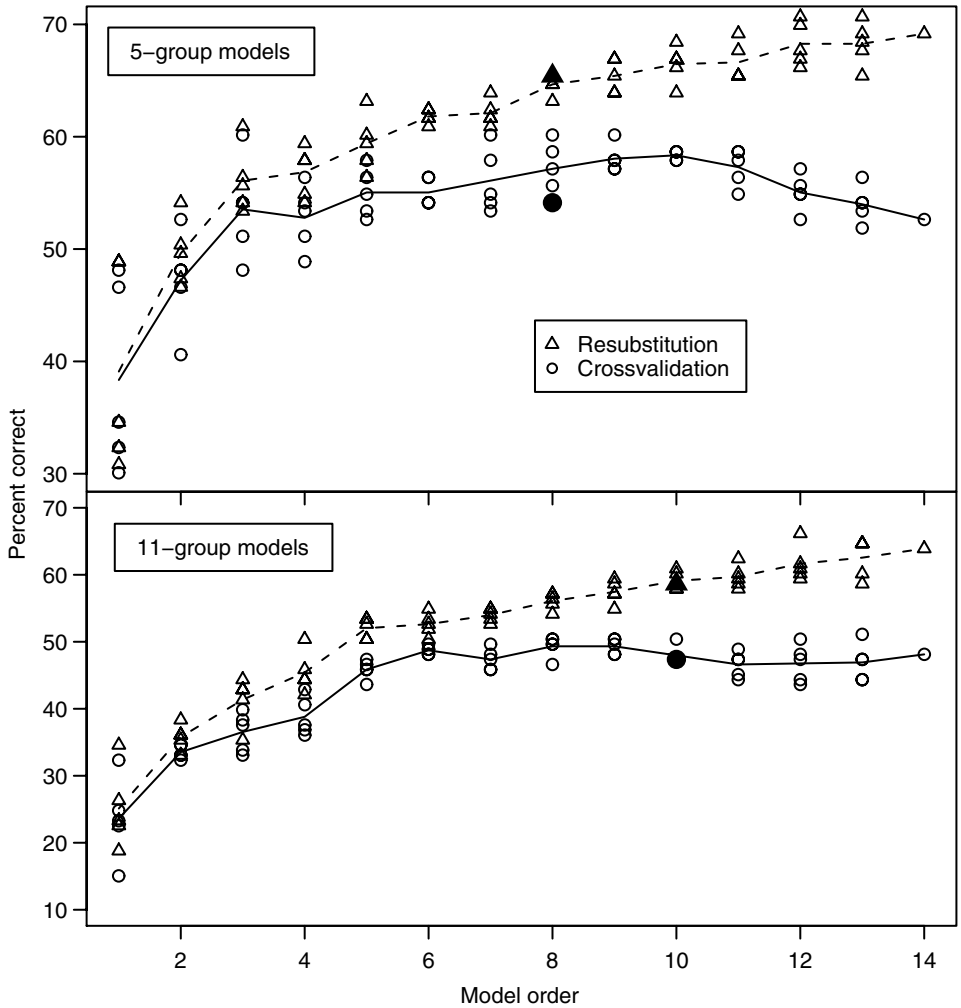

Fig. 2 Percentage of 133 training sites correctly classified using resubstitution and also leave-one-out cross-validation. Classification accuracies of 66 best DF models are given for the five-group classification and of another 66 models for the 11-group classification. Larger, solid symbols denote best models selected by stepwise DF analysis. Lines connect the mean cross-validation (solid lines) and resubstitution (dashed lines) accuracy for each model order.

tion, some of the best models identified by the best-subsets method had higher cross-validation accuracies than did the best stepwise-selected models, for both the five-group and 11-group cases (Fig. 2).

Overall predictive model performance

For five-group predictive models, overall precision and accuracy at training sites increased rapidly [that is, RMSE(*O*/*E*) decreased] with the order of the best DF models, up to about eighth order (Fig. 3a). Larger DF models offering little additional improvement. At training sites, the RMSE's from DF models larger than about fourth order were below the null-model baseline, and between about 30% to 50% of the distance to the replicate-sampling lower baseline (Fig. 3a). In other words, these models could account for between 30% and 50% of the standard deviation explainable by any predictive model. For the 11-group models, the decline in RMSE(*O*/*E*) at training sites with model order was not as rapid as for the five-group models, and the decline in RMSE continued up to about the 12th order (Fig. 3b). The minimum RMSE achieved by the 11-group models was slightly higher than that achieved by five-group models.

Overall predictive model performance at the 66 validation sites was markedly inferior to training site performance. RMSE(*O*/*E*) for five-group models de-

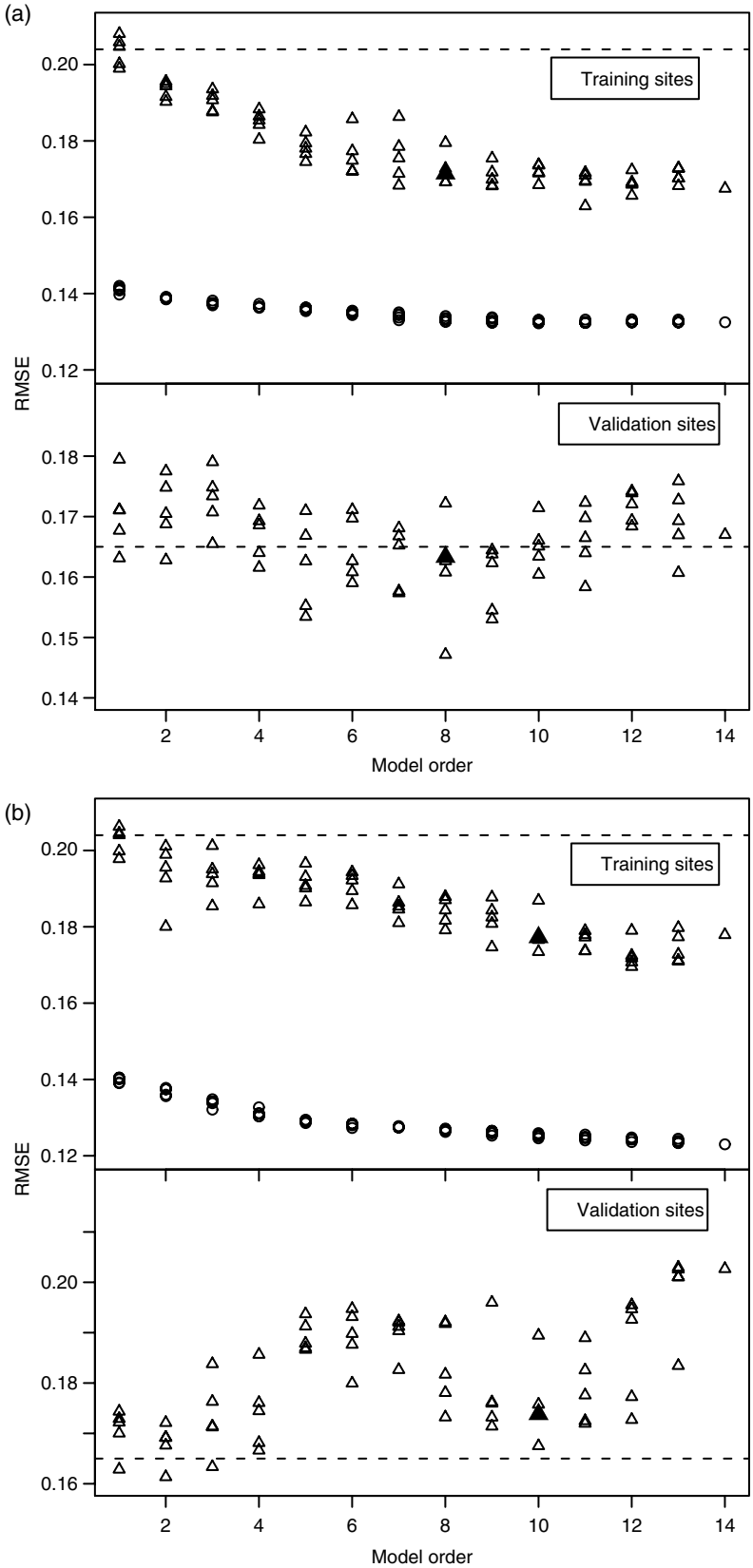

Fig. 3 Root mean squared errors (RMSE) of *O*/*E* (triangles), from predictive models based on (a) 66 best DF models and the five-group classification and (b) another 66 DF models and the 11-group classification. Larger, solid triangles denote best models selected by stepwise DF analysis. Dashed lines denote RMSE of the null model. Open circles denote replicate sampling standard deviation (SD_R) for each model.

clined gradually up to about the eighth order, but then began to increase again, a sign that these larger models were overfitted (Fig. 3a). Note the large number of five-group models having lower accuracy and precision (higher RMSE) than the null model at validation sites. Performance of 11-group models was inferior to five-group models at validation sites (Fig. 3b). RMSE actually increased with increasing DF model order, indicating that model performance on set-aside sites steadily declined as model complexity increased. RMSE's for all but a few of the lowest-order 11-group models were larger than that of the null model (Fig. 3b). This suggests that, from the standpoint of independent validation data, nearly all of the 11-group models were overfitted.

At both validation and training sites, and for both the five-group and 11-group models, there were several DF models among the best-model subsets that yielded smaller RMSE's than did the best model selected by the stepwise forward algorithm (Fig. 3).

Relations between DF model performance and overall predictive model performance

For the five-group DF models, both Wilks' Λ and cross-validated classification accuracy were clearly related to overall predictive model performance at training sites, as measured by the RMSE(*O*/*E*) (Fig. 4, upper panels). The correlation with RMSE was stronger for Wilks' Λ than for classification accuracy, for both the five-group and the 11-group sets of models (Fig. 4; Table 2). But for the best-performing models (highest classification accuracy, lowest Wilks' Λ and lowest RMSE), there was considerable variation

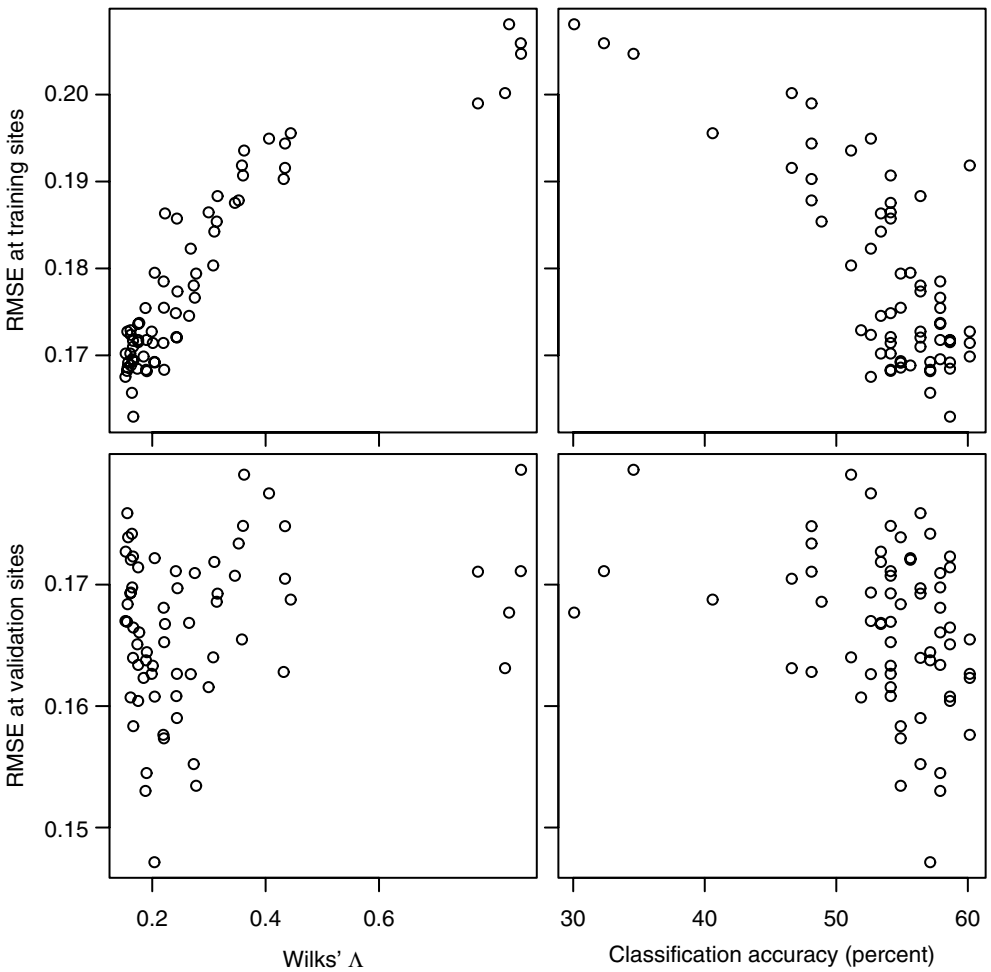

Fig. 4 RMSE of *O*/*E* at training sites (upper panels) and validation sites (lower panels), versus the performance of 66 DF models for five site groups at training sites, as measured by Wilks' Λ (left panels) and cross-validated classification accuracy (right panels).

Table 2 Spearman rank correlation coefficients between RMSE's of *O*/*E* at training and validation sites, and measures of DF model performance at training sites, across 66 best DF models based on five site groups and another 66 models based on 11 site groups

Site type	Site groups	Wilks' Λ	Cross-validated classification accuracy
Training	5	0.85	−0.57
Training	11	0.90	−0.43
Validation	5	0.11	−0.33
Validation	11	−0.51	0.43

in RMSE across numerous DF models having about the same levels of classification accuracy or Wilks' Λ (Fig. 4). This same pattern was observed in the 11-group models (plot not shown).

Predictive model performance at validation sites was only weakly related to DF performance at training sites, for the set of five-group DF models (Fig. 4, lower panels; Table 2). The best-performing DF models, as measured by either classification accuracy or Wilks' Λ, yielded RMSE's at validation sites that spanned nearly the entire RMSE range. For the 11-group case, better-performing DF models (higher classification accuracy and lower Wilks' Λ) tended to yield degraded predictive models (higher RMSE's) at validation sites, a result predictable from Fig. 3b (Table 2).

Predictor importance

Site elevation was the single most prevalent predictor, appearing in every best five-group and 11-group model that contained two or more predictors (Fig. 5). Julian day of sampling was also included in about 80% of our best models. At the other extreme, site latitude and ecoregion location indicators were each included in fewer than 25% of best models. Catchment area and stream power appeared in over 70% of the best five-group models, but these two predictors were less prevent in the 11-group models. Longitude and the igneous rock indicator appeared in even higher percentages of 11-group models than did sampling day.

Discussion

Based on Fig. 3, our best predictive models employed five groups and those sixth to eighth-order DF models that yielded low RMSE's at both training and valid-

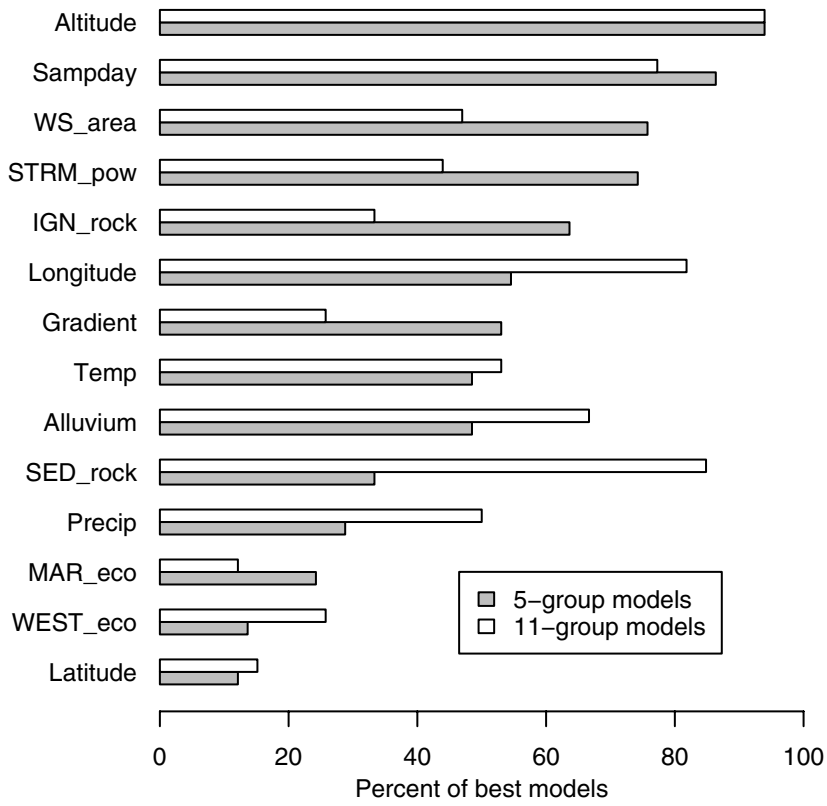

Fig. 5 Percentage of models in which each candidate predictor variable was included, for 66 best five-group models and another 66 best 11-group models. Predictors are defined in Table 1.

ation sites. These models reduced the RMSE of *O*/*E* at training sites to only about 0.17, relative to the null-model value of 0.20. This is not a large reduction, but it does represent about 50% of the RMSE that could be explained by any predictive model, relative to the SD_R lower baseline.

We emphasise that our overall model performance statistics are based solely on the *O*/*E* comparison of assemblage species richness. Performance statistics for other indices comparing the observed and expected assemblages, such as tolerance-weighted richness measures or a species-by-species matching of predicted occurrence probabilities with observed occurrences (Clarke *et al*., 1996), may give a different view of overall model quality. Our best-subsets software can be adapted by users to report alternative performance measures.

For the Oregon data, our best-subsets approach revealed that, for a given set of site groups, several DF models fit the data nearly equally well. In such cases, model performance alone is not a sensible basis for choosing a single final model (McKay & Campbell, 1982). One can winnow a collection of best-performing models by examining the predictor variables included in each DF component. Individual predictors may be preferred because of their ease and accuracy of measurement. In addition, certain subsets of predictors may be desirable because they rarely lead to predictive extrapolation, as evaluated by a chi-squared test of the distance in predictor space between a new site and the training data sites (Moss *et al*., 1987; Clarke *et al*., 2003). It is not necessary for a winnowing process to ultimately produce a single 'best' predictive model. For example, recent versions of RIVPACS provide users with a system of alternative models having nearly equal performance (Moss *et al*., 1987; Wright, 2000). As another approach, one could estimate *O*/*E* from several best models, then average the estimates, weighted perhaps by model quality (Burnham & Anderson, 2002).

The single DF model selected by the stepwise method for five groups was among our chosen sixth to eighth-order models, but it did not give the best overall performance. In addition, Figs 2 & 3 suggest that the stepwise-selected model may be overfitted. We strongly discourage modellers from relying solely on stepwise DF analysis to select a single best model. If the stepwise method is the only automated variable selection tool accessible to modellers, then we recommend manually fitting and evaluating additional DF models that are 'near' to the stepwise-selected model, as well as a few lower-order models that were visited during the forward stepping sequence. Predictors that enter a DF model at early steps in the forward algorithm tend to have greater predictive value and can be included in all manually selected models. For example, site altitude was selected at the first step for our five-group and 11-group models, and it was also the most prevalent predictor among best models for both groupings.

Our choice of prevalence to measure predictor importance in multiple best models may have been reasonable, but, like other such measures, it was also somewhat *ad hoc* (Kruskal, 1987; Budescu, 1993). Our measure could be upgraded by calculating the weighted prevalence for each predictor, using weights derived from a measure of model quality (Burnham & Anderson, 2002). Further research might identify a single measure of predictive model quality that is analogous to Burnham & Anderson's (2002) use of the Akaike information criterion (AIC), and it also might suggest a scaling of quality weights that is analogous to that of Burnham & Anderson (2002).

For our Oregon example, a measure of group separation for DF predictors (Wilks' Λ) was more closely correlated with RMSE of *O*/*E* than was DF classification accuracy. A weak relationship between classification accuracy and overall predictive model performance might be expected because a predictive model employs predicted group membership probabilities only to weight species occurrence rates, rather than to assign sites to individual groups as done by the BEAST model (Reynoldson *et al*., 1997). Although classification accuracy may be more easily interpreted, we recommend that Wilks' Λ be used for selecting the DF models most likely to yield high predictive model performance. The high correlations between Wilks' Λ and RMSE at training sites (Table 2) support the choice of Wilks' Λ to screen the set of all possible models in our best-subsets algorithm. However, as Fig. 4 suggests, Wilks' Λ is not a perfect indicator of overall predictive model performance, especially for the better DF models. We encourage model developers to choose among alternative DF models based on the statistics of *O*/*E* rather than on DF model performance alone.

Finally, Fig. 3 shows that many of our models performed quite well on training data, but then

performed poorly, even worse than the null model, on data from an independent set of validation sites. Fig. 3b also suggests that model performance at validation sites can be degraded by using too many site groups as well as too many DF predictors. Our validation results indicate that overfitting is an issue for the entire structure of a predictive model, not just for its DF component.

Sample size permitting, we recommend that predictive model developers set aside at least 20 reference samples or sites to provide an independent, realistic assessment of overall model quality. For many model development projects, reference site samples will be too precious for use only as validation data (Harrell, 2001). In these cases, cross-validation of the entire model-building process might be a more efficient, but still realistic, way to estimate overall predictive model performance.

Acknowledgments

We thank Lester Yuan, Phil Larsen, Bob Ozretich and two anonymous referees for valuable suggestions. This research was funded by the US Environmental Protection Agency. The manuscript has been subjected to the agency's peer and administrative review and has been approved for publication as an EPA document. CPH was supported by EPA Science To Achieve Results (STAR) grants R-82863701 and R-82863701. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

References

Budescu D.V. (1993) Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, **114**, 542–551.

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

Chessman B.C. (1999) Predicting the macroinvertebrate faunas of rivers by multiple regression of biological and environmental differences. *Freshwater Biology*, **41**, 747–757.

Clarke R.T., Wright J.F. & Furse M.T. (2003) RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. *Ecological Modelling*, **160**, 219–233.

Clarke R.T., Furse M.T., Wright J.F. & Moss D. (1996) Derivation of a biological quality index for river sites: comparison of the observed with the expected fauna. *Journal of Applied Statistics*, **23**, 311–332.

Commission for Environmental Cooperation (1997) *Ecological Regions of North America: Toward a Common Perspective*. Montreal, Quebec, Canada, 71 pp.

Daly C., Neilson R.P. & Phillips D.L. (1994) A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *Journal of Applied Meteorology*, **33**, 140–158.

Drake D.L. (2004) *Selecting Reference Condition Sites – an Approach for Biological Criteria and Watershed Assessment*. Technical Report WAS04–002. Oregon Dept. of Environmental Quality, Watershed Assessment Section, Portland, OR.

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

Harrell Jr F.E. (2001) *Regression Modeling Strategies*. Springer-Verlag, New York.

Hawkins C.P. (in press) Quantifying biological integrity by taxonomic completeness: evaluation of a potential indicator for use in regional- and global-scale assessments. *Ecological Applications*.

Hawkins C.P. & Carlisle D.M. (2001) Use of predictive models for assessing the biological integrity of wetlands and other aquatic habitats. In: *Bioassessment and Management of North American Freshwater Wetlands* (Eds D.P. Batzer & S.A. Wissinger), pp. 59–83. John Wiley and Sons, New York, NY.

Hawkins C.P., Norris R.H., Hogue J.N. & Feminella J.W. (2000) Development and evaluation of predictive models for measuring the biological integrity of streams. *Ecological Applications*, **10**, 1456–1477.

Herlihy A.T., Larsen D.P., Paulsen S.G., Urquhart N.S. & Rosenbaum B.J. (2000) Designing a spatially balanced, randomized site selection process for regional stream surveys: the EMAP Mid-Atlantic pilot study. *Environmental Monitoring and Assessment*, **63**, 95–113.

Ihaka R. & Gentleman R. (1996) R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 239–314.

Johnson R.A. & Wichern D.W. (1988) *Applied Multivariate Statistical Analysis*, 2nd edn. Prentice-Hall, Englewood Cliffs, NJ, U.S.A.

Johnson R.K. (2000) RIVPACS and alternative statistical modeling techniques: accuracy and soundness of principles. In: *Assessing the Biological Quality of Fresh Waters: RIVPACS and Other Techniques* (Eds J.F. Wright, D.W. Sutcliffe & M.T. Furse), pp. 323–332. Freshwater Biological Association, Ambleside, England, U.K.

Joy M.K. & Death R.G. (2002) Predictive modeling of freshwater fish as a biomonitoring tool in New Zealand. *Freshwater Biology*, **47**, 2261–2275.

Klemm D.J., Blocksom K.A., Fulk F.A. ***et al.*** **(2003)** Development and evaluation of a macroinvertebrate biotic integrity index (MBII) for regionally assessing Mid-Atlantic Highlands streams. *Environmental Management*, **31**, 656–669.

Kruskal W. (1987) Relative importance by averaging over orderings. *American Statistician*, **41**, 6–10.

Legendre P. & Legendre L. (1998) *Numerical Ecology*, 2nd edn. Elsevier, Amsterdam.

Linke S., Norris R.H., Faith D.P. & Stockwell D. (2005) ANNA: a new prediction method for bioassessment programs. *Freshwater Biology*, **50**, 147–158.

Marchant R., Hirst A., Norris R.H., Butcher R., Metzeling L. & Tiller D. (1997) Classification and prediction of macroinvertebrate assemblages from running waters in Victoria, Australia. *Journal of the North American Benthological Society*, **16**, 664–681.

McCune B. & Grace J.B. (2002) *Analysis of Ecological Communities*. MjM Software Design, Gleneden Beach, OR, U.S.A.

McKay R.J. & Campbell N.A. (1982) Variable selection techniques in discriminant analysis, Part 1 (description) and Part 2 (allocation). *British Journal of Mathematical and Statistical Psychology*, **35**, 1–41.

Moss D., Furse M.T., Wright J.F. & Armitage P.D. (1987) The prediction of the macro-invertebrate fauna of unpolluted running-water sites in Great Britain using environmental data. *Freshwater Biology*, **17**, 41–52.

Moss D., Wright J.F., Furse M.T. & Clarke R.T. (1999) A comparison of alternative techniques for prediction of the fauna of running-water sites in Great Britain. *Freshwater Biology*, **41**, 167–181.

Omernik J.M. (1987) Ecoregions of the conterminous United States. *Annals of the Association of American Geographers*, **77**, 118–125.

Ostermiller J.D. & Hawkins C.P. (2004) Effects of sampling error on bioassessments of stream ecosystems: application to RIVPACS-type models. *Journal of the North American Benthological Society*, **23**, 363–382.

Parsons M. & Norris R.H. (1996) The effect of habitat-specific sampling on biological assessment of water quality using a predictive model. *Freshwater Biology*, **36**, 419–434.

Rencher A.C. (2002) *Methods of Multivariate Analysis*, 2nd edn. John Wiley and Sons, New York.

Reynoldson T.B., Bailey R.C., Day K.E. & Norris R.H. (1995) Biological guidelines for freshwater sediment based on BEnthic Assessment of SedimenT (the BEAST) using a multivariate approach for predicting biological state. *Australian Journal of Ecology*, **20**, 198–219.

Reynoldson T.B., Norris R.H., Resh V.H., Day K.E. & Rosenberg D.M. (1997) The reference condition: a comparison of multimetric and multivariate approaches to assess water-quality impairment using benthic macroinvertebrates. *Journal of the North American Benthological Society*, **16**, 833–852.

Rice J.A. (1988) *Mathematical Statistics and Data Analysis*. Wadsworth & Brooks-Cole, Pacific Grove, CA, U.S.A.

Van Sickle J., Hawkins C.P., Larsen D.P. & Herlihy A.H. (2005) A null model for the expected macroinvertebrate assemblage in streams. *Journal of the North American Benthological Society*, **24**, 178–191.

Waite I.R., Herlihy A.T., Larsen D.P. & Klemm D.J. (2000) Comparing strengths of geographic and nongeographic classifications of stream benthic macroinvertebrates in the Mid-Atlantic Highlands, USA. *Journal of the North American Benthological Society*, **19**, 429–441.

Walker G.W. & MacLeod N.S. (1991) *Geologic Map of Oregon*. scale 1:500,000. United States Geological Survey, Reston, VA.

Wright J.F. (2000) An introduction to RIVPACS. In: *Assessing the Biological Quality of Fresh Waters: RIVPACS and Other Techniques* (Eds J.F. Wright, D.W. Sutcliffe & M.T. Furse), pp. 1–23. Freshwater Biological Association, Ambleside, England, U.K.

Wright J.F., Furse M.T. & Armitage P.D. (1993) RIVPACS: a technique for evaluating the biological water quality of rivers in the UK. *European Water Pollution Control*, **3**, 15–25.

Wright J.F., Moss D., Furse M.T. & Armitage P.D. (1984) A preliminary classification of running-water sites in Great Britain based on macroinvertebrate species and the prediction of community type using environmental data. *Freshwater Biology*, **14**, 211–256.

(Manuscript accepted 19 October 2005)