A Tiered Method for Discriminant Function Analysis Models for the Reference Condition Approach: Model Performance and Assessment

Authors: Reynoldson, Trefor B., Strachan, Stephanie, and Bailey, John L.

Source: Freshwater Science, 33(4) : 1238-1248

Published By: Society for Freshwater Science

URL: https://doi.org/10.1086/678947

A tiered method for discriminant function analysis models for the Reference Condition Approach: model performance and assessment

Trefor B. Reynoldson[1,4], **Stephanie Strachan**[2,5], **and John L. Bailey**[3,6]

[1]Institute for Applied Ecology, University of Canberra, Australian Capital Territory 2601, Australia
[2]Environment Canada, Water Science and Technology Directorate, Water Quality Monitoring and Surveillance Division, 201–401 Burrard Street, Vancouver, British Columbia, Canada V6C 3S5
[3]Ontario Ministry of Environment and Climate Change, Cooperative Freshwater Ecology Unit, Laurentian University, Sudbury, Ontario, Canada P3E 2C6

Abstract: Reference Condition Approach (RCA) predictive models are used to assess a test site against reference sites probabilistically matched based on habitat. These models are the basis of several major national stream bioassessment programs in the UK, Australia, and Canada. In the usual approach to developing predictive models, discriminant function analysis (DFA) is used to assign a test site to a group of matched reference sites. These groups typically are established by classification of a macroinvertebrate assemblage and matched to the habitat attributes in a single-step DFA model. We examined an alternative to standard DFA in which a series of tiered models are used. This tiered method constructs a model for the 1st division in a hierarchical classification, and then develops models for each further step in the hierarchical classification. We examined the method with 3 training and validation data sets. Validation data consisted of data from reference sites and those same sites after they underwent simulated impairment. We compared the tiered approach to the standard approach based on prediction accuracy and Type 1 and Type 2 error rates for each data set. The tiered DFA models were similar to or slightly better than the standard single-step DFA models in correctly matching validation sites to reference groups, but this improvement in accuracy did not necessarily translate into improved bioassessment error rates.

Key words: Reference Condition Approach, BEAST, AUSRIVAS, CABIN, simulated impairment, bioassessment, predictive modeling, tiered DFA

Reference Condition Approach (RCA) predictive models are a key component of several major national stream bioassessment/monitoring programs. First developed and applied in the UK as part of the River Invertebrate Prediction and Assessment Classification System (RIVPACS; Wright et al. 1984), these models also underpin national programs in Australia (Australian River Assessment System [AUSRIVAS]; Simpson and Norris 2000) and Canada (Canadian Aquatic Biomonitoring Network [CABIN]; Reynoldson et al. 1999) and have been tested and applied in Spain (Pardo et al. 2014), Portugal (Feio et al. 2006), and Scandinavia (Johnson 2003). The basis of these predictive models is the probabilistic relationship between biological assemblages—usually benthic macroinvertebrates—and environmental variables at reference sites. These models are used to match a test site to a collection of reference sites, and assessment is made by comparing the test-site assemblage with assemblages from the matched reference sites.

The most common method for building RCA models for stream assessment has 2 main steps. First, reference sites from a study area are classified into groups based on their biological assemblages. These classifications have the underlying assumption that the groups represent different natural invertebrate assemblages within a geographic region. The 2nd step is to develop a quantitative relationship between a set of predictor habitat variables and the biological classification. Discriminant function analysis (DFA) is most frequently used in this step. DFA identifies a set of

E-mail addresses: [4] Present address: Canada Creek, Nova Scotia, Canada B0P 1V0, trefor.reynoldson@gmail.com; [5]stephanie.strachan@ec.gc.ca; [6]jbailey@laurentian.ca

DOI: 10.1086/678947. Received 14 April 2013; Accepted 23 June 2014; Published online 7 October 2014.
Freshwater Science. 2014. 33(4):1238–1248. © 2014 by The Society for Freshwater Science.

habitat variables that show the strongest relationship with the classification created from the biological assemblage. DFA then establishes a set of coefficients for the selected set of habitat variables that allows an exposed test site to be assigned a probability of belonging to each of the biological groups created by the classification.

Once a model has been developed, it is used to assign a test site to a reference group for the bioassessment step. The bioassessment compares the test-site assemblage to the reference-site assemblage to determine whether the test site is in reference condition. Two methods are commonly used for this comparison. The RIVPACS/AUSRIVAS method uses the weighted probability that a test site belongs to a group to determine the likelihood of a taxon's occurrence and expected taxonomic richness (Wright et al. 1984). The Benthic Assessment of Sediment method (BEAST) (Reynoldson et al. 1995) compares the position of the test site in ordination space to the distribution of reference sites in the group to which the test site has the highest probability of belonging.

The nature of classification of the invertebrate assemblages in the RCA is hierarchical (Fig. 1A, B). The entire collection of reference sites is split further and further until a classification is accepted. The decision regarding what represents a final classification is partially subjective, but considers several attributes of the classification including: the number of sites in a group, e.g., Reynoldson and Wright (2000) suggested a minimum group size of 10 sites; among- and within-group dissimilarities and similarities; the distribution of sites in ordination space, and the accuracy and performance of preliminary DFA classification. These attributes are all considered in an iterative procedure for determining the optimal classification and models. Once a final set of groups is established, DFA is used to determine the subset of variables that best classify this grouping. Six groups are shown in the classification in Fig. 1A. Use of the standard DFA approach would result in a single model that assigns sites to the 6 groups created by the classification step (Fig. 1A). However, as RCA predictive modeling has evolved, the ability of DFA to assign sites to the correct biological group has become more problematic. The accuracy of the prediction declines, and the number of predictor variables required tends to increase as the number of sites to be assigned increases. This problem is illustrated in data sets from the UK and North America (Table 1), and methods to improve the prediction accuracy of these models are under consideration.

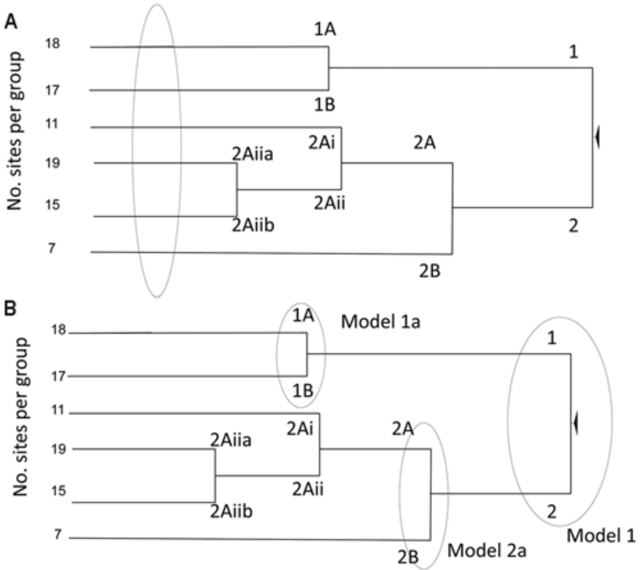

Figure 1. Hypothetical classification of a reference data set showing 6 reference groups and 2 modeling approaches A.—A single model discriminating groups in a one-step discriminant function analysis (DFA). B.—Tiered DFA models for different levels in the classification.

Table 1. Ability of standard discriminant function analysis models of varying complexity to correctly predict sites (model accuracy). Data for the River Invertebrate Prediction and Classification System (RIVPACS) model from the UK are from Reynoldson and Wright (2000).

Model	Sites	Predictors used	Model accuracy (%)
Yukon 1	90	9	55
Yukon 4	286	14	46
Fraser 1	127	10	63
Fraser 4	389	17	49
RIVPACS UK 1	268	11	66
RIVPACS UK 4	614	12	52

We described a new approach to DFA based on the use of a set of tiered models (Fig. 1B) that attempts to partition and exclude superfluous variation in the habitat. We used 2 measures of model performance to test whether this method is an improvement over the standard single-DFA model. First, we tested the accuracy of the model based on a training data set, and second, we tested the error in the assessment. We used 3 data sets from 2 habitat types and 2 continents to test whether the performance of the 2 models was in general agreement and whether attributes of a data set affected the model performance.

METHODS

DFA identifies the best subset of candidate predictor variables (in this case, environmental variables) to use to separate groups of objects (in this case, the groups of sites formed by the invertebrate assemblage). In the standard approach, all the groups are assigned in a single-DFA step. We used the new tiered-DFA approach to assign groups hierarchically.

We conducted all DFAs in SYSTAT (version 11.0; Systat, Chicago, Illinois). We used forward-stepwise analysis with the *p*-value for entry and removal set at 0.15 and the membership priors set as equal. After the initial forward-stepwise analysis, we used an iterative process to refine the set of predictors with the *F*-values for entry and removal as a guide to obtain the best possible classification. All classification results were reported from cross-validation.

In the tiered approach, DFA models are created for each branch of the classification (Fig 1B), so that a model (model 1 in Fig. 1B) is built to discriminate group 1 (35 sites) from group 2 (52 sites). A 2nd model is built to separate groups 1A and 1B (model 1a in Fig. 1B). This 2nd model does not include the 52 sites in group 2 and, therefore, does not have to explain all the variability associated with those 52 sites. Similarly, a model is created for group-2 sites, excluding the 35 group-1 sites, and a model is built to predict subsequent groups (model 2a in Fig. 1B) following the same procedure until the classification is resolved. Each step in this tiered DFA method focuses on only part of the classification, thereby eliminating the variability among all the sites that must be explained in the standard single-model approach.

We constructed tiered models for each of 3 different data sets with the objective of improving the accuracy of predicting sites to the biological groups. Once these models were developed, we used them to assign validation (test) sites to a reference group. We conducted standard BEAST assessment using ordination of the test site with its predicted set of reference sites and a 90% confidence ellipse constructed around the reference sites. For this comparison, a site was considered to be in reference condition if it was inside the 90% ellipse and out of reference condition if it was outside the 90% ellipse. We chose semi-strong hybrid multidimensional scaling (Belbin 1991), which uses linear regression for association values <0.9 and ordinal regression for higher association values. We used 2 or 3 axes depending on the stress value (the correlation between the association matrix and the ordination). If stress was <0.15 with 2 axes, we used 2 axes. If stress was >0.15, then we used a 3rd axis.

The method was applied to the same 3 freshwater reference-site data sets used by other authors in this special series (Bailey et al. 2014). Two data sets were from large regions in North America: the Yukon River basin, Yukon Territory (YT) (323,800 km^2), and the Laurentian Great Lakes (GL) (244,000 km^2) in Canada. The 3rd data set, from the Upper Murrumbidgee River basin in the Australia Capital Territory Region (ACT), represented a much smaller geographic area (12,000 km^2). The data included benthic invertebrate assemblage and environmental variables. Sample collection, taxonomic resolution, and types of habitat data collected varied among data sets (Bailey et al. 2014). Each data set was divided into a *training* set of reference sites used to build assessment models and a *validation* set of reference sites used to test the model performance. Further details on the data sets are available in Bailey et al. (2014).

The validation sites were used in 2 ways to test the performance of the tiered DFA predictive models modeling approach: 1) ability to assess validation sites (undisturbed reference sites) as in reference condition, and 2) ability to assess simulated-impairment validation sites as not in reference condition. Erroneous assessment of a validation site (D0) as disturbed was defined as a Type 1 error, i.e., the number of reference validation sites (D0) that fell outside the 90% ellipse. We assessed the same validation sites after applying mild (D1), moderate (D2), and severe (D3) simulated impairment by eutrophication. These sites were known to be disturbed, so failure of the model and assessment process to identify them as such was defined as a Type 2 error, i.e., the number of simulated-impairment validation sites within the 90% ellipse.

Our assessments were based on the assumption that validation and training assemblages had similar distributions of biological assemblages. We tested this assumption by comparing the distributions of validation and training sites in ordination space. Ideally, validation site assemblages should be within the range of biological variation observed in the training data set. We also tested whether the models predicted the validation sites as belonging to the correct group. Normally, there is no absolute way of knowing to which reference group a test site belongs. However, in this case, the training and validation data were from a single set of reference sites, so we were able to do a classification of both the training and validation data for each data set. This procedure enabled us to determine with a high level of confidence the reference group with which a validation site was most strongly associated and to check whether it was assigned independently to that group by the model.

In summary, the measures we used to describe the performance of the tiered modeling approach were: 1) cross-validation rates with the training data set, i.e., the accuracy with which the models assigned training sites to known groups, based on the error rates from the individual submodels and the combined errors from each submodel required to classify a site (e.g., Fig. 1B: accuracy model 1 × accuracy model 1a); 2) the accuracy of assigning validation sites to the parent group based on the classification of the validation sites by the model; and, 3) the Type 1 and 2 error rates with validation and simulated-impairment sites.

RESULTS

Tiered models

Construction of tiered models with the 3 training data sets produced a set of models for each data set. Their use required a stepwise procedure to estimate the probability of a site belonging to a reference group.

The YT training data set, with 4 groups (Fig. 2A–C) and 3 submodels required the fewest steps (Table 2). The first step, model Y1 with 7 variables correctly (cross-

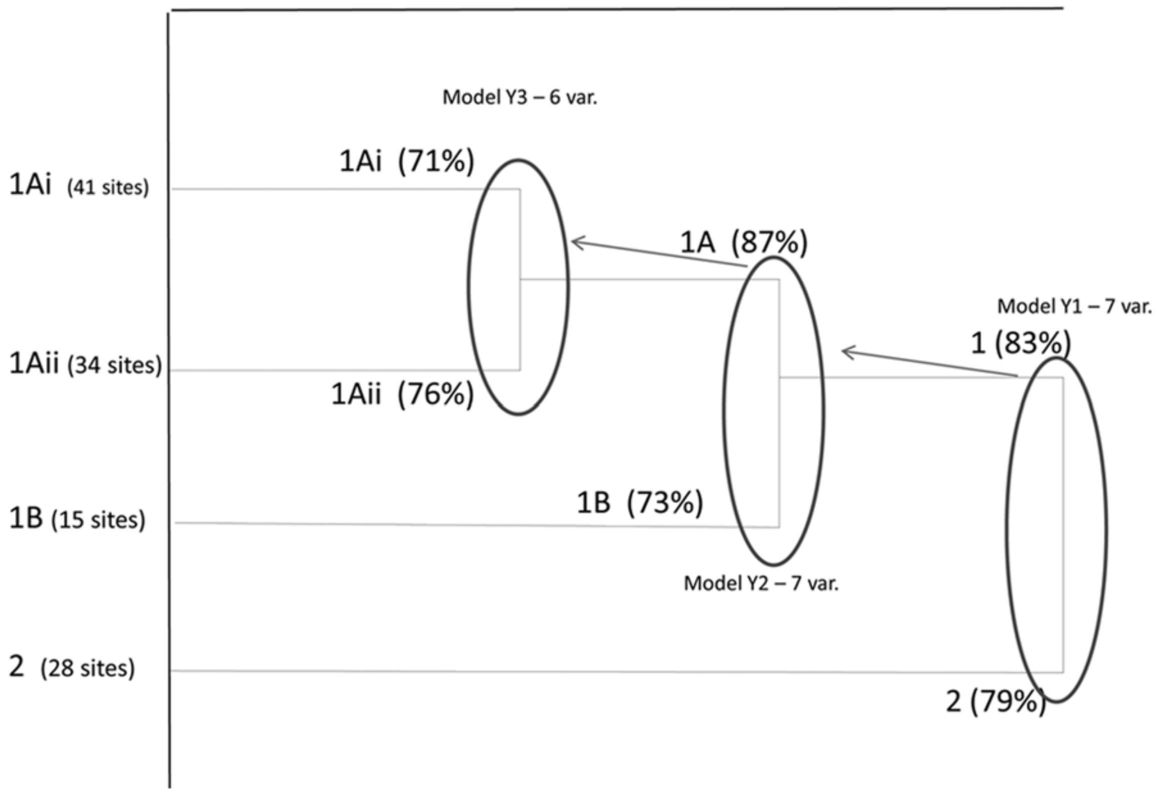

Figure 2. Classification of 118 training reference sites from the Yukon Territory (YT) data set showing 4 groups and the 3 tiered discriminant function analysis models used to classify to the groups.

validation) predicted 82% of the sites to 2 major groups (Fig. 2A, Table 2). A site predicted to group 2 (79% accuracy) required no further action. However, a site predicted to group 1 required a 2nd prediction step with model Y2. This model had an overall accuracy of 84% with 7 predictor variables. A site predicted to group 1B (accuracy 73%) required no further action. However, a site predicted to group 1A required a 3rd prediction step in model Y3. This model had 6 variables and an overall accuracy of 73% (Group 1Ai: 71%; Group 1Aii: 76%) (Table 2). However, these accuracies are for individual models. In application with new test data, inaccuracies can compound through each modeling step. For example, a site that was in either Group 1Ai or 1Aii would require the use of 3 models, so the misclassification could compound. To capture this compounding effect, classification rates are multiplied. Thus, a site predicted to Group 1Aii has an actual correct classification rate calculated as the classification rates from model Y1 × model Y2 × model Y3 (0.83)(0.87)(0.76) = 55% (Table 2).

Table 2. Performance of tiered models for the Yukon Territory (YT) data set in assigning training sites to reference groups.

Model	Sites correctly predicted (% accuracy)		Predictor variables	Cross-validation accuracy
Y1	Group 1: 75 of 90 (83%)	Group 2: 22 of 28 (79%)	7: longitude, sedimentary and volcanic bedrock, ultramafic/metamorphic bedrock, unforested land cover, June rainfall, annual snowfall, channel depth	82%
Y2	Group 1A: 65 of 75 (87%)	Group 1B: 11 of 15 (73%)	7: stream density, alpine land cover, % lakes in catchment, minimum June temperature, maximum June temperature, conductivity, substrate embeddedness	84%
Y3	Gp 1Ai: 29 of 41 (71%)	Gp 1Aii: 26 of 34 (76%)	6: latitude, sedimentary and volcanic bedrock, stream density, % wetlands, conductivity, dominant substrate	73%

Table 3. Summary of performance of tiered discriminant function analysis (DFA) models and a standard single-DFA model in assigning training sites to the correct group (cross-validation %) for the Yukon Territory (YT), Laurentian Great Lakes (GL), and Australian Capital Territory (ACT) data sets.

	Tiered model				
	Within levels		All levels		
Data set	Submodel (range)	Total	Groups	Average	Standard model (Strachan and Reynoldson 2014)
YT (*n* = 118)	73–84%	75%	51–79%	64%	58%
GL (*n* = 124)	76–92%	81%	39–62%	52.2%	62%
ACT (*n* = 87)	67–82%	70%	38–61%	51%	56%

We calculated these classification rates for each tiered model (Table 3).

The GL training data set with 7 groups (Fig. 3A–C) was more complex and required a total of 6 submodels (Table 4, Fig. 3A–C). However, any site required only 2 or 3 steps for a final prediction, with the sequence being model GL1, then either model GL1A for groups 1Ai, 1Aii, and 1B followed by model GL1B for groups 1Ai and 1Aii (Table 4, Fig. 3A–C). If a site was predicted to group 2 by model GL1, then model GL2 was used, followed by either model GL2A for groups 2Ai and 2Aii or model GL2B for groups 2Bi and 2Bii. None of the models used >5 variables and model GL2B required only 3 variables (Table 4, Fig. 3A–C). Individual submodels ranged in accuracy from 76 to 92%, but for all the required steps accuracy ranged from 39–62% (Table 3).

The ACT training data set with 6 groups contained 4 submodels (Table 5, Fig. 4), and each site required either 2 or 3 steps for complete prediction (Table 5, Fig. 4). The 1st step used model A1 (Table 5) with 4 variables and an overall accuracy of 82% (group 1: 74%; group 2: 87%). Sites predicted to group 1 required the use of model A1A (4 variables; 74% accuracy) to separate group 1A from group 1B (Table 5, Fig. 4). Sites predicted to

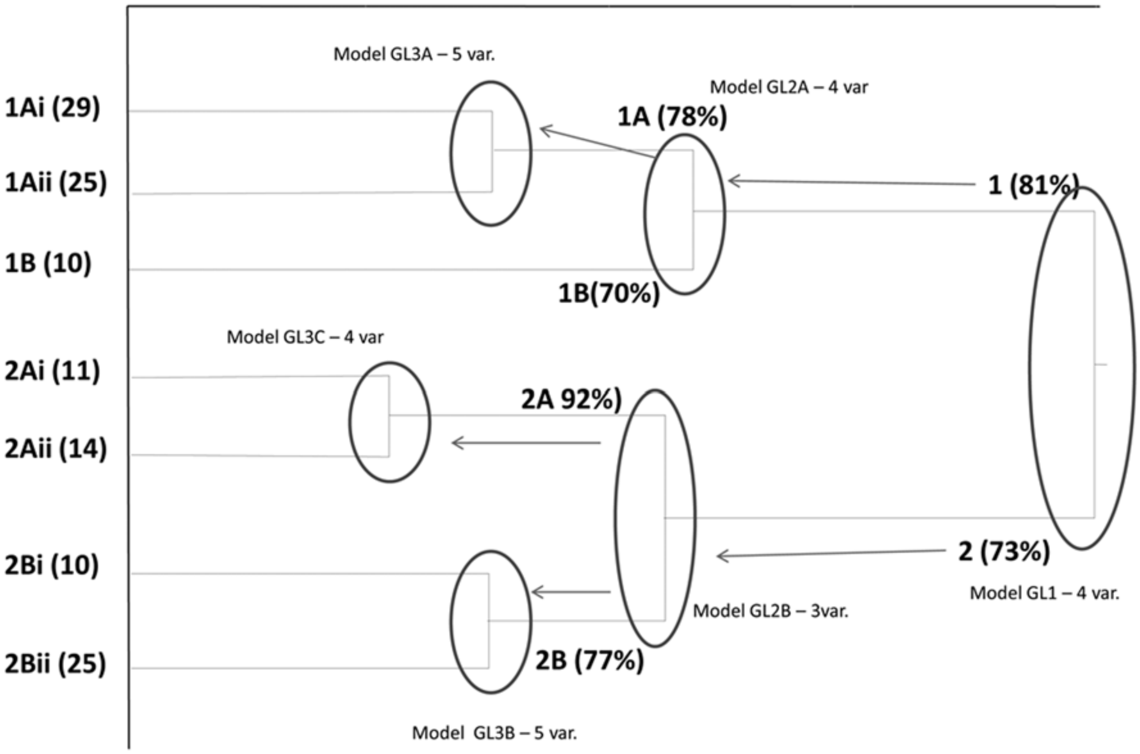

Figure 3. Classification of 124 training reference sites from the Laurentian Great Lakes (GL) data set showing 7 groups and the 6 tiered discriminant function analysis models used to classify to the groups.

Table 4. Performance of tiered models for the Laurentian Great Lakes (GL) data set in assigning training sites to reference groups. TOC = total organic C.

Model	Groups sites correctly predicted (% accuracy)		Predictor variables	Cross-validation accuracy
GL1	Group 1: 52 of 64 (81%)	Group 2: 44 of 60 (73%)	4: depth, silt %, 75th percentile particle size, $MgO_{(sediment)}$.	77%
GL2A	Group 1A: 42 of 54 (78%)	Group 1B: 7 of 10 (70%)	4: longitude, $alkalinity_{(water),}$ $TN_{(sediment)}$, $P_2O_{5(sediment)}$	77%
GL3A	Group 1Ai: 21 of 29 (72%)	Group 1Aii: 20 of 25 (80%)	5: $alkalinity_{(water)}$, $pH_{(water)}$, 75th percentile particle size, TOC $_{(sediment)}$, $Al_2O_{3(sediment)}$	76%
GL2B	Group 2A: 23 of 25 (92%)	Group 2B: 27 of 35 (77%)	3: longitude, depth, $alkalinity_{(water)}$	83%
GL3B	Group 2Bi: 7 of 10 (70%)	Group 2Bii: 22 of 25 (88%)	5: latitude, depth, $alkalinity_{(water)}$, 25th percentile particle size, $LOI_{(sediment)}$	83%
GL3C	Group 2Ai 10 of 11 (91%)	Group 2Aii 13 of 14 (93%)	4: latitude, TOC $_{(sediment)}$, $Na_2O_{(sediment)}$, $MnO_{(sediment)}$	92%

group 2 required model A2 and, if predicted to group 2A, a final step with model A2A that predicted sites to 1 of 3 groups (Table 3, Fig. 4). Correct classification rates ranged from 38 to 61% (Table 3).

Model performance

The cross-validation performance of correctly assigning the training sites for the tiered models ranged from 67 to 82% for the ACT (Table 5), 76 to 92% for the GL (Table 4), and 73 to 84% for the YT data sets (Table 2). Within the individual models for the various models in the classification this performance is substantially better than the typical RCA models we have constructed which typically fall in the 50 to 60% accuracy range (Reynoldson et al. 2000, 2001). However, when the classification rates are adjusted to account for the use of each model step required to acquire a final classification then the models perform slightly better (YT) than or similarly (GL, ACT) to single models (Table 3). Strachan and Reynoldson (2014) showed that the accuracy of the standard method with these data sets was typically 56 to 62%. The tiered approach was markedly better than the standard method (Table 3) in terms of correctly assigning sites to groups. Fourteen to 19% more training sites were correctly assigned (Table 3), but because of the additive effects of misclassifications the classification performance with test sites was not as clear.

The group membership of the validation sites was known, so we were able to compare the observed accuracies (i.e., the groups to which the sites were predicted) with the expected accuracies based on the model error rates. These data are shown for the YT validation sites (Table 6). Of the 13 validation sites belonging to group 1Ai only 6 (46%) were predicted to that group, but we

Table 5. Performance of tiered models for the Australian Capital Territory (ACT) data set in assigning training sites to reference groups.

Model	Sites correctly predicted (% accuracy)			Predictor variables	Cross-validation accuracy
A1	Group 1: 26 of 35 (74%)	Group 2: 45 of 52 (87%)		4: longitude, bankfull width, $alkalinity_{(water)}$, % $boulder_{(substrate)}$	82%
A2A	Group 1A: 12 of 18 (67%)	Group 1B: 14 of 17 (82%)		4: latitude, catchment area, bankfull width, % $gravel_{(substrate)}$	74%
A2B	Group 2A: 36 of 45 (80%)	Group 2B: 5 of 7 (71%)		5: %riffle, % $boulder_{(substrate)}$, % $cobble_{(substrate)}$, % $pebble_{(substrate)}$, % $gravel_{(substrate)}$	79%
A3	Group 2Ai: 6 of 11 (55%)	Group 2Aii: 14 of 19 (74%)	Group 2Aiii: 10 of 15 (67%)	7: altitude, bankfull width, bankfull height, water velocity, pool-riffle ratio, % $bedrock_{(substrate)}$, % $boulder_{(substrate)}$	67%

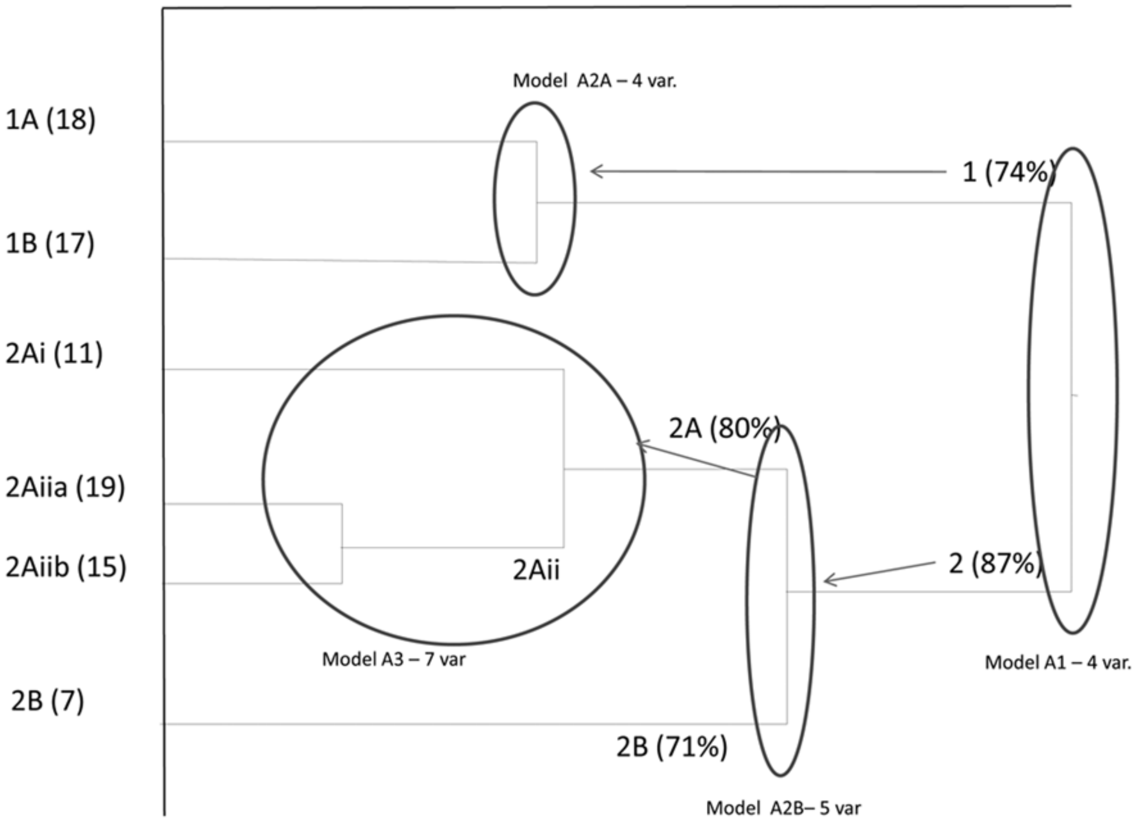

Figure 4. Classification of 87 training reference sites from the Australian Capital Territory (ACT) data set showing 6 groups and the 4 tiered discriminant function analysis models used to classify to the groups.

would expect 7 (55% from the combined error rates in the tiered model). In general the model was less accurate than anticipated. Only 15 of an expected 25 sites were correctly assigned (Table 6).

Site assessment

Assessments were conducted on the validation sites, undisturbed sites known to be in reference condition (D0 sites), and on simulated-impairment validation sites (D1–D3) (Table 7). For the ACT data set, 14 of the 20 validation sites (D0) were assessed as disturbed (equivalent to a Type 1 error rate = 70%). Sixteen D1 sites were correctly assessed as disturbed, and 4 were incorrectly assessed as not disturbed (Type 2 error rate = 20%). The Type 2 error rates for D2 and D3 sites were 15 and 10%, respectively (Table 7). For the GL data set, the Type 1 error rate was 47.5%, and the Type 2 error rates were 52.5, 62.5, and 57.5% for D1, D2, and D3 sites, respectively (Table 7). For the YT data set, the Type 1 error rate was 45%, and the Type 2 error rates were 57.5, 37.5, and 22.5% for the D1, D2, and D3 sites, respectively (Table 7).

We have no way of knowing a priori what the Type 2 error rates should be, and this error is, in a sense, a mea-

Table 6. Comparison of inferred group membership of Yukon Territory (YT) validation sites from classification to that predicted by the model.

	Predicted to group				
Group to which validation site belongs	1Ai	1Aii	1B	2	Total
1Ai	6 (46%)	2	1	4	13
1Aii	2	4 (33%)	1	5	12
1B	1	2	1 (20%)	1	5
2	4	2	0	4 (40%)	10
Expected accuracy	7 (55%)	6 (51%)	4 (72%)	8 (79%)	25

Table 7. Summary of the assessments of validation and simulated-impairment sites for the Australian Capital Territory (ACT), Laurentian Great Lakes (GL), and Yukon Territory (YT) data sets. For each data set, the results are presented (Table 3) for the number of validation sites assessed as undisturbed (D0) and disturbed (D1, D2, D3), using the Benthic Assessment of Sediment (BEAST) 90% ellipse, for each level of disturbance. Error rates (%) are equivalent to Type 1 error for the D0 sites and the Type 2 error for the D1–D3 sites. For comparison the error rates are reported for the standard Benthic Assessment of Sediment (BEAST) method by Strachan and Reynoldson (2014).

	Level of disturbance			
Data set	None (D0)	Mild (D1)	Moderate (D2)	Severe (D3)
ACT (*n* = 20)				
Assessed as undisturbed	6	4	3	2
Assessed as disturbed	14	16	17	18
Error rate (%)	70.0	20.0	15.0	10.0
*BEAST error (%)	70.0	20	20	5
GL (*n* = 40)				
Assessed as undisturbed	21	21	25	23
Assessed as disturbed	19	19	15	17
Error rate (%)	47.5	52.5	62.5	57.5
*BEAST error (%)	30	65	60	55
YT (*n* = 40)				
Assessed as undisturbed	22	23	15	9
Assessed as disturbed	18	17	25	31
Error rate (%)	45.0	57.5	37.5	22.5
*BEAST error (%)	53	43	25	25

sure of the power of the assessment. However, we would expect a Type 1 error rate for D0 validation sites of ~10% because α is set at 0.10 by the 90% probability ellipse. In fact, the Type 1 error actually ranged between 45 and 70% (Table 7) and 51 of the 100 validation sites (for the 3 data sets) were assessed as not in reference condition (Table 7). At least 2 explanations exist for the high Type 1 error rate. First, the validation sites might not have been representative of the training data set. Second, the models might have incorrectly assigned validation sites to reference groups so that they were not assessed with the appropriate set of reference sites.

As described by Bailey et al. (2014), the validation sites were selected so as to be within the central range (25th–75th percentile) of the habitat characteristics of the entire data set. However, this criterion does not guarantee that they are necessarily in the same range of the benthic assemblage. We plotted both training and validation sites in nonmetric multidimensional scaling (NMDS) ordination space to compare the distribution of the invertebrate assemblages at training and validation sites (Fig. 5A–C). For the ACT data set, only 2 of the 20 (10%) validation sites were outside the training 90% ellipse (Fig. 5A), and the average similarity of the training and validation sites to the training site median differed little. For the GL data set, 9 of 40 (22.5%) validation sites were outside the 90% ellipse (Fig. 5B), and for the YT data set, only 3 of 40 (8%) sites were outside the 90% ellipse (Fig. 5C). For the GL data set, the validation sites were more different from the training sites than would be expected (>2× as many sites as expected were outside the 90% ellipse; Fig. 5B), and the selection of the validation sites may explain part of the high Type 1 error rate (Table 7). However, this was not the case for the ACT or YT data sets, where the number of validation sites outside the 90% ellipse was equal to or less than expected (Fig. 5A,C).

The 2nd explanation for high Type 1 errors was inaccurate matching of sites to reference groups. Our comparison of the groups to which the validation sites were predicted relative to the group to which they actually belonged (Table 6) suggests inaccurate matching as a cause for the observed errors.

DISCUSSION

We examined 2 aspects of the tiered modeling approach: the accuracy of the model and the error in assessment. Commonly used RCA methods use the reference group(s) predicted for a test site as the basis of comparison in the site-assessment step. Both the observed/expected (O/E) score (RIVPACS, AUSRIVAS) and the ordination approach (CABIN) use the probability that a test site belongs to a reference group. Therefore, the greater the accuracy in

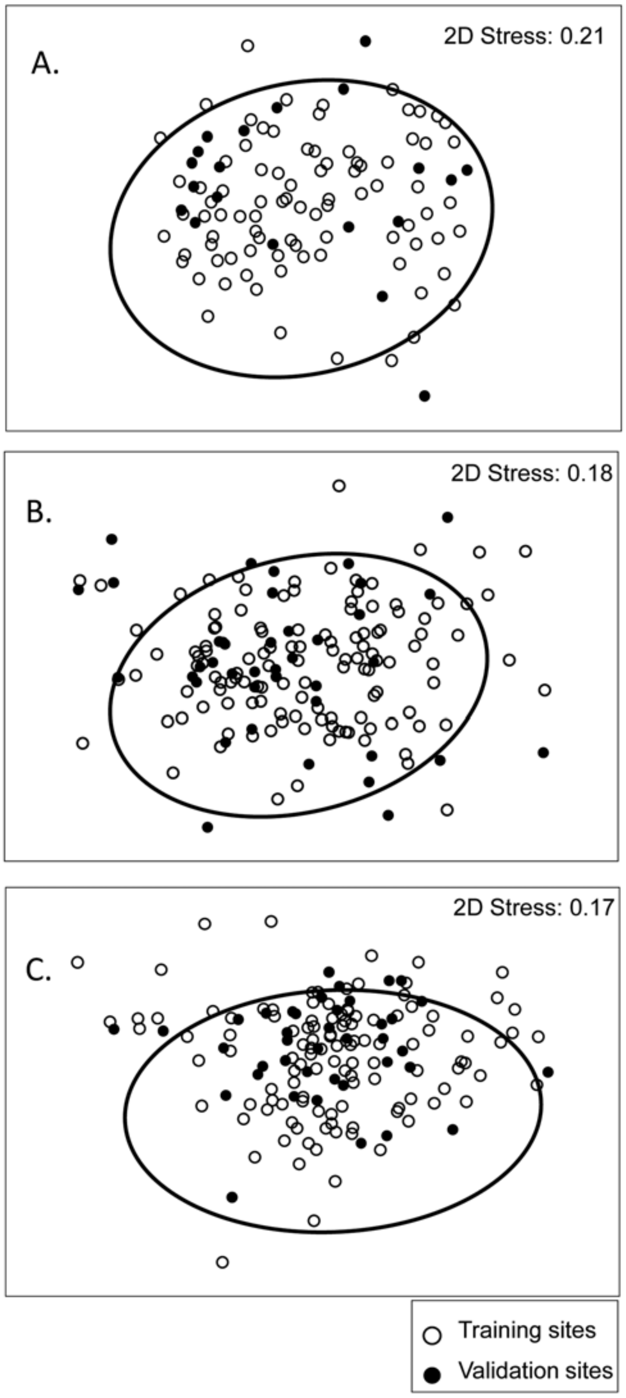

Figure 5. Nonmetric multidimensional scaling (NMDS) ordination of reference data sets from the Australian Capital Territory (ACT) (A), Laurentian Great Lakes (GL) (B), and Yukon Territory (YT) (C). For each data set the training (open) and validation (solid) sites are shown with a 90% probability ellipse constructed around training sites only.

matching the test site to a reference group, the more confidence a user may have in the assessment.

The individual tiered models provided much higher accuracy than the single-step model. Fourteen to 19% more training sites were correctly classified by the tiered models than by the standard approach. From this perspective, the tiered models performed better than the standard approach. However, when the accumulated misclassifications accrued through multiple tiered models, the performances of the tiered and standard models were similar. Tiered models with as few levels as possible are desirable.

Comparison of the tiered and standard models on the basis of assessment of validation sites yielded equivocal results. Type 1 error rates ranged from 45 to 70%. Given the α level of 0.10 (90% probability ellipse), these Type 1 error rates are much higher than the 10% expected. The comparison with the standard BEAST method (Strachan and Reynoldson 2014) showed that Type 1 error rates were the same between methods for the ACT data, higher with the tiered method for the GL data (47.5 vs 30%), and lower with the tiered method for the YT data set (45 vs 53%) (Table 7). Type 2 error rates ranged from 10% (ACT, D3) to 62.5% (GL, D2) (Table 7).

Each data set had particular issues. With the ACT data set, the Type 1 error rate was high, the trend for the assessment of D1 to D3 sites was as expected, but the difference among the 3 levels of simulated impairment was small. However, the ACT data consisted of relative abundances. Therefore, a change in total abundance, which in ordination-based assessment contributes substantially to the assessment, would have no effect on the assessment and might explain the lack of discrimination among the levels of disturbance. The GL data set was particularly problematic because more D0 sites (19) than D3 sites (17) were assessed as different from reference. The selection of validation sites resulted in a higher Type 1 error than expected, and it appeared that the greater the disturbance, the fewer the number of sites assessed as out of reference. The YT data set behaved as expected. More sites were assessed as disturbed as the level of disturbance increased, but the Type 1 error rate was high.

The selection of validation sites is part of the issue with the GL data set, but the other major concern in terms of the assessments is that many validation sites appear to be incorrectly matched to groups. To remove the error associated with the incorrect assignment of validation sites, we recalculated the error rates for only those sites that were correctly predicted to the group to which they belong based on classification of the complete data set (Table 8). The results have error rates much closer to those expected, particularly for the YT and ACT data sets. The YT data set did have high Type 2 error for D1 sites. However, the error rates for the GL data set remained high, and the Type 2 error rate remained inconsistent with expectations (higher Type 2 error with D2 than with D1 sites). In the case of the GL model, the classification itself may have to be re-examined.

Table 8. Number of correctly matched validation (D0) and simulated-impairment validation sites (D1–D3) in the Australian Capital Territory (ACT), Laurentian Great Lakes (GL), and Yukon Territory (YT) data sets assessed as different from reference. Type 1 and Type 2 error rates are in parentheses.

Disturbance	None (D0)	Mild (D1)	Moderate (D2)	Severe (D3)
ACT (*n* = 5)	1 (20%)	3 (40%)	3 (40%)	5 (0%)
GL (*n* = 15)	6 (40%)	6 (60%)	2 (87%)	5 (67%)
YT (*n* = 15)	2 (13%)	2 (87%)	7 (53%)	11 (27%)

Several DFA models have been reconstructed to include new reference-site data. For models constructed for the Fraser River (British Columbia), the Yukon, and the UK (Reynoldson and Wright 2000), the capacity of models to predict sites accurately decreased as the number of reference sites increased and the models become more complex (Table 1). This situation is not surprising because including more reference sites in a model increases the variation and complexity that must be partitioned with a single-DFA model. Our experience when presenting a model to users has been that a model with 50% accuracy does not inspire confidence, particularly in the case of ordination-based assessment where the test site is compared only to the group with the highest probability of occurrence. Low prediction accuracy creates the perception that assessment accuracy is poor. However, the basis for such an opinion must be considered in light of the number of groups. For 2 groups, 50% prediction accuracy would be no improvement over a null model, but with 10 groups, it would be 4 times better than a null model. Furthermore, ordination-based assessment carries a high degree of redundancy or robustness because many reference-site groups overlap in ordination space so that often, regardless of a low probability of prediction to a reference group, sites in adjacent groups share similar ordination space and, therefore, have comparable reference sites.

Based on our analyses alone, we find it difficult to recommend the tiered method over the standard method. We have no way to state absolutely that one model is an improvement over another, but the desirable attributes of RCA models for model builders are fewer predictor variables, more groups, and higher accuracy. The initial classification in the tiered approach is much superior to the classification in the standard approach. However, the potential accumulation of misclassifications through the levels may result in a similar performance between the 2 methods, particularly if several tiered models are required to resolve a classification. Classification with a single-tier model had high accuracy (79%), but accuracy was 60 ± 8% (SD) when 2 models were required and 50 ± 5% with 3 tiers. On the other hand, tiered models used more reference groups for the data sets explored in this series than reported for the standard approach (Strachan and Reynoldson 2014). Thus, their resolution with DFA models was more complex and provided greater partitioning of the natural variation with potentially higher sensitivity in assessment. Moreover, the tiered models required fewer predictor variables. The standard approach used 8 (GL) or 9 (ACT, YT) variables (Strachan and Reynoldson 2014), whereas the tiered models used 3 to 7 variables (Tables 2, 4, 5). Models built with the tiered DFA method allowed more groups to be similarly or more accurately predicted with fewer predictor variables compared with the standard DFA method.

In summary, the model prediction performance with the training sites can be better with the tiered method than with the standard approach using a single model. However, with the data used in our study, this improvement does not yield lower error rates. Both Type 1 and 2 errors were similar to that produced by the single-step method (Strachan and Reynoldson 2014). The overall poor performance, particularly the high Type 1 error rates, are problematic. They were related, in part, to the selection of the validation sites (GL), but also to inaccuracies in matching of sites to reference groups. Thus, approaches using validation and back-checking of model performance are critical when building models, and results of these tests should be reported. Tests with simulated impairment data sets, or some equivalent, should be standard when testing assessment performance. We also suggest that both standard and tiered models be considered. However, we see no obvious way to decide a priori which approach is likely to result in more accurate prediction, and prediction accuracy seems to be critical for the quality of the assessments. Some compromises were made in assembling these data sets, and these compromises, particularly those related to availability of habitat variables, could have reduced model performance. However, it also should be acknowledged that this exercise probably is one of the few examples where an assessment approach has been so rigorously tested.

ACKNOWLEDGEMENTS

We thank Susan Nichols, Evan Harrison, and the Environment and Sustainable Development Directorate, ACT Government, Canberra, Australia for the ACT (Murrumbidgee River) data set; John Bailey and the Yukon Government, DFO and UWO for the Yukon data set; and Lee Grapentine and Environment Canada for the Great Lakes data set. We also thank the Institute for Applied Ecology, University of Canberra, for a Professorial Fellowship to the senior author during which much of the work was undertaken. Don Jackson and an anonymous referee provided input that substantially contributed to improving the final manuscript. Most important, we wish to recognize Richard Norris, to whom this special issue is dedicated, for his passion and expertise in bioassessment.

LITERATURE CITED

Bailey, R. C., S. Linke, and A. G. Yates. 2014. Bioassessment of freshwater ecosystems using the Reference Condition Approach: comparing established and new methods with common data sets. Freshwater Science 33:1204–1211.

Belbin, L. 1991. Semi-strong hybrid scaling, a new ordination algorithm. Journal of Vegetation Science 2:491–496.

Feio, M. J., T. B. Reynoldson, and M. A. S. Graça. 2006. Effect of seasonal and inter-annual changes in the predictions of the Mondego River model at three taxonomic levels. International Review of Hydrobiology 91:509–520.

Johnson, R. 2003. Development of a prediction system for lake stony-bottom littoral macroinvertebrate communities. Archiv für Hydrobiologie 158:517–540.

Pardo, I., C. Gómez-Rodríguez, R. Abraín, E. García-Roselló, and T. B. Reynoldson. 2014. An invertebrate predictive model (NORTI) for streams and rivers: sensitivity of the model in detecting stress gradients. Ecological Indicators 45:51–62.

Reynoldson, T. B., R. C. Bailey, K. E. Day, and R. H. Norris. 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., M. Bombardier, D. B. Donald, H. O'Neill, D. M. Rosenberg, H. Shear, T. Tuominen, and H. H. Vaughan. 1999. Strategy for a Canadian aquatic biomonitoring network. NWRI 99-248. Environment Canada, Burlington, Ontario.

Reynoldson, T. B., K. E. Day, and T. Pascoe. 2000. The development of the BEAST: a predictive approach for assessing sediment quality in the North American Great Lakes. Pages 165–180 *in* J. F. Wright, D. W. Sutcliffe, and M. T. Furse (editors). Assessing the biological quality of freshwaters. RIVPACS and other techniques. Freshwater Biological Association, Ambleside, UK.

Reynoldson, T. B., D. M. Rosenberg, and V. H. Resh. 2001. Comparison of models predicting invertebrate assemblages for biomonitoring in the Fraser River catchment, British Columbia. Canadian Journal of Fisheries and Aquatic Sciences 58:1395–1410.

Reynoldson, T. B., and J. F. Wright. 2000. The reference condition: problems and solutions. Pages 293–303 *in* J. F. Wright, D. W. Sutcliffe, and M. T. Furse (editors). Assessing the biological quality of freshwaters. RIVPACS and other techniques. Freshwater Biological Association, Ambleside, UK.

Simpson, J. C., and R. H. Norris. 2000. Biological assessment of river quality: development of AusRivAS models and outputs. Pages 125–142 *in* J. F. Wright, D. W. Sutcliffe, and M. T. Furse (editors). Assessing the biological quality of freshwaters. RIVPACS and other techniques. Freshwater Biological Association, Ambleside, UK.

Strachan, S. A., and T. B. Reynoldson. 2014. Performance of the standard CABIN method: comparison of BEAST models and error rates to detect simulated degradation from multiple data sets. Freshwater Science 33:1225–1237.

Wright, J. F., D. Moss, P. D. Armitage, and M. T. Furse. 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 14:221–256.