Using Test Site Analysis and two Nearest Neighbor Models, ANNA and RDA, to Assess Benthic Communities with Simulated Impacts

Authors: Sarrazin-Delay, Chantal L., Somers, Keith M., and Bailey, John L.

Source: Freshwater Science, 33(4) : 1249-1260

Published By: Society for Freshwater Science

URL: https://doi.org/10.1086/678702

Using Test Site Analysis and two nearest neighbor models, ANNA and RDA, to assess benthic communities with simulated impacts

Chantal L. Sarrazin-Delay[1,4], Keith M. Somers[2,5], and John L. Bailey[3,6]

[1]Living with Lakes Center, Laurentian University, Sudbury, Ontario, Canada, P3E 2C6
[2]Dorset Environmental Science Centre, Ontario Ministry of the Environment and Climate Change, Dorset, Ontario, Canada, P0A 1E0
[3]Living with Lakes Center, Ontario Ministry of the Environment and Climate Change, Sudbury, Ontario, Canada, P3E 2C6

Abstract: Reference Condition Approach bioassessment programs have been in place in the northern and Muskoka regions of Ontario, Canada, for many years. Assessments are carried out regularly to evaluate and monitor the effects of a variety of activities, including mining, forestry, and cottage development. These programs are run by the Co-operative Freshwater Ecology Unit (CFEU) at Laurentian University in Sudbury, Canada, and the Dorset Environmental Science Centre (DESC) in Muskoka, Canada. We applied 2 bioassessment methods used at the CFEU and DESC to 3 data sets that were subjected to simulated impact by nutrient enrichments to compare their performance with a number of other bioassessment methods. We used Assessment by Nearest Neighbour Analysis (ANNA) and a Redundancy Analysis (RDA) variation of ANNA with Test Site Analysis (TSA) to identify subsets of reference sites to compare with a given simulated impact test site based on habitat matching. We compared the benthic macroinvertebrate (BMI) communities and evaluated the differences between the validation and training sites to assess the degree of impairment. After assessing all impacted sites, we calculated Type 1 and Type 2 error rates. ANNA and RDA separated sites with different levels of simulated impact in an Australian data set of diverse benthic macroinvertebrate communities distributed along a habitat gradient. In contrast, our assessments did not perform well with 2 data sets for which the simulation did not behave as expected, perhaps because of impoverished communities.

Key words: reference condition, benthic macroinvertebrates, Assessment by Nearest Neighbour, redundancy analysis, simulated impact

The Canadian Fisheries Act (FA) provides protection for fish and their habitat. As part of the Environmental Effects Monitoring (EEM) requirements of the Metal Mining Effluent Regulations of the FA, mining companies are required to monitor the biological effects of their discharge on receiving waters. This requirement includes monitoring the benthic macroinvertebrate (BMI) community (Walker et al. 2003). BMI communities from sites exposed to mine effluent are compared to communities from unexposed reference areas and an effect is defined as a statistically significant difference between exposed and unexposed communities (Glozier et al. 2002).

An increasingly used study design for such comparisons is the Reference Condition Approach (RCA) in which a database is developed of reference sites where human disturbance has been minimal, i.e., no point-source pollution, logging activity, etc. (Simon 1991, Omernik 1995). Habitat is used to match an exposed site to a number of reference sites, and the BMI community at the test site of interest is compared to that found at the appropriate reference sites (Bailey et al. 2004, Bowman and Somers 2005).

These types of assessments are used in Ontario by the Ontario Benthos Biomonitoring Network (OBBN) (Jones et al. 2005) at the Dorset Environmental Centre and the Freshwater Invertebrate Research Network of Northern Ontario (FIRNNO) (Reynoldson et al. 2005, Sarrazin-Delay et al. 2006) at the Cooperative Freshwater Ecology Unit. These networks were established to define reference conditions for benthos in Ontario. The resultant RCA databases include BMI abundance data, water chemistry and site-, chan-

E-mail addresses: [4]csarrazindelay@laurentian.ca; [5]keith.somers@ontario.ca; [6]jbailey@laurentian.ca

DOI: 10.1086/678702. Received 15 April 2013; Accepted 22 April 2014; Published online 25 September 2014.
Freshwater Science. 2014. 33(4):1249–1260.

nel-, and watershed-level habitat data for hundreds of reference and test sites. The habitat variables are used at the site-matching step of RCA assessment (Bowman and Somers 2005). In the Benthic Assessment of Sediment (BEAST) approach, clustering based on BMI followed by Discriminant Functions Analysis (DFA) is used as the site-matching strategy (Reynoldson et al. 1995, 1997). The BEAST approach assumes discrete groupings of reference sites, but in reality, BMI communities may span a continuum. As a result, test sites may be classified into the incorrect cluster of reference sites (Bowman and Somers 2005). A similar clustering method is used for Australian River Assessment System (AUSRIVAS; Simpson and Norris 2000). In contrast, groups of reference sites are not assumed in the Assessment by Nearest Neighbour Analysis (ANNA; Linke et al. 2005). In ANNA, the test-site BMI community is simply compared to the BMIs from reference sites that most resemble the test site with respect to their habitat variables. Broadly, ANNA involves: 1) matching a test site to appropriate reference sites based on closest distance in multivariate habitat space, its nearest neighbors (NNs), and 2) comparison of the BMI community at a test site to the communities from the matched NN subset of reference sites to determine degree of impairment. In general, ANNA modeling is simple, requires fewer steps than BEAST and AUSRIVAS, and is conducive to incorporating new reference data as they become available.

The BEAST and AUSRIVAS approaches are based on the assumption that clusters or types of BMI communities exist at reference sites, whereas the ANNA approach assumes that the reference sites belong to 1 large group. BMI communities from reference sites also can span a gradient of habitat conditions (Bailey et al. 1998). Multivariate methods that model community changes along gradients are available (Legendre and Legendre 1998). Redundancy Analysis (RDA) or Canonical Correspondence Analysis (CCA) could be used to model community–habitat relationships and subsequently to match test sites with neighboring reference sites (Verdonschot 1995, Bowman and Somers 2005). As a complement to ANNA, we used RDA to model the community–habitat relationship for reference sites and used these models to select NNs to evaluate test sites, much like the ANNA approach.

We applied ANNA and RDA assessments to 3 data sets as part of a collection of studies in which the ability of commonly used bioassessment methods to correctly classify sites with simulated enrichment impacts was evaluated (Bailey et al. 2014). Simulated impacts of varying severity were applied to reference-site data by deliberately altering abundance and richness of macroinvertebrates from a randomly selected subset of reference sites as described by Bailey et al. (2014). We assessed reference sites with 1 of 4 levels of simulated impacts (none to severe) with a series of biological metrics in a Test Site Analysis (TSA; Bowman and Somers 2006). A test site was deemed to be impaired when any of the metrics fell outside the normal range for the reference sites (e.g., Kilgour et al. 1998, Glozier et al. 2002). We used the 3 data sets to evaluate the ability of ANNA and RDA methods to assess impaired sites correctly.

METHODS

Data sets

Data consisted of BMI and habitat data from 3 independent data sets from Yukon Territory (YT) streams, Laurentian Great Lakes (GL) nearshore areas, and Australian Capital Territory (ACT) streams (Bailey et al. 2014). We selected data sets from areas that differed in geography and habitat and that were obtained with differing sampling methods to test the various assessment methods rigorously.

YT data set In the Yukon region of northern Canada, BMIs were sampled with a travelling kick-and-sweep technique at 158 reference streams between 2004 and 2006. Samples were subsampled in the laboratory to a fixed 300-count with abundance values extrapolated to the entire sample. Forty-two habitat variables were recorded at each site.

GL data set Great Lakes nearshore lake sites were sampled over the period of 1991 to 2010 with a box core, mini-box core, or Ponar sampler and subsequently corrected for area sampled. Entire samples were counted; when excessive amounts of material were encountered, subsampling was done using a Marchant box with fixed count of 200 prior to 2000 and 300 in subsequent years. Fixed count data were scaled up to the whole sample and expressed /m^2. Twenty-five variables were used to characterize the habitat of the 164 reference lake sites.

ACT data set Australian Capital Territory reference-stream riffle sites were sampled in spring 1994 and 1995 with a 10-m-long kick-and-sweep method. Samples were subsampled to a minimum 200 count and BMIs were expressed as relative abundance. Data were not scaled to the entire sample. In this case, 26 habitat variables were recorded for each of the 107 sites.

Simulated impacts

For each data set, a subset (40 for YT and GL, 20 for ACT) of the reference sites was randomly selected to be assessed as test (validation) sites. Each of these sites was subjected to 4 levels of simulated impact to create 4 sets of validation sites in which abundance and richness had been altered to simulate the effects of no (D0), mild (D1), moderate (D2), and severe (D3) enrichment of each site (Table 1; Bailey et al. 2014). The simulated impacts preferentially removed sensitive families at progressively stronger levels. Taxon tolerances used to alter the BMI data were based on the Hilsenhoff family-level tolerance value (TV; Hilsenhoff 1988) for North American sites

Table 1. Simulated impact treatment applied to minimally disturbed validation sites to create 4 levels of disturbance. Effects were applied to sensitive, intermediate, and tolerant families in benthic macroinvertebrate communities.

Impact severity	Code	Effects applied to validation minimally disturbed sites
Unimpacted	D0	None
Mild	D1	Sensitive: −25% abundance, −10% taxa
		Intermediate: unchanged
		Tolerant: +75% abundance
Moderate	D2	Sensitive: −75% abundance, −50% taxa
		Intermediate: −50% abundance, −20% taxa
		Tolerant: −25% abundance, −10% taxa
Severe	D3	Sensitive: −100% taxa
		Intermediate: −75% abundance, −50% of taxa
		Tolerant: −50% abundance, −20% taxa

(YT, GL) or Stream Invertebrate Grade Number (SIGNAL) tolerance value (Chessman et al. 1997) for ACT sites. Sensitive taxa had TV 1–4 or SIGNAL 7–10, intermediate had TV 5–6 or SIGNAL 4–6, and tolerant had TV 7–10 or SIGNAL 1–3.

Data preparation

For each data set, we prepared the habitat data by replacing missing values for the habitat variables with the mean value of the variable, removing outliers, and centering and standardizing variables (Fig. 1). We did not include habitat variables related to the simulated enrichment impact in the analyses. In the GL data set, water nutrients, dissolved O_2, and temperature at the lake bottom were excluded. The GL data set also had an outlier that was excluded, site 1216. This site was the only one with a coarse substrate consisting of gravel and sand, whereas substrate at all other sites consisted of silt and sand. We included all ACT and YT habitat and BMI data in modeling and site assessment.

Site matching

Site matching is central to the RCA. We used 2 site-selection methods: ANNA and RDA. For both methods, the 1st step was to reduce the habitat variables to a set of linearly uncorrelated variables with principal components analysis (PCA; Fig. 1). We centered and standardized the habitat data (z-score) then used the *Biplot* add-in in Excel 2007 (Microsoft, Redmond, Washington; Lipkovich and Smith 2001) to run the PCAs. We used the broken-stick model (Jackson 1993) to determine the number of nontrivial axes to retain for site matching. For ANNA, Euclidean distance of nontrivial PCA axis scores between each validation site and all of the training sites was used to select 30 NNs. In other words, the 30 training sites closest to a validation site in habitat ordination space were selected to assess the validation site.

For the RDA, we used the same PCA scores in multiple regressions with each of the 4 Environmental Effects Monitoring (EEM) metrics (density, family richness, evenness, and Bray–Curtis distance) as the dependent variables (i.e., $\text{Metric}_1 = b_0 + b_1x_1 + b_2x_2 + \ldots b_nx_n$, where x_n values are the habitat PCA axis scores and b_n values are the slopes) to calculate the expected (predicted) metric values for each training site. We rotated the predicted training-site values using PCA to produce RDA scores (Legendre and Legendre 1998) and projected the predicted validation-site values onto the RDA ordination. We used Euclidean distance between each validation site and all of the training sites in the RDA to identify the 30 NNs. The Euclidean distance was weighted by explained variation for each of the axes. Therefore, the axis that explained most of the variations weighed more heavily on the results.

Community metrics

We calculated 4 summary metrics (as described by the EEM program; Walker et al. 2003) for the training and validation sites: total BMI density, family richness, Simpson's evenness (0 = community dominated by 1 or 2 families, 1 = even distribution of families), and Bray–Curtis distance (0 = same community, 1 = different community) (Fig. 1; Glozier et al. 2002). In the case of YT and ACT data sets, timed sampling did not allow calculation of density, so abundance, standardized by sampling effort, was used. For the sake of simplicity, future reference to this metric, regardless of data set, will be referred to as density. Assumptions of normality were satisfied by using $\log_{10}$(density). Transformation was not necessary for richness, evenness, and Bray–Curtis distance.

TSA

Thirty training-site NNs for each validation site were used to calculate the NN mean for each of the biological metrics (Fig. 1). We defined the effect size (ES) as the value for a validation-site metric falling outside the normal range of variation among the 30 NNs (mean ± 2 SD, 30 NNs), i.e., outside the cloud enclosing 95% of the variation in the NN sites as described by Kilgour et al. (1998). The ES was calculated as $(X_{val} - \bar{X}_{NN})/SD_{NN}$, where X_{val} is the metric value for the validation site, $\bar{X}_{NN}$ is the mean metric value for the 30 NNs, and SD_{NN} is the standard deviation of the 30 NNs. The univariate noncentral F test (ES > normal range), calculated using the *πface* add-in (Excel 2007; Microsoft, Redmond, Washington; Lenth 2003), was used as a measure of the overall biological difference between each validation site and its NNs for each of the metrics. This approach is called

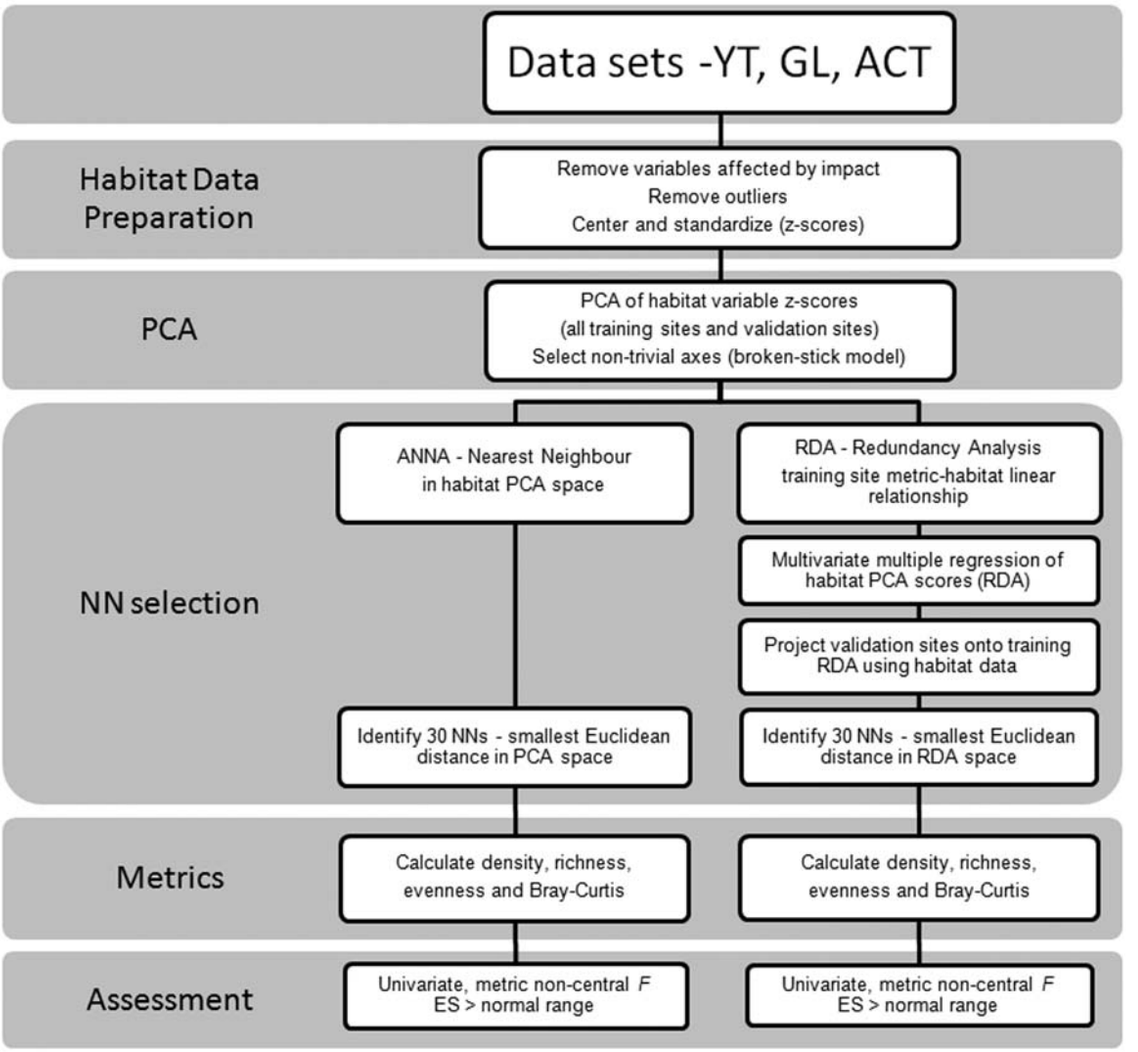

Figure 1. Flow chart describing steps used in the bioassessment of simulated impact sites using Assessment by Nearest Neighbour Analysis (ANNA) and Redundancy Analysis (RDA) in the Reference Condition Approach context. YT = Yukon Territory, GL = Laurentian Great Lakes, ACT = Australian Capital Territory, PCA = principal components analysis, NN = nearest neighbor, ES = effect size.

Test Site Analysis (TSA; Bowman et al. 2003, Bowman and Somers 2006). As per EEM requirements, a test site was assessed as impaired if its BMI community fell outside of the normal range for *any one* of the EEM biological endpoints based on the 30 NNs.

RESULTS

PCA ordinations and site matching

Four nontrivial axes explained 57.8% of the variation for the YT data set (Table S1, Fig. 2A–C). A geography/climate gradient was evident on axis 1, with latitude and June temperature loading negatively and longitude and January snowfall and temperature loading positively. Climate and watershed-scale variables, including precipitation, June temperatures, stream order, and perimeter of the upstream watershed, loaded on axis 2. A climate/geology gradient involving June and total rainfall and sedimentary geology characterized axis 3. Axis 4 captured a watershed size gradient with drainage area, perimeter, stream length and wetted width loading on this axis. Sites were clustered in site-score plots, especially on plots of axes 1 and 2 (Fig. 2A) and of axes 1 and 3 (Fig. 2B), which were heavily loaded with climate/geography variables. The validation sites fell within the scatter of the training sites on all axes. For each of the validation sites, 30 NNs were selected to assess the BMI community.

The GL PCA (Table S2, Fig. 3A, B) had 3 nontrivial axes that explained 56.1% of the variation in the habitat. Axis 1 presented a nutrient/substrate size gradient with

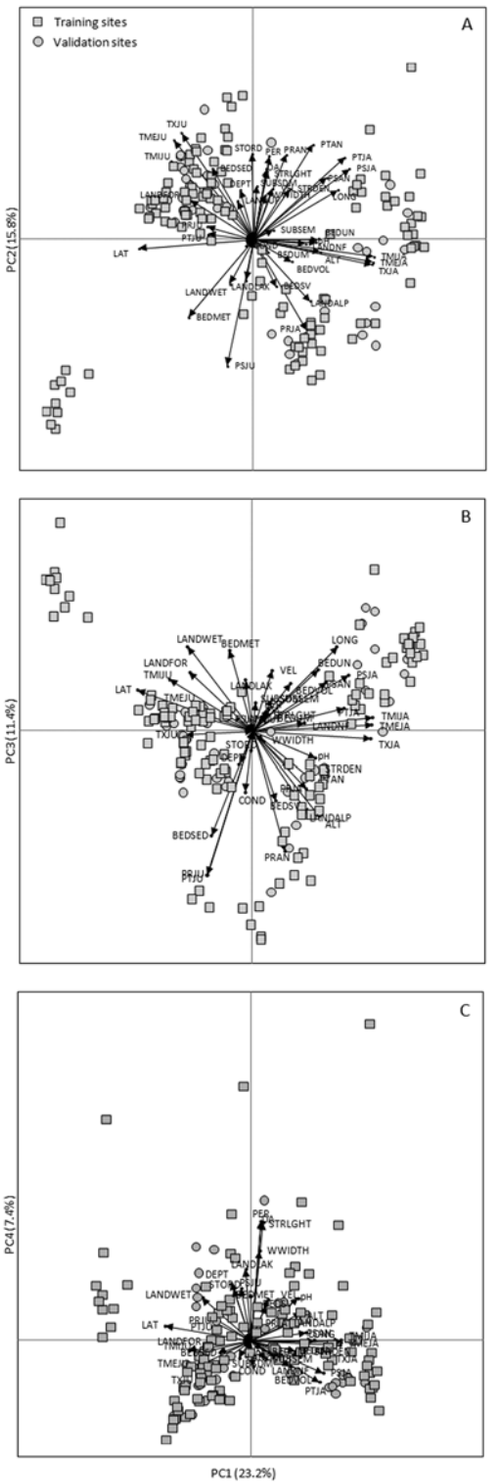

Figure 2. Yukon Territory habitat ordination plots showing site principal components analysis (PCA) scores for PC2 vs PC1 (A), PC3 vs PC1 (B), and PC4 vs PC1 (C). See Table S1 for loading scores and habitat variable codes. Numbers in parentheses are % variation explained by the axis. Vectors indicate strength of relationship of habitat variables with the PC axis.

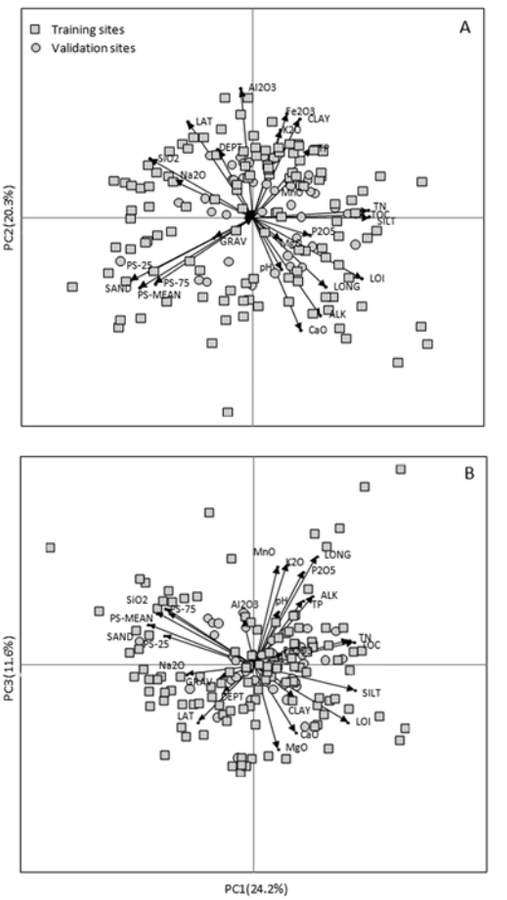

Figure 3. Laurentian Great Lakes habitat ordination plots showing site principal components analysis (PCA) scores for PC2 vs PC1 (A) and PC3 vs PC1 (B). See Table S2 for loading scores and habitat variable codes. Numbers in parentheses are % variation explained by the axis. Vectors indicate strength of relationship of habitat variables with the PC axis.

sediment nutrients and % silt loading positively and substrate size (% sand and particle size) loading negatively. On axis 2, latitude and Al/Ca gradients were evident with Al_2O_3, Fe_2O_3, and clay loading positively and alkalinity and CaO loading negatively. Axis 3 was characterized by longitude and sediment chemistry (P_2O_5, MnO, and K_2O). Plots of these PCA scores revealed a continuous distribution or single group of training and validation sites. All validation sites fell within the scatter for the training sites. Within this group, 30 NNs were selected for each validation site.

For the ACT data set (Table S3, Fig. 4A–C), 4 PCA axes explained 60.2% of the variance. The 1st axis captured a substrate/flow gradient with cobble, boulder, and water velocity loading positively and gravel loading negatively. Axis 2 reflected watershed size characteristics (stream order, distance from source, watershed area, bank width) and presence of bedrock. Axis 3 was a substrate gradient with peb-

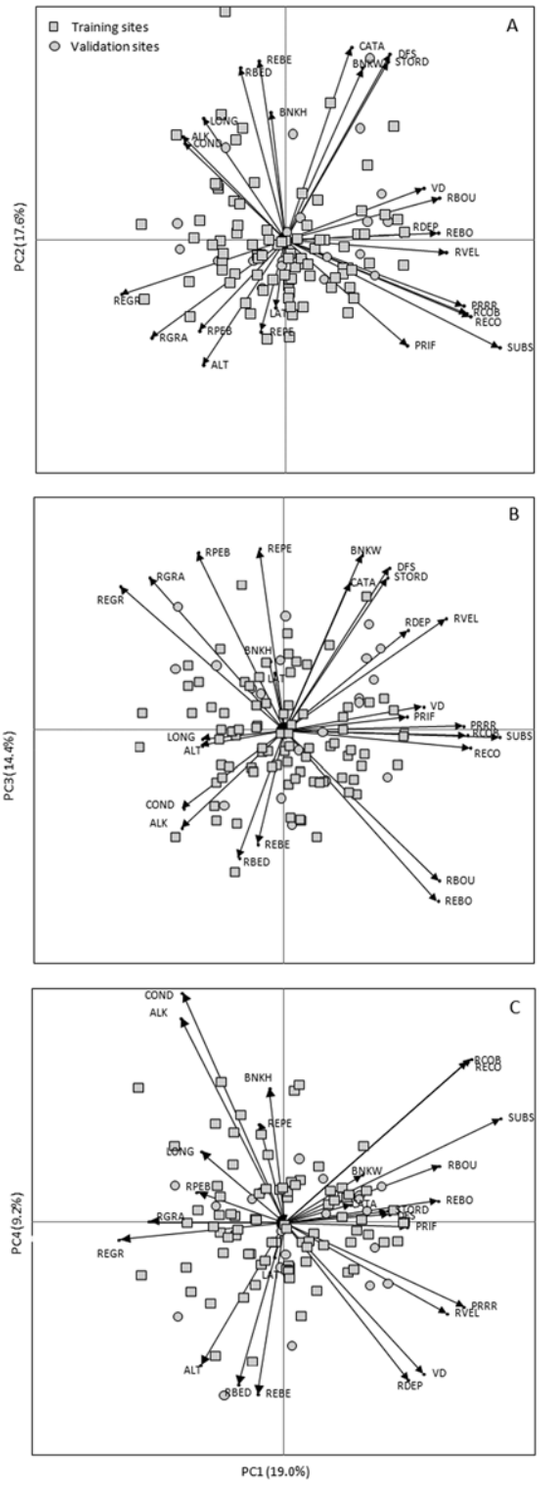

Figure 4. Australian Capital Territory habitat ordination plots showing site principal components analysis (PCA) scores for PC2 vs PC1 (A), PC3 vs PC1 (B), and PC4 vs PC1 (C). See Table S3 for loading scores and habitat variable codes. Numbers in parentheses are % variation explained by the axis. Vectors indicate strength of relationship of habitat variables with the PC axis.

ble and gravel loading positively and boulder loading negatively. Axis 4 was an alkalinity/conductivity gradient. The ACT plots revealed one large group of sites with a few validation sites falling outside most of the training sites especially on the positive end of axes 2 and 3 (i.e., sites with larger watershed areas and higher stream order). Thirty NN training sites were selected to assess each of the validation sites.

Community metrics

Total BMI density was highest at GL sites with a mean density of ~17,000 BMI/site, whereas mean densities were much lower at YT sites (714 BMI/site). YT densities were similar between D0 and D1 sites with a subsequent reduction at D2 and D3 sites (Fig. 5A). A slight increase in GL density was observed from the D0 and D1 sites with a gradual reduction in density for D2 and D3 sites (Fig. 5B). Variability in density was high among training sites and among categories of validation sites. YT and GL densities were organism counts extrapolated to the entire sample, but ACT density was number of BMIs subsampled (minimum 200). Mean density at ACT D0 sites was 228 BMIs with a steady decrease, as simulated impact severity increased, to a mean of 36 individuals at D3 sites (Fig. 5C).

The YT data set had 59 families and an average richness of 10 families/site (Fig. 5D). Fewer families were encountered in the GL data set with 54 families and an average richness of 8 families/site (Fig. 5E). Training-site richness was highest in the ACT data set with 67 families across all sites and an average richness of 18 families at individual sites (Fig. 5F).

Low evenness was found at ACT (0.29; Fig. 5I) and YT (0.35; Fig. 5G) D0 sites, whereas highest evenness was found at GL training sites (mean = 0.43; Fig. 5H). For all data sets, Simpson's evenness changed very little with simulated impact severity, except for an increase at D3 sites. Bray–Curtis distance was higher at YT (0.66; Fig. 5J) and GL (0.64; Fig. 5K) training sites and lowest at ACT training sites (0.43; Fig. 5L). Bray–Curtis distance increased with increasing simulated impact severity for all data sets.

TSA and model evaluation

The percentage of sites evaluated as impaired (% impaired) (noncentral F test, ES > normal range of 95% of NN) generally increased with simulated impact severity with ANNA and RDA models (Fig. 6A–C). A striking increase in % impaired with simulated impact severity was seen for ACT sites with both models (10% of D0 sites and 100% of D3 sites assessed as impaired; Fig. 6C). Percent impaired YT sites also increased with simulated impact severity (Fig. 6A), but did not approach the levels seen for ACT D3 sites. Percent impaired GL sites did not change with simulated impact severity and was ~10%

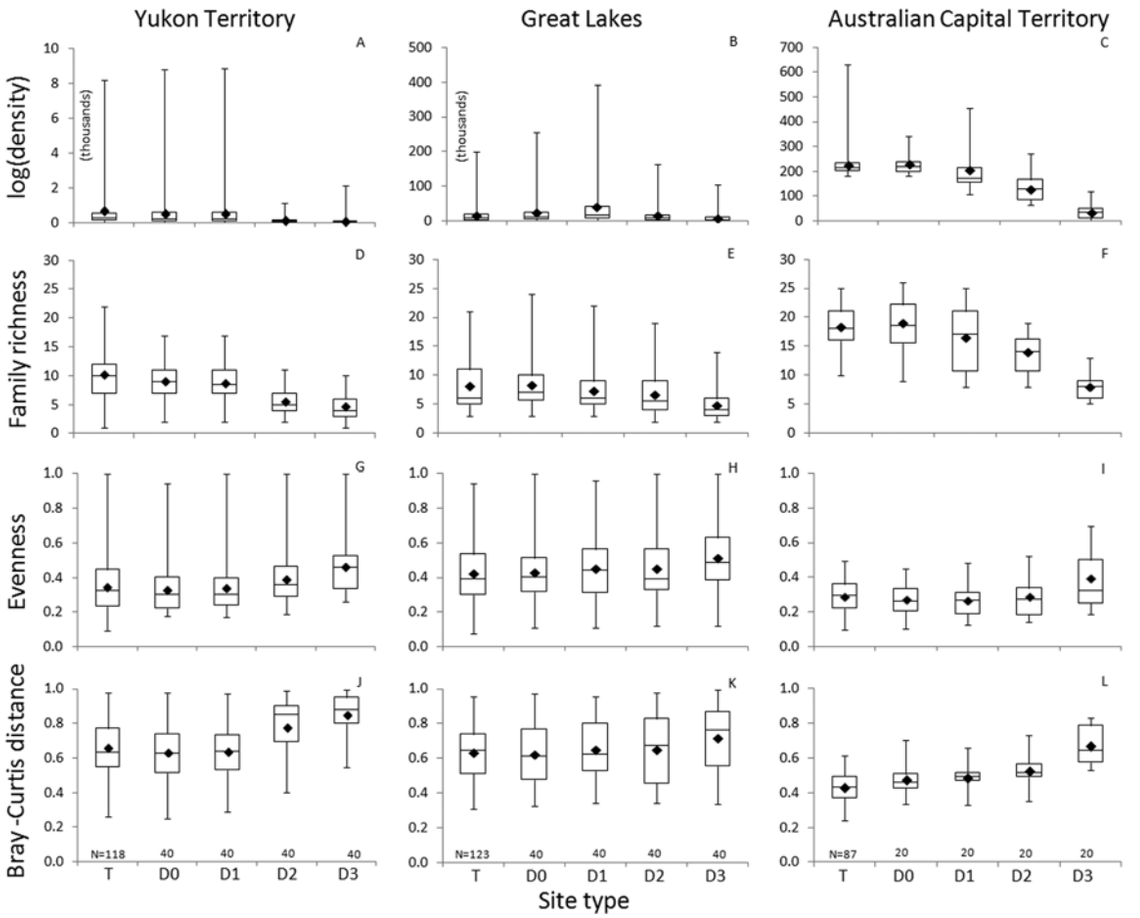

Figure 5. Box-and-whisker plots showing the relationship of benthic macroinvertebrate community density (A–C), family richness (D–F), evenness (G–I), and Bray–Curtis distance (J–L) to simulated impact severity for the Yukon Territory (A, D, G, J), Laurentian Great Lakes (B, E, H, K), and Australian Capital Territory (C, F, I, L) data sets for each site type. Diamonds are means, central horizontal lines are medians, box ends are quartiles, and whiskers show ranges. T = training site, D0 = unimpacted validation site, D1 = mildly impacted validation site, D2 = moderately impacted validation site, D3 = severely impacted validation site.

regardless of impact severity. For all data sets, RDA and ANNA models had low Type 1 errors that ranged from 5 to 12.5% of D0 sites assessed as impaired (Table 2). Type 2 errors (impacted sites assessed as equal to reference) were high in most cases, but Type 2 errors were low for ACT D2 and D3 sites.

ESs for each metric are presented as 1 : 1 plots for RDA and ANNA models (Fig. 7A–L). The closer a site falls to the line, the more similar the assessment by the 2 models. A reduction of BMI density with increasing simulated impact severity was observed in all 3 data sets, and D3 sites characteristically fell within the negative ES range (below the D0 site mean). Agreement between modeling methods was good for all data sets, but the relationship began to break down for D3 sites (negative ES) in the YT (Fig. 7A) and ACT (Fig. 7C) data sets. ANNA was slightly better than the RDA in assessing D3 sites as impaired in the YT data set. ESs were very large for the ACT data set. When simulated impacts were applied (especially at the D3 level), densities were as low as 2 individuals.

Richness decreased with simulated impact severity (i.e., at D3 sites with negative ESs) (Fig. 7D–F). ANNA and RDA models assessed sites equally, except that sites in the YT data set sometimes were assessed as more impaired by one of the approaches (i.e., more scatter around the 1 : 1 line). Evenness was assessed similarly by both models (Fig. 7G–I). However, the increase in evenness with simulated impact was unexpected. A severely impacted site (D3) typically would have low evenness indicated by a negative ES. Instead, D3 sites fell mostly at the positive end of the plots, whereas D0 sites fell on the negative end. Increasing Bray–Curtis distance (i.e., differences in communities indicating impairment) produced positive ESs (Fig. 7J–L). For Bray–Curtis distance, impacted sites generally were assessed differently by the 2 models, and the plots had greater scatter around the 1 : 1 line relative to the other metrics.

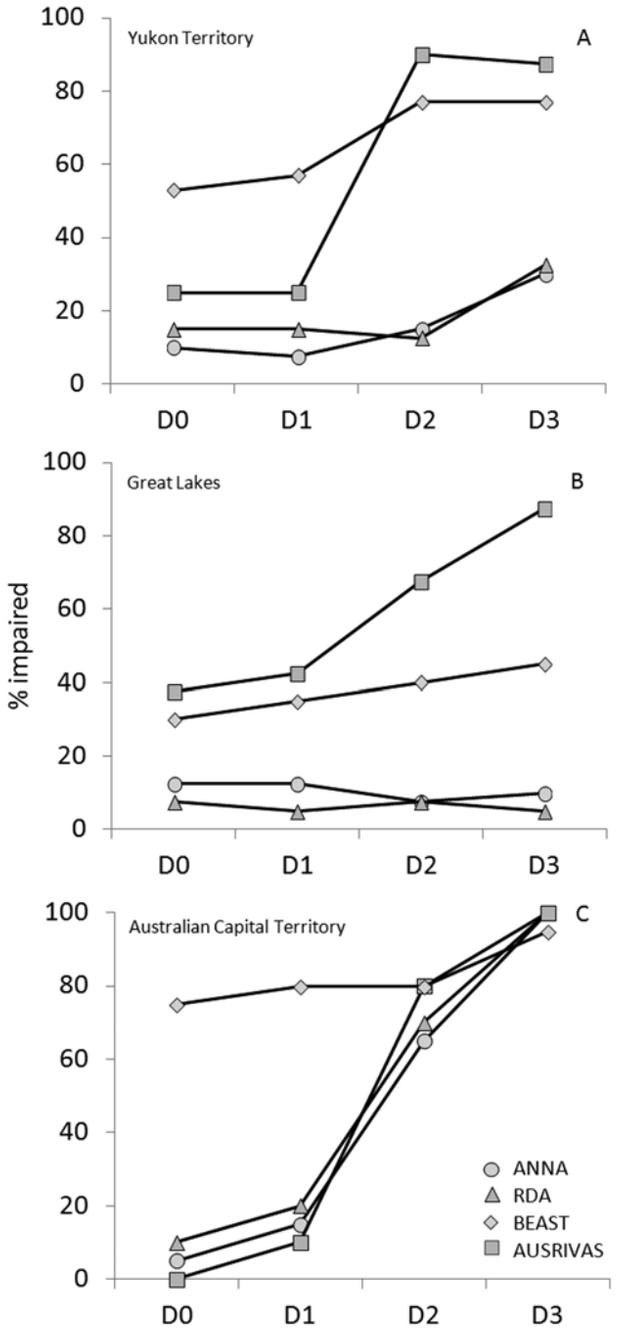

Figure 6. Percentage of sites assessed as impaired for Yukon Territory (A), Great Lakes (B), and Australian Capital Territory (C) using 4 modeling methods, Assessment by Nearest Neighbour Analysis (ANNA), Redundancy Analysis, (RDA) Benthic Assessment of Sediment (BEAST; data from Reynoldson et al. 2014), and Australian River Assessment System (AUSRIVAS; data from Nichols et al. 2014) for unimpacted (D0), mildly (D1), moderately (D2), and severely (D3) impacted validation sites.

Assessment results using ANNA and RDA were compared to the results of 2 national-program models (BEAST and AUSRIVAS) for the same 3 data sets. Percent impaired at each severity level were very similar for the ANNA, RDA, and AUSRIVAS models (Fig. 6A–C; Nichols et al. 2014), but the BEAST model assessed a larger percentage of D0 and D1 sites as impaired (Reynoldson et al. 2014) especially for the YT and ACT data sets. For the YT and GL data sets, more sites were assessed as impaired with increasing simulated impact severity with BEAST and AUSRIVAS, whereas fewer sites were assessed as impaired with increasing simulated impact severity with ANNA and RDA. Percent impaired sites in the GL data set did not increase with simulated impact severity with ANNA and RDA (~10% impaired regardless of the simulated impact severity).

Table 2. Percentage of sites incorrectly assessed using the Assessment of Nearest Neighbour Analysis (ANNA) and Redundancy Analysis (RDA) models, reported as Type 1 (validation sites assessed as impaired) and Type 2 (simulated impact sites assessed as equal to reference) errors.

		Type 1 (%)	Type 2 (%)		
		D0	D1	D2	D3
Australia (ACT)	ANNA	5	85	35	0
	RDA	10	80	30	0
Yukon Territory (YT)	ANNA	10	92.5	85	70
	RDA	15	85	87.5	67.5
Great Lakes (GL)	ANNA	12.5	87.5	92.5	90
	RDA	7.5	95	92.5	95

DISCUSSION

ACT training sites (n = 87) were distributed along a continuum of habitat without clustering (Fig. 4A–C), a pattern that allowed good site matching by the NN methods, ANNA and RDA. The responses to simulated impacts corresponded to increased % impaired when the ANNA and RDA models were applied (Fig. 6C). The rich benthic communities (18 families) with low variability resulted in simulated impact treatments that successfully mimicked actual impacts. As a consequence, impacted sites were classified correctly as impaired. The AUSRIVAS model performed equally well (Nichols et al. 2014) (Fig. 6C), but BEAST did not and >80% of sites were assessed as impaired regardless of simulated impact severity (Reynoldson et al. 2014).

YT training sites (n = 118) were clustered in 4 distinct groups along a habitat continuum (Fig. 2A–C), a pattern that can lead to poor site matching by NN methods (i.e., validation sites are compared to training sites that are somewhat dissimilar in habitat). Furthermore, this data set was unusual in that some training sites had quite low density (8–8200 BMIs, mean = 714 BMI/site) and richness (1–22, mean = 10 families/site). Poor site matching led to poor site assessments. For example, some D0 sites at the high end of the richness range were assessed as impaired, whereas their D2 and D3 counterparts were assessed as

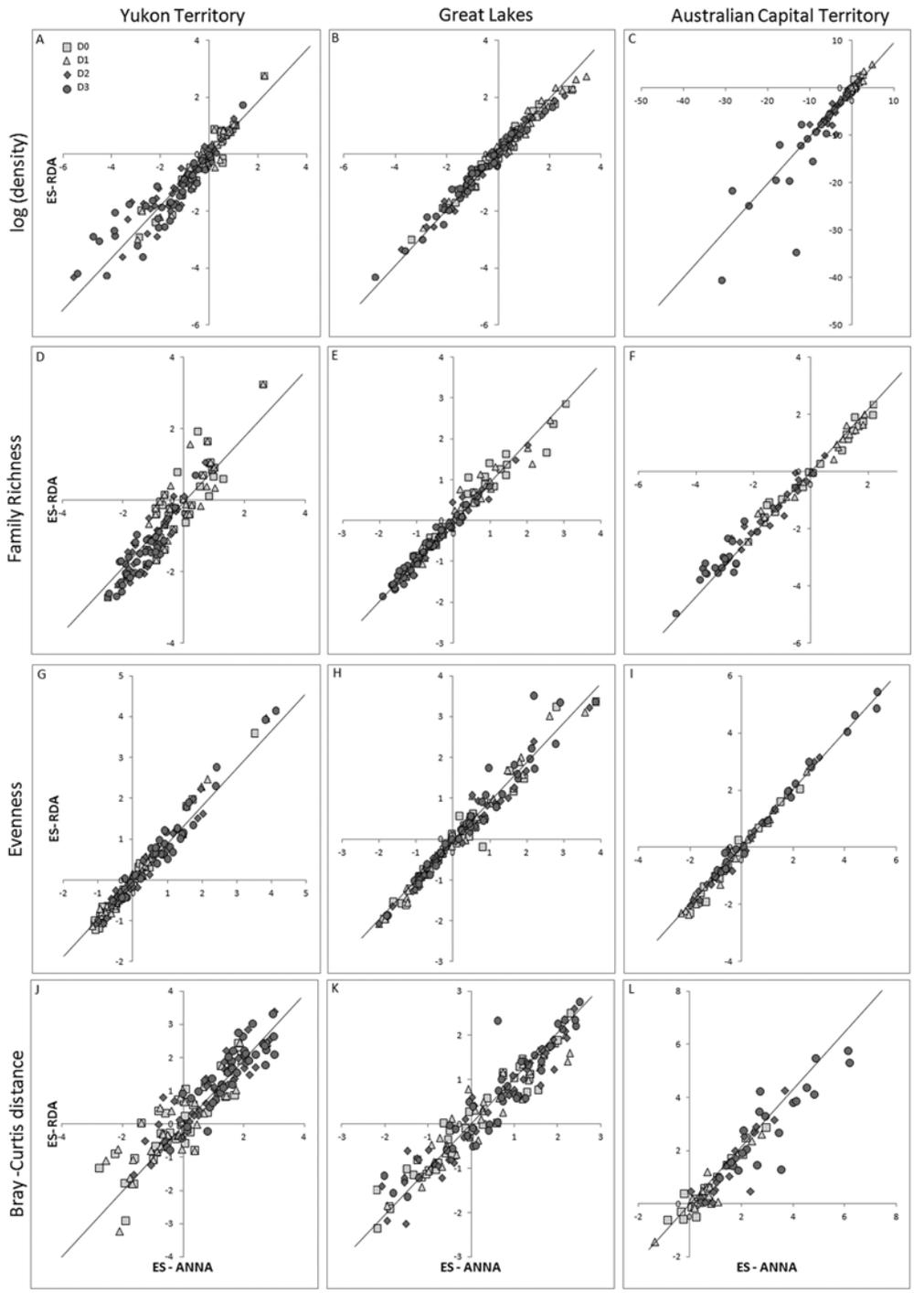

Figure 7. Effect size (ES; distance of validation metric value from mean value for 30 nearest neighbors) for benthic macroinvertebrate community density (A–C), family richness (D–F), evenness (G–I), and Bray–Curtis distance (J–L) for the Yukon Territory (A, D, G, J), Great Lakes (B, E, H, K), and Australian Capital Territory (C, F, I, L) data sets using 2 modeling methods (Redundancy Analysis [RDA] and Assessment by Nearest Neighbour Analysis [ANNA]). Symbols represent varying levels of simulated impacts (unimpacted, mildly, moderately, and severely impacted).

unimpaired even though higher richness is generally an indication of an unimpaired community. Density and evenness changed little over the range of simulated impact severity, and although richness and Bray–Curtis distance did show the expected response to simulated impacts, metric values overlapped considerably across site types (Fig. 5A, D, G, J), which made detection of impairment problematic. In fact, low percentages of impacted sites were assessed as impaired by both ANNA and RDA models (Fig. 6A, Table 2). Distinguishing D0 from D3 sites was difficult with the

BEAST model (Reynoldson et al. 2014), although % impaired sites did increase with simulated impact severity (Fig. 6A). In contrast, AUSRIVAS (Nichols et al. 2014) distinguished D0 and D1 sites from D2 and D3 sites (Fig. 6A) as expected.

GL training sites (n = 123) were distributed along a habitat continuum without clustering (Fig. 3A, B), allowing good site matching with ANNA and RDA. However, low-diversity communities at training sites and high variation in metric values (Fig. 5B, E, H, K) resulted in inability to distinguish effects of simulated impact severity in site assessments (Fig. 6B). GL training-site BMI communities were dominated by Oligochaeta (e.g., Tubificidae), Diptera (e.g., Chironomidae), and Dreissenidae even before simulated impacts were applied. The simulated impacts removed sensitive families first, so it had minimal effect on GL community metrics. Thus, detection of impairment by ANNA, RDA, and the BEAST model (Reynoldson et al. 2014) was poor. The AUSRIVAS model (Nichols et al. 2014) was able to detect impairment, but a higher than expected percentage of unimpacted (D0) sites was assessed as impaired.

Some modeling methods (e.g., BEAST and AUSRIVAS) are based on the assumption that distinct groups of sites exist. Groups were evident for the YT data set (Fig. 2A–C), and AUSRIVAS successfully distinguished YT impacted sites, as did BEAST to a lesser extent. However, often sites are distributed along a continuum of habitat variables, as seen in the ACT and GL data sets (Figs 3A, B, 4A–C). In this situation, models that do not assume clusters or discrete types of communities, such as ANNA and RDA, should be more sensitive than grouping models (Bowman and Somers 2005, Linke et al. 2005). The ANNA and RDA models successfully distinguished sites with varying simulated impacts for the ACT data set, which displayed a continuum or gradient of habitat conditions. Poorer performance with both the GL and YT data sets suggests that clustering of sites (e.g., YT) was not the only problem with the application of the NN approaches to these data sets. Subsample data were scaled up to full samples (in contrast to ACT), thereby increasing BMI density to levels at which simulated impacts were insufficient to create distinct differences. We think that simulated impacts were less apparent in the GL and YT data sets, and this situation led to correct classification of fewer impaired sites in these 2 data sets.

The response of the EEM metrics to enrichment is generally well known: density increases, richness decreases, evenness decreases, and Bray–Curtis distance increases (Rosenberg and Resh 1993, Mackie 2004). The enrichment impact simulated here successfully produced the expected trends (Fig. 5A–L), except that density was systematically reduced rather than increased as simulated impact severity increased (D0–D3) and evenness was slightly greater at D3 than at D0 sites. The outcome of this model-testing exercise depends on the successful application of simulated impacts to the data sets (e.g., Reynoldson et al. 1997), but an evaluation of the success of the simulated impacts is not presented here.

High variability in the BMI data also may have led to poor ability to distinguish impacted sites. Ranges presented in Fig. 5A–L are inflated compared to values used to evaluate a validation site for which only the metric values for a subset of 30 NN sites would be used. Nevertheless, high variability in these data sets may have contributed to our inability to distinguish impacted sites. Four levels of simulated impacts should have been distinct with respect to all metrics, with little to no overlap in their ranges. Somewhat distinct metric values were achieved for the ACT data set (Fig. 5C, F, I, L), leading to successful separation of impact levels (Fig. 6C). However, the simulation does not appear to have successfully created distinct levels of impact for the GL data set and, to a lesser extent, the YT data set. The range of the GL metrics overlapped considerably and only subtle changes were seen in metric values as simulated impact severity increased from D0 to D3 (Fig. 5B, E, H, K). Therefore, it is not surprising that the models did not tease apart the impact levels for the GL data set (Fig. 6B). Furthermore, YT and GL training sites generally had low richness before simulated impacts were applied. Low richness at training sites led to assessment of some D0 and D1 sites as impaired because of higher than expected richness at these validation sites, whereas the corresponding D2 and D3 sites were assessed as unimpaired.

The objective of our study was to identify which bioassessment method was best at detecting simulated impacts. ANNA and RDA produced similar results that agreed with our expectations for the ACT data set with a habitat continuum and a rich BMI community. With the YT and GL data sets, ANNA and RDA produced results that corresponded only weakly to the simulated impact gradient. We think that the simulated impacts did not sufficiently alter BMI communities with naturally low diversity and high variability (YT streams) or dominated by enrichment-tolerant taxa (GL nearshore sites). Thus, assessment results were data-set specific and may depend on the metrics that were used. We used metrics specified by the Canadian federal EEM program, but these metrics did not respond to the simulated impacts as expected. Given these results, the best method depends on the particular scenario, and use of several different methods might be wise when conducting a bioassessment. Future comparative analyses should include additional metrics (e.g., correspondence analysis scores) and multivariate assessments (e.g., D in TSA), on both real and simulated data.

ACKNOWLEDGEMENTS

We thank the Yukon Government, Department of Fisheries and Oceans Canada, and University of Western Ontario for gra-

ciously providing the Yukon data set. Permission to use the Australian Capital Territory data set was given by Environment and Sustainable Development Directorate, ACT Government, Canberra, Australia. L. Grapentine of Environment Canada supplied the Great Lakes data set. We extend special thanks to T. Reynoldson for coordinating the data set distribution, GHOST Environmental Consulting, Inc. for financial support, and K. Fram for technical support. We also thank M. Bowman, C. Jones, B. Keller, B. Kilgour, and M. White for discussions on implementation of ANNA and RDA methods.

LITERATURE CITED

Bailey, R. C., M. G. Kennedy, M. Z. Dervish, and R. M. Taylor. 1998. Biological assessment of freshwater ecosystems using a reference condition approach: comparing predicted and actual benthic invertebrate communities in Yukon streams. Freshwater Biology 39:765–774.

Bailey, R. C., S. Linke, 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.

Bailey, R. C., R. H. Norris, and T. B. Reynoldson. 2004. Bioassessment of freshwater ecosystems using the reference condition approach. Kluwer Academic Publishers, Boston, Massachusetts.

Bowman, M. F., and K. M. Somers. 2005. Considerations when using the reference condition approach for bioassessment of freshwater ecosystems. Water Quality Research Journal of Canada 40:347–360.

Bowman, M. F., and K. M. Somers. 2006. Evaluating a novel Test Site Analysis (TSA) bioassessment approach. Journal of the North American Benthological Society 25:712–727.

Bowman, M. F., K. M. Somers, and R. A. Reid. 2003. A simple method to evaluate whether a biological community has been influenced by anthropogenic activity. Pages 62–72 *in* K. Hedley, S. Roe, and A. J. Niimi (editors). Proceedings of the 30th Annual Aquatic Toxicology Workshop: September 28 to October 1, 2003, Ottawa, ON. Canadian Technical Reports of Fisheries and Aquatic Sciences 2510:62–72.

Chessman, B. C., J. E. Growns, and A. R. Kotlash. 1997. Objective derivation of macroinvertebrate family sensitivity grade numbers for the SIGNAL biotic index: application to the Hunter River system, New South Wales. Marine and Freshwater Research 48:159–172.

Glozier, N. E., J. M. Culp, T. B. Reynoldson, R. C. Bailey, R. B. Lowell, and L. Trudel. 2002. Assessing metal mine effects using benthic invertebrates for Canada's environmental effects program. Water Quality Research Journal of Canada 37:251–278.

Hilsenhoff, W. L. 1988. Rapid field assessment of organic pollution with a family-level biotic index. Journal of the North American Benthological Society 7:65–68.

Jackson, D. A. 1993. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecology 74:2204–2214.

Jones, C., K. M. Somers, B. Craig, and T. B. Reynoldson. 2005. Ontario Benthos Biomonitoring Network: protocol manual. Ontario Ministry of Environment, Dorset Environmental Science Centre, Dorset, Ontario. (Available from: www.saugeenconservation.com)

Kilgour, B. W., K. M. Somers, and D. E. Matthews. 1998. Using the normal range as a criterion for ecological significance in environmental monitoring and assessment. Ecoscience 5:542–550.

Legendre, P., and L. Legendre. 1998. Numerical ecology. 2nd English edition. Elsevier, Amsterdam, The Netherlands.

Lenth, R. V. 2003. πface for Excel. Department of Statistics and Actuarial Science, University of Iowa, Iowa City, Iowa. (Available from: http://www.stat.uiowa.edu/~rlenth/)

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

Lipkovich, I., and E. P. Smith. 2001. Biplot and singular value decomposition macros for Excel. Journal of Statistical Software 7(5):1–13.

Mackie, G. L. 2004. Applied aquatic ecosystem concepts. 2nd edition. Kendall/Hunt, Dubuque, Iowa.

Nichols, S. J., T. B. Reynoldson, and E. T. Harrison. 2014. Evaluating AUSRIVAS predictive model performance for detecting simulated eutrophication effects on invertebrate assemblages. Freshwater Science 33:1249–1260.

Omernik, J. M. 1995. Ecoregions: a spatial framework for environmental management. Pages 49–62 *in* W. S. Davis and T. P. Simon (editors). Biological assessment and criteria. Tools for water resource planning and decision making. Lewis Publishers, Boca Raton, Florida.

Reynoldson, T. B., R. C. Bailey, K. E. Day, and R. H. Norris. 1995. Biological guidelines for freshwater sediments 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., C. I. Brereton, W. Keller, and C. Sarrazin-Delay. 2005. Development of a northern Ontario benthic invertebrate reference condition approach (RCA) biomonitoring network to meet metal mining environmental effects monitoring requirements. Phase One. Cooperative Freshwater Ecology Unit, Department of Biology, Laurentian University, Sudbury, Ontario. (Available from: www3.laurentian.ca/livingwithlakes/)

Reynoldson, T. B., R. H. Norris, V. H. Resh, K. E. Day, and D. M. Rosenberg. 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.

Reynoldson, T. B., S. Strachan, and J. L. Bailey. 2014. A tiered method for discriminant function analysis models for the Reference Condition Approach: model performance and assessment. Freshwater Science 33:1238–1248.

Rosenberg, D. M., and V. H. Resh. 1993. Introduction to freshwater biomonitoring and benthic macroinvertebrates. Pages 1–9 *in* D. M. Rosenberg and V. H. Resh (editors). Freshwater biomonitoring and benthic macroinvertebrates. Chapman and Hall, New York.

Sarrazin-Delay, C. L., M. F. Bowman, W. Keller, and K. M. Somers. 2006. Summary of the model-building results for the northern Ontario benthic invertebrate RCA biomonitoring initiative: 2003–2006. Cooperative Freshwater Ecology Unit, Department of Biology, Laurentian University, Sudbury, Ontario. (Available from: www3.laurentian.ca/livingwithlakes/)

Simon, T. P. 1991. Development of biological integrity expectations for the ecoregions of Indiana. I. Central Corn Belt Plain. EPA 905/9-91/025. Region V, Environmental Sciences Division, Monitoring and Quality Assurance, US Environmental Protection Agency, Chicago, Illinois.

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 fresh water: RIVPACS and other techniques. Freshwater Biological Association, Cumbria, UK.

Verdonschot, P. F. M. 1995. Typology of macrofaunal assemblages: a tool for the management of running waters in The Netherlands. Hydrobiologia 297:99–122.

Walker, S. L., S. C. Ribey, L. Trudel, and E. Porter. 2003. Canadian environmental effects monitoring: experiences with pulp and paper and metal mining regulatory programs. Environmental Monitoring and Assessment 88:311–326.