Ecological Indicators 10 (2010) 527–537

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

The Index of Biological Integrity and the bootstrap: Can random sampling error affect stream impairment decisions?

Christine L. Dolph [a,*], Aleksey Y. Sheshukov [b], Christopher J. Chizinski [c], Bruce Vondracek [d], Bruce Wilson [e]

[a] Water Resources Science Program, University of Minnesota, 200 Hodson Hall, 1980 Folwell Ave, University of Minnesota, St. Paul, MN, 55108, USA
[b] Department of Biological and Agricultural Engineering, 49 Seaton Hall, Kansas State University, Manhattan, KS, 66506, USA
[c] Department of Fisheries, Wildlife, and Conservation Biology, University of Minnesota, 1980 Folwell Ave, St. Paul, MN, 55108, USA
[d] U.S. Geological Survey, Minnesota Fish and Wildlife Cooperative Research Unit[1], 200 Hodson Hall, 1980 Folwell Ave, University of Minnesota, St. Paul, MN, 55108, USA
[e] Department of Bioproducts and Biosystems Engineering, 1390 Eckles Avenue, University of Minnesota, St. Paul, MN, 55108, USA

ARTICLE INFO

Article history:
Received 3 April 2009
Received in revised form 29 September 2009
Accepted 1 October 2009

Keywords:
Bioassessment
Multimetric
Variability
Nonparametric
Streams
Fish
Rare

ABSTRACT

Multimetric indices, such as the Index of Biological Integrity (IBI), are increasingly used by management agencies to determine whether surface water quality is impaired. However, important questions about the variability of these indices have not been thoroughly addressed in the scientific literature. In this study, we used a bootstrap approach to quantify variability associated with fish IBIs developed for streams in two Minnesota river basins. We further placed this variability into a management context by comparing it to impairment thresholds currently used in water quality determinations for Minnesota streams. We found that 95% confidence intervals ranged as high as 40 points for IBIs scored on a 0–100 point scale. However, on average, 90% of IBI scores calculated from bootstrap replicate samples for a given stream site yielded the same impairment status as the original IBI score. We suggest that sampling variability in IBI scores is related to both the number of fish and the number of rare taxa in a field collection. A comparison of the effects of different scoring methods on IBI variability indicates that a continuous scoring method may reduce the amount of bias in IBI scores.

1. Introduction

Biological indicators are now an integral part of water quality monitoring programs worldwide (Furse et al., 2006; Yagow et al., 2006; Marchant et al., 2006; Borja et al., 2008). In the United States, multimetric indices such as the Index of Biological Integrity (IBI) have been widely adopted by water management agencies as the primary tool for assessing the biological condition of streams and lakes (Karr and Chu, 1999; EPA, 2002). These IBIs consist of a suite of metrics that reflect the taxonomic composition, trophic relationships, and abundance and condition of organisms within an aquatic community, and thus aim to convey an integrated picture of ecosystem health (Karr and Yoder, 2004).

Increasingly, resource managers use IBIs to assess whether surface waters fulfill the requirements of their designated uses under the Clean Water Act (EPA, 2002). If IBI scores do not meet expected criteria, streams can be added to the 303(d) impaired waters list and scheduled for subsequent management action under the Total Maximum Daily Load (TMDL) program. In Minnesota, for example, 10% of surface waters designated as impaired in 2008 were listed primarily on the basis of poor IBI scores, in conjunction with other relevant habitat, water chemistry, and catchment data (MPCA, 2008). The increasing prominence of biological indices in management decisions warrants careful evaluation of their strengths and limitations, particularly regarding their susceptibility to error.

Whereas the application of IBIs to new geographic contexts and scales continues to receive broad attention in the scientific literature (e.g., Stoddard et al., 2008; Wang et al., 2008; Pinto et al., 2009), only a few studies have directed attention to the inherent statistical properties of these indices—i.e., their precision and accuracy, as well as their sensitivity to variation in biological samples (Fore et al., 1994; Carlisle and Clements, 1999; Blocksom, 2003). The best known of these efforts was undertaken by Fore et al. (1994), who used a bootstrapping approach to quantify the response of IBIs to random sampling variation. Bootstrapping, originally described by Efron (1979, 2003), is a computer-intensive statistical technique used to estimate variability of a statistic when the actual distribution is unknown, such as when that statistic is determined from a single random sample.

* Corresponding author. Tel.: +1 612 868 1565.
E-mail addresses: dolph008@umn.edu (C.L. Dolph), ashesh@ksu.edu (A.Y. Sheshukov), chizi001@umn.edu (C.J. Chizinski), bvondrac@umn.edu (B. Vondracek), wilson@umn.edu (B. Wilson).

[1] The Unit is jointly sponsored by the U.S. Geological Survey, the University of Minnesota, the Minnesota Department of Natural Resources, the U.S. Fish and Wildlife Service, and the Wildlife Management Institute.

1470-160X/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ecolind.2009.10.001

We believe the application of bootstrapping to the problem of IBI variability warrants further consideration for a number of reasons. First, the increased availability of large biomonitoring datasets and advances in computing processing speed allow for a more robust estimate of index variability. Moreover, the metrics included in a multimetric index, and the way in which those metrics are scored, may vary depending upon the unique geographic location to which an index is applied (Simon and Lyons, 1995). The quantification of variability associated with an alternative set of IBIs (i.e., IBIs designed for Minnesota streams) can therefore reveal new information about the ways in which these indices may respond to variability in biological data. Finally, given that IBIs are increasingly relied upon to determine whether site water quality is impaired, estimates of IBI variability need to be examined in an updated context that considers the current criteria upon which such impairment decisions are made.

The broad goals of this research were to assess (1) whether fish IBI scores used in Minnesota can provide a consistent indication of stream quality, given random sampling variability; and (2) whether the IBI score represents an appropriate policy tool on which to base yes/no decisions about stream impairment. Specifically, we sought to address the following research questions: What is the range of possible IBI scores that a stream site may receive given random sampling errors in fish samples (i.e., how sensitive is the IBI to these errors)? Are stream impairment decisions based on IBI scores likely to change as the result of these errors? What aspects of fish samples might be related to IBI sensitivity? Are IBIs for some types of streams more sensitive than others? Do different scoring systems for component metrics of the IBI improve index accuracy (i.e., reduce index bias)? How does random sampling error in IBI scores compare to IBI variability over time? To evaluate these questions, we used bootstrapping to mimic the effects of random sampling on a set of IBIs developed by the Minnesota Pollution Control Agency (MPCA) for fish communities in the St. Croix and Upper Mississippi River basins, Minnesota. We then quantified the IBI's sensitivity to random sampling errors, and demonstrated how this sensitivity relates to IBI-based impairment decisions for Minnesota streams. Finally, we

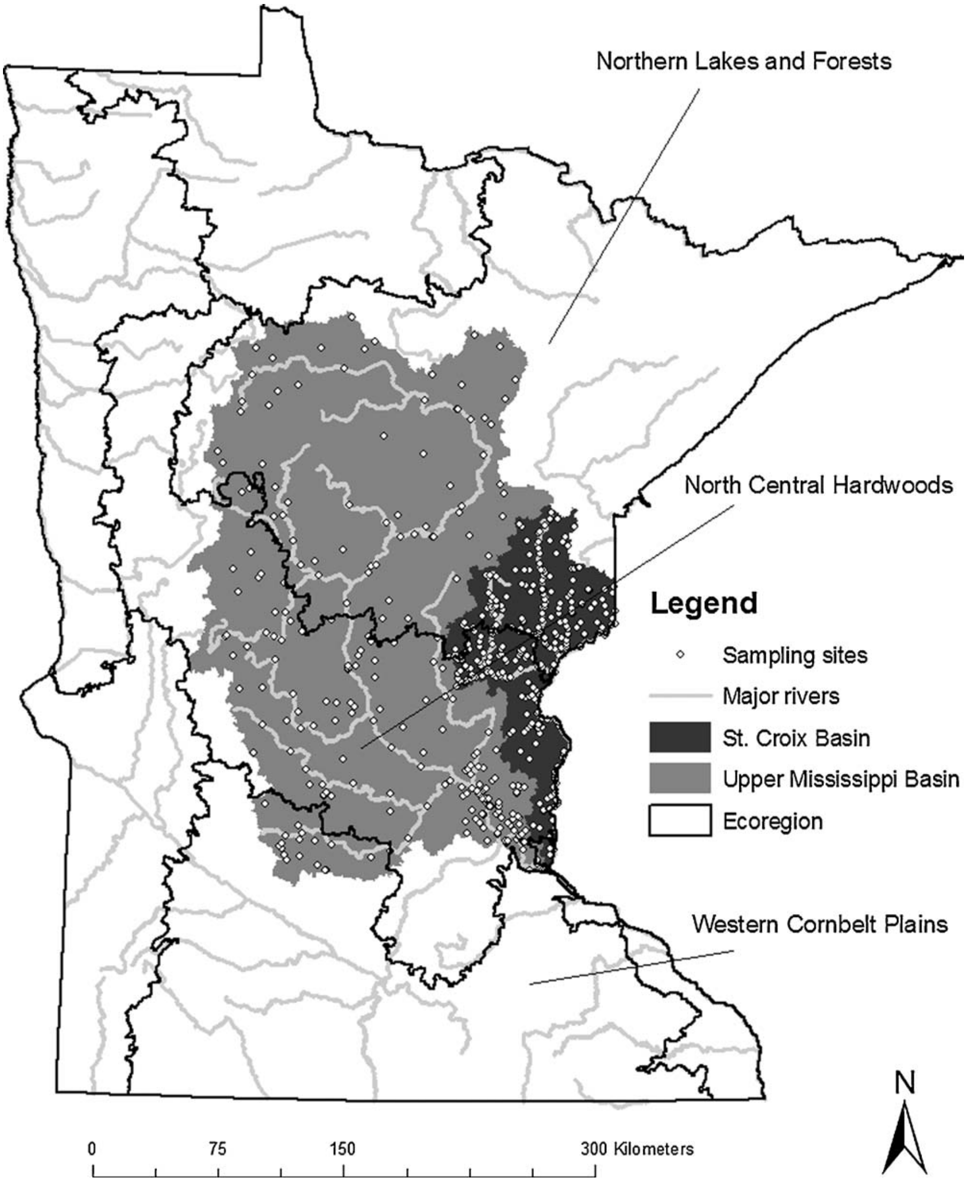

Fig. 1. Stream sites in the Upper Mississippi ($n = 181$) and St. Croix ($n = 207$) river basins from which quantitative fish samples were collected by the Minnesota Pollution Control Agency between 1996 and 2006. In some cases, more than one sample was collected from a stream site over time, yielding a total of 513 samples across all 388 stream sites.

Table 1
Variables used by the MPCA to classify warmwater stream sites. Ecoregion was used only to further classify large rivers sites (drainage area >691 km^2) within the St. Croix basin.

Stream class	Major drainage basin	Drainage area size (km^2)	Stream size class	Ecoregion	# site visits used in this study
1	St. Croix	<51	Headwater		77
2	St. Croix	51–138	Small		86
3	St. Croix	138–691	Moderate		73
4	St. Croix	>691	River	Northern Lakes and Forests	32
5	St. Croix	>691	River	North Central Hardwoods	35
6	Upper Mississippi	<13	Headwater		25
7	Upper Mississippi	13–90	Small		70
8	Upper Mississippi	90–512	Moderate		57
9	Upper Mississippi	>512	River		58

identified specific characteristics of fish samples and of the IBI itself that correlated with IBI sensitivity, including whether continuous or discontinuous methods were used to score the IBI's component metrics.

2. Methods and materials

2.1. Study sites and data collection

We used fish IBI scores from stream sites in two of Minnesota's major river basins (Fig. 1): the St. Croix (n = 303 site visits at 207 unique sites) and Upper Mississippi (n = 210 site visits at 181 unique sites). Stream sites included in this study spanned a broad range of IBI scores and drainage size classes, as well as three different ecoregions: Northern Lakes and Forests (NLF), North Central Hardwoods (NCH), and Western Cornbelt Plains (WCBP).

The IBIs used to assess the health of fish communities in these basins were developed by the MPCA as part of the state's biological monitoring program (Niemela and Feist, 2000, 2002). These IBIs were developed exclusively for warmwater streams; thus, no coldwater streams are included in our analysis. To stratify natural variability across warmwater streams, the MPCA classified these stream sites by major river basin, drainage area size, and ecoregion, resulting in 9 separate stream classes (Table 1). Drainage area size classes are closely related to stream order and discharge (Allan and Castillo, 2008), and were chosen to minimize differences in maximum species richness among headwater, small, moderate, and large river systems (Table 1; Niemela and Feist, 2000, 2002).

Separate IBIs were developed for each stream class, with each IBI comprised of a slightly different set of metrics and metric scoring criteria (Table 2). The only exceptions were classes 3 and 4 IBIs, which used the same metrics and scoring criteria, and class 5 IBIs, which also used the same set of metrics as classes 3 and 4, but slightly different scoring criteria for two of the 10 component metrics (total no. of species and no. of sensitive species). For class 5 streams, higher numbers of total and sensitive species were required to receive a given metric score, relative to classes 3 and 4 streams. For example, 29 or more species were required for the total no. of species metric to receive a score of 10 in class 5 streams, whereas only 23 species were required for this metric to score a 10 in classes 3 and 4 streams.

All individual metric scores ranged from 0 to 10; however, some metrics were scored using five categories (0, 2, 5, 7, 10), whereas others were scored using only three (0, 5, 10). One metric (no. of individuals per meter) was scored using only two categories (0 or 10). IBIs in all classes had a possible range of 0–100, and most IBIs consisted of 10 metrics. Where IBIs consisted of fewer than 10 metrics (stream classes 1, 2, and 6), the sum of metric scores was normalized to a 100 point scale (Niemela and Feist, 2000, 2002).

Fish samples used in IBI calculations were collected by the MPCA during summer low-flow conditions from 1996 to 2006. Fish were captured by electrofishing, and were identified to the lowest possible taxonomic level (typically species). Four types of electrofishing equipment were used to collect fish, with selection of equipment type dependent on stream size and accessibility (Niemela and Feist, 2000, 2002). In small, wadeable streams (<8 m wide), a backpack electrofisher was used to sample streams in an upstream direction, with one person carrying the electrofishing gear and one person collecting fish with a dip net. Where possible, almost all of the available habitat was sampled, but in streams >3 m wide, the sampling crew weaved among habitat types. In larger wadeable streams (>8 m wide), a stream electrofisher was used, in which a generator and control box were secured in a canoe, attached to two anodes. The canoe was pulled upstream by one person, two people deployed the anodes, and two people dip-netted for fish, again weaving among available habitat types. In small but unwadeable streams, a mini-boom electrofisher was used to sample streams. This unit consisted of a generator and control box placed into a small jon-boat, and connected to a single anode. A driver directed the boat downstream and weaved through different habitat types as a single person dip-netted for fish from the bow. Finally, a boom electrofisher was used to sample large, unwadeable rivers. Two crew members collected fish with dipnets from the bow, as the boat driver conducted three separate sampling passes in a downstream direction: one along each shoreline and one along the middle of the channel.

Table 2
Metrics used by the MPCA to calculate the fish Index of Biological Integrity (IBI) for warmwater streams in the St. Croix and Upper Mississippi River basins.

Metric	Anticipated response to disturbance	Stream classes using metric in IBI score
Total no. of species	Decrease	All classes
No. of benthic invertivore species	Decrease	2–5
No. of darter species	Decrease	3–5
No. of darter, sculpin, and madtom species	Decrease	8–9
No. of invertivore species[a]	Decrease	1; 6–9
No. of minnow species[a]	Decrease	1–2; 7
No. of omnivore species	Increase	3–5
No. of sensitive species	Decrease	2–5; 7–9
No. of wetland species[a]	Decrease	6–8
No. headwater species[a]	Decrease	1
% of total abundance comprised of individuals of the two most abundant taxa	Increase	1–2; 6–7
% of individuals with deformities, lesion, or tumors	Increase	All classes
% of individuals classified as omnivore species	Increase	9
% of individuals classified as lithophilic spawners	Decrease	1–5; 7–9
% of individuals classified as piscivores	Decrease	3–5; 8–9
% of individuals classified as tolerant	Increase	All classes
No. of individuals per meter[a]	Decrease	All classes

[a] Metrics exclude species that are considered tolerant to disturbance.

2.2. Bootstrapping

Bootstrapping creates replicate samples from a single sample by randomly resampling from the original sample with replacement (Chernick, 1999; Efron, 2003; Manly, 2007). In the case of a fish sample from a single stream site, for example, the bootstrapping algorithm randomly selects one individual specimen at a time, adds it to a new replicate sample, and replaces it in the original sample, where it again has the same probability of being selected as any other specimen. Resampling is repeated in this manner until the number of individuals in the replicate sample is equivalent to the number of individuals in the original sample. The result is a series of samples that contain ''collections of fish that could have been caught at the same site and time by electrofishing'' (Fore et al., 1994), differing only by random variation. An IBI score can be determined for each individual bootstrap replicate sample, and the mean and variance can be calculated across all bootstrap replicates for a given site visit. The end result is the range of possible IBI scores that a stream site could receive, given variability in the fish collection that may arise from random sampling effects.

In this study, we created 1000 bootstrap samples for each original fish sample using the statistical software R. This number of replicates is generally considered sufficient for confidence interval generation (Chernick, 1999; Carpenter and Bithell, 2000). Of the many methods available for determining confidence intervals from bootstrap data (Carpenter and Bithell, 2000), we used the percentile method based on simplicity of use and to enable comparisons with an earlier study by Fore et al. (1994). The primary disadvantage of the percentile method is poor performance with small samples and asymmetric distributions (Chernick, 1999).

To estimate a 95% percentile confidence interval for an IBI score at a given stream site, we first sorted IBI scores from the replicate bootstrap samples into ascending order. For 1000 replicate scores, the 25th ordered value represents the lower bound of the confidence interval, and the 975th ordered value represents the upper bound (Carpenter and Bithell, 2000). Confidence interval length was determined by subtracting the lower bound from the upper bound value.

2.3. Implications of variability for impairment status

The MPCA derives impairment thresholds for IBI scores from a set of reference sites (i.e., sites that are relatively unimpacted by human disturbance) in each stream class. The lowest IBI score in the range of all IBI scores measured at reference sites in a given stream class is taken as the impairment threshold for that class (MPCA, 2007). The MPCA also designates a confidence region around the impairment threshold (approximately ±9 points for sites in the St. Croix River basin and ±13 points for sites in the Upper Mississippi River basin) that is based on variability of IBI scores at least impacted (i.e., reference) sites over time (MPCA, 2007). IBI scores falling above and below this confidence region are considered 'unimpaired' and 'impaired', respectively; IBI scores falling within the confidence region are considered 'potentially impaired', with additional evidence needed to verify status.

In this study, our objective was to evaluate whether stream impairment decisions based on IBI scores are likely to change as the result of random sampling error. We therefore determined how many of the 1000 IBI scores generated for each site visit indicated a different impairment status ('impaired', 'potentially impaired', or 'unimpaired') than the original IBI score. This number was divided by 1000 to calculate the proportion of bootstrap scores that diverged from the original score for a given site visit. Finally, these proportions were averaged across all 513 site visits.

2.4. Identifying covariates of IBI sensitivity

In addition to quantifying the IBIs' sensitivity to random sampling errors, we sought to identify aspects of fish samples that were related to this sensitivity. In particular, we sought to investigate whether IBI sensitivity was affected by aspects of community abundance or richness. We used simple linear regression to assess the relationship between confidence interval length and several possible covariates, including the total number of fish in a sample (i.e., sample size), the total number of taxa in a sample (i.e., species richness), Pielou's evenness, and the number of species that occurred as single individuals in a sample (i.e., the number of singletons). Pielou's evenness was calculated using the vegan: community ecology package in R (Oksanen et al., 2009). By including the number of singletons as a covariate, we sought to capture the effect of rare species on IBI variability. Using the segmented package in R (Muggeo, 2008), we also conducted breakpoint regression analysis (Muggeo, 2003) to further identify whether there were any abrupt changes in IBI sensitivity with increasing sample size. Finally, we used analysis of variance (ANOVA) to determine whether the number of singletons varied among stream classes.

2.5. Comparative performance of IBIs

We evaluated whether IBI confidence interval length varied among the nine different stream classes using ANOVA followed by Dunnett's modified Tukey–Kramer (DTK) multiple comparison test (Dunnett, 1980), to determine whether some combinations of metrics were more sensitive to random sampling error than others. The DTK test was selected because it is appropriate for testing multiple pairwise comparisons when sample sizes or sample variances are unequal, and was applied using the DTK package in R, with $\alpha = 0.05$ (Lau, 2009).

2.6. Continuous scoring of metrics and bias among IBI scores

Component metrics of the fish IBIs used in this study are currently scored in a discontinuous fashion; that is, metric values typically receive a discrete score of 0, 2, 5, 7, or 10. This type of scoring method was originally proposed by Karr (1981), and has been incorporated into many subsequent IBIs. However, Fore et al. (1994) suggested that a discontinuous scoring system resulted in bias among IBI scores (i.e., in scores that underestimate the integrity of high quality sites and overestimate the quality of poor quality sites). In this study, we developed a continuous scoring method for IBI metrics, and evaluated whether the use of this method reduced bias among replicate IBI scores compared to the discontinuous method used by the MPCA.

To score metrics continuously, a linear piecewise polynomial was defined in which continuous scores were anchored to the discontinuous discrete scores at the midpoint of the metric values for a given discrete score (i.e., continuous and discrete scores were the same at this midpoint value), and were everywhere continuous (de Boor, 1978). We then recalculated IBI scores for each bootstrap replicate sample using this new scoring system. Bias associated with the discontinuous and continuous scoring methods was determined for each stream site by subtracting the resulting IBI score calculated from the original sample from the mean IBI score calculated across the 1000 bootstrap replicate samples. Paired two-sided *t*-tests were used to evaluate differences between mean bootstrap IBI scores and original scores.

2.7. Random sampling error vs. variability over time

In the St. Croix River basin, fish samples were available over four consecutive years at 12 unique stream sites. We compared confidence intervals derived from bootstrapping for these sites with those determined by calculating variance over time, with the goal of understanding the relative contributions of random sampling error and temporal variation to overall variability in IBI scores. Confidence intervals for IBI scores calculated from repeat visits over time were determined by $\bar{x} \pm t^{*}(s/\sqrt{n})$, where $\bar{x}$ = mean IBI score, t^{*} = critical value from the $n - 1$ student's t distribution, s = the standard deviation of the repeat IBI scores, and n = the number of repeat visits for each site. Repeat visits over multiple years were not available for sites in the Upper Mississippi River basin.

3. Results

3.1. IBI variability and the impairment threshold

IBI confidence interval length (i.e., IBI sensitivity) in response to random sampling error ranged from 0 to 40 points (mean = 11) across all 513 stream site visits included in this study. In 510 of these site visits (99.4%), this random sampling variability was not sufficient to change impairment status from 'unimpaired' to 'impaired', or vice versa, relative to the site's original IBI score (Fig. 2). In other words, for these sites a bootstrap replicate sample never received an IBI score classified as 'impaired' if the original score had been classified as 'unimpaired', or vice versa. For the remaining three site visits, 0.3%, 0.4%, and 1.0% of replicate samples were classified as 'impaired' when the original sample was classified as 'unimpaired'.

Although bootstrap IBI scores rarely indicated the opposite impairment status compared to the original score, 19.8% of a site visit's replicate samples, on average, were classified as 'potentially impaired' when the original sample was classified as 'unimpaired' (Fig. 2). Conversely, when the original IBI score was classified as 'impaired', 2.2% of replicate samples, on average, indicated that the site was 'potentially impaired'. Finally, when the original score was 'potentially impaired', 1.7% and 7.4% of replicate samples, on average, gave IBI scores that would be considered 'unimpaired' and 'impaired', respectively. Overall, 11.3% of bootstrap replicate samples yielded IBI scores that resulted in a different impairment outcome than the original IBI score. For sites with original scores that fell within 20 points of the impairment threshold, 16.0% of bootstrap replicate samples yielded a different impairment outcome than the original sample.

3.2. Covariates of IBI sensitivity

Confidence interval length was not significantly related to species richness or Pielou's evenness for sites in either river basin. However, confidence interval length did increase significantly with increasing numbers of singletons for sites in the St. Croix (linear regression, r^2 = 0.12, $p < 0.001$) and the Upper Mississippi (r^2 = 0.06, $p < 0.001$, Fig. 3). In turn, the number of singletons was strongly related to stream class, with increasing numbers of singletons occurring in progressively larger stream size classes in both major basins (ANOVA, F = 34.53, d.f. = 8504, $p < 0.001$, Fig. 4).

In addition, linear regression indicated a significant negative relationship between confidence interval length and the number of fish in the original sample for sites in St. Croix basin (r^2 = 0.09, $p < 0.001$), and the Upper Mississippi basin (r^2 = 0.04, p = 0.004). Subsequent analysis using breakpoint regression (Muggeo, 2003) indicated a threshold in confidence interval length when samples contained approximately 160–170 fish (Fig. 5). Confidence interval lengths were more likely to be less than 10 points for sample sizes beyond the breakpoint value. Sixty-two of 303 (20%) samples from the St. Croix basin and 73 of 210 (35%) samples from the Upper Mississippi basin had total numbers of fish that were less than the breakpoint values. For these sites, mean confidence interval lengths were 13 (St. Croix) and 14 (Upper Mississippi) points.

Confidence interval length was significantly related to IBI score in the Upper Mississippi basin (linear regression, r^2 = 0.25, p = 0.021, Fig. 6), but not in the St. Croix (r^2 = 0.005, p = 0.225). Original IBI score was also correlated with the total number of fish in the original sample in both basins (linear regression, r^2 = 0.27, $p < 0.001$ and r^2 = 0.24, $p < 0.001$, for the St. Croix and Upper Mississippi, respectively, Fig. 7).

3.3. Comparative IBI performance

IBI confidence interval lengths were significantly different among the nine separate stream classes (ANOVA, F = 2.787, d.f. = 7505, p = 0.007, Fig. 8). A multiple pairwise comparison test (DTK test, α = 0.05) indicated that IBI scores in stream class 5 (large rivers located in the St. Croix basin and the NCH ecoregion) and class 8 (moderate-sized streams located in the Upper Mississippi basin) had larger confidence intervals than streams in class 3 (moderate-sized streams located in the St. Croix basin); no other significant differences between stream classes were found.

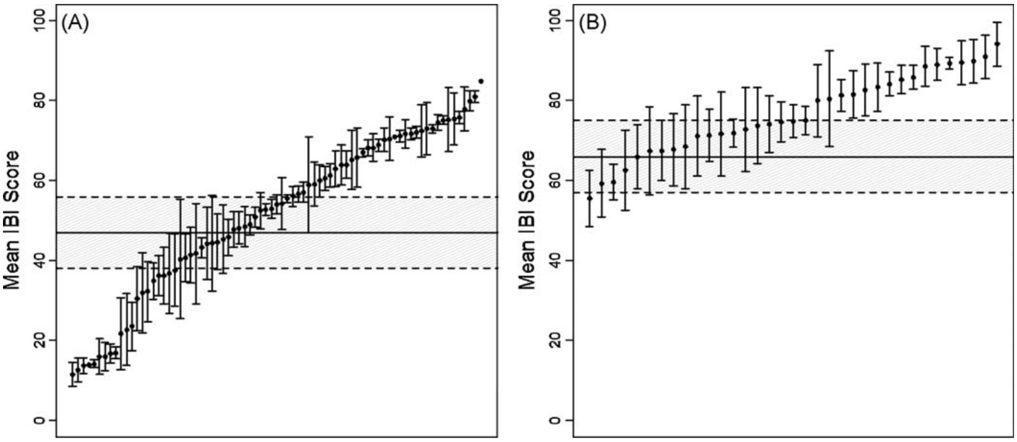

Fig. 2. Examples of confidence interval lengths generated by bootstrapping relative to the impairment threshold used to make water quality management decisions for (A) streams in class 1 (St. Croix basin; drainage area <51 km^2); (B) streams in class 5 (St. Croix basin; drainage area >691 km^2; North Central Hardwoods ecoregion). Similar relationships were obtained for sites in other drainage area and basin classes (not shown). Circles are mean IBI scores sorted into ascending order; error bars represent 95% percentile confidence intervals around the mean IBI score. Solid lines represent thresholds used by the Minnesota Pollution Control Agency to determine whether sites should be considered impaired.

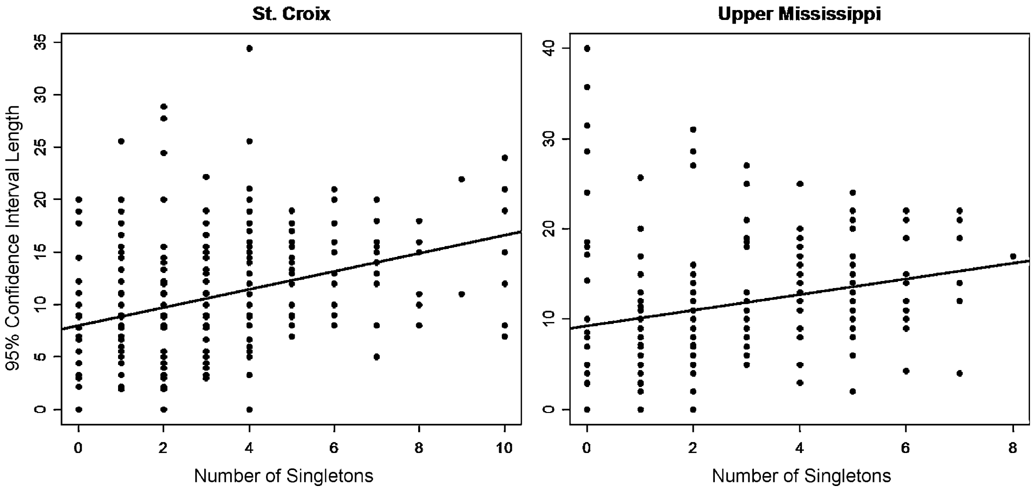

Fig. 3. Confidence interval length as a function of the total number of singletons in the original sample for sites in the St. Croix and Upper Mississippi River basins. Solid lines are statistically significant linear regression models.

3.4. Metric scoring and bias

Across all sites, mean IBI scores of bootstrap replicates were significantly lower (i.e., exhibited negative bias) than the original IBI score (paired t-test, $t = 21.09$, d.f. = 512, $p < 0.001$). Mean IBI scores were 2.5 points lower, on average, than the original scores (range = −13.4 to +8.1). Bias was significantly and negatively related to IBI score (linear regression, $r^2 = 0.08$, $p < 0.001$, Fig. 9A). Although replicate samples tended to have lower IBI scores than original samples, replicate samples for sites with low original scores were more likely to exhibit zero bias (i.e., replicate samples gave approximately the same score relative to the original sample), whereas sites with high original scores had almost uniformly negative bias (i.e., replicate samples underestimated the score).

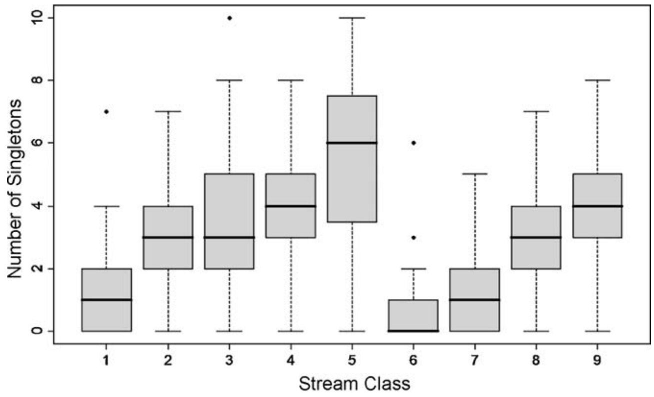

Fig. 4. Distribution of the number of singletons within each stream class. Boxes represent first and third quartiles, black lines are medians, whiskers are 1.5× the Interquartile Range, circles are outliers.

When IBIs were recalculated using the continuous scoring method, a significant negative correlation between IBI score and bias was still evident (Fig. 9B; linear regression, $r^2 = 0.15$, $p < 0.001$). However, mean bias was substantially reduced to −0.06 (range = −8.3 to +12.0), and mean IBI scores of bootstrap replicates were no longer significantly different from the original IBI scores (paired t-test, $p = 0.60$).

3.5. Comparison with variability over time

Estimates of confidence levels derived from repeat visits over time were compared with those derived from bootstrap replicates for 12 stream sites (Fig. 10). For poor to moderate quality sites

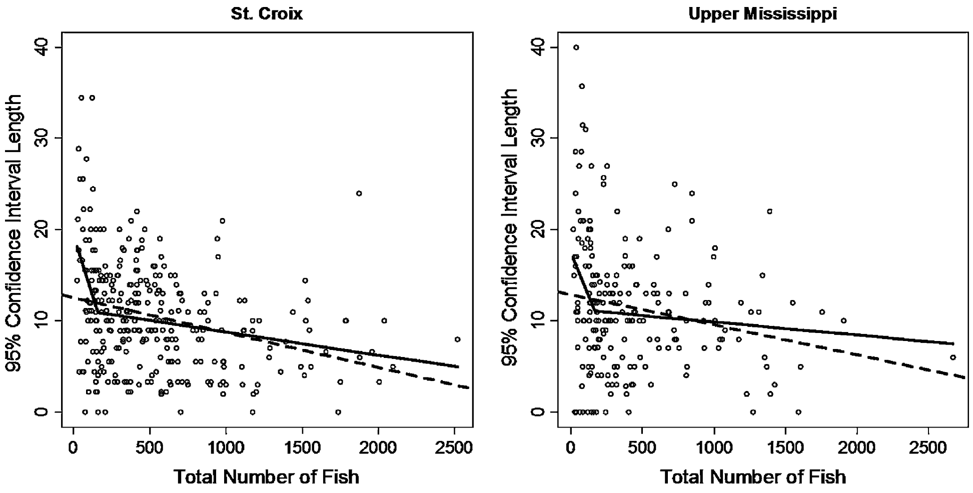

Fig. 5. Confidence interval length as a function of the total number of fish in the original sample for sites in the St. Croix and Upper Mississippi River basins. Dashed lines are statistically significant linear regression models; solid lines are statistically significant breakpoint regression models. Breakpoint regression indicated estimated break points of 160 fish (95% CI, 108–212 fish) and 171 fish (95% CI, 77–265 fish) for samples collected in the St. Croix and Upper Mississippi basins, respectively.

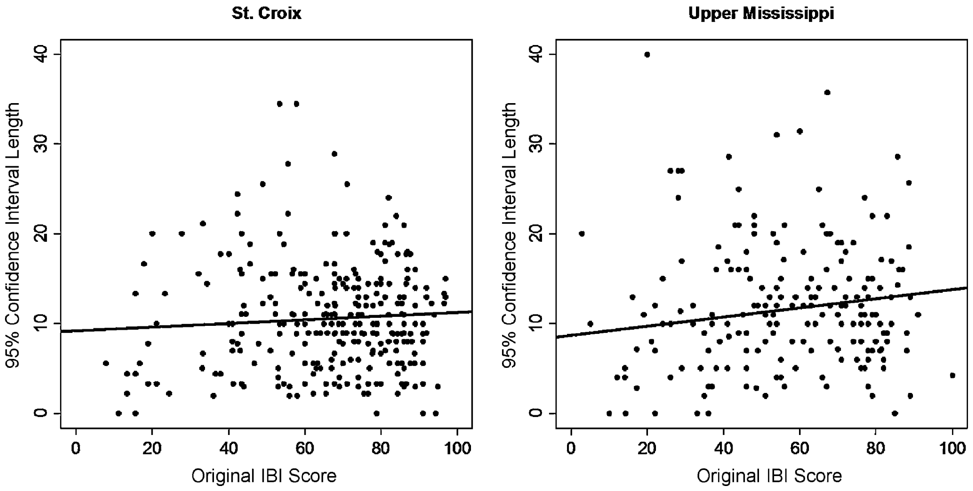

Fig. 6. Confidence interval length as a function of original IBI score. The relationship is significant for sites in the Upper Mississippi, but not for sites in the St. Croix.

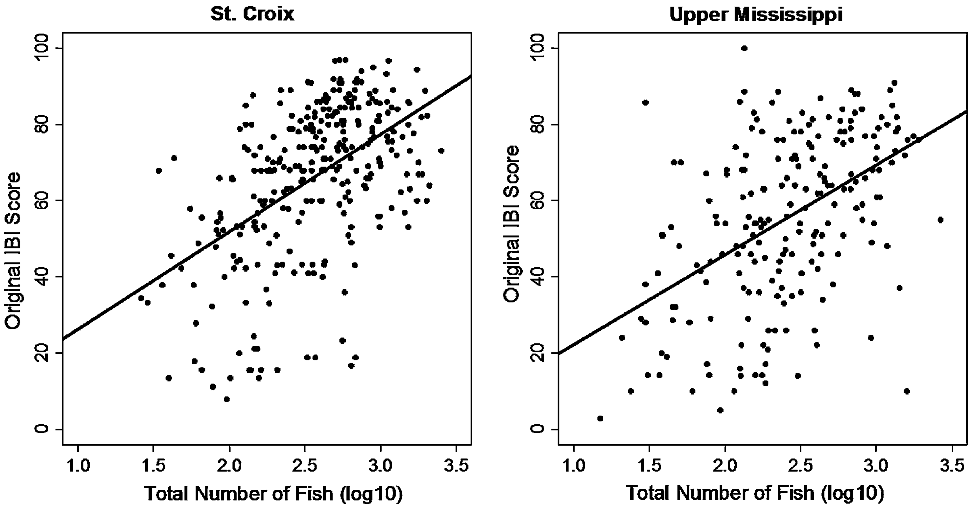

Fig. 7. Original IBI score as a function of total number of fish ($\log_{10}$ scale) in the original field sample. In both basins, solid lines indicate statistically significant relationships.

(IBI < 70), estimates of temporal variability tended to exceed variability attributed to random sampling error. For higher quality sites (IBI > 70), both estimates of variability were in close agreement, except for the two highest quality sites, in which temporal variance greatly exceeded that due to random sampling error. Both of these sites were rivers with large drainage areas (D.A. > 691 km^2).

4. Discussion

4.1. Can impairment status be affected by random sampling error?

In this study, we found that nearly a quarter of fish IBI scores for Minnesota streams varied by more than 15 points as the result of random sampling error, with the most variable score having a

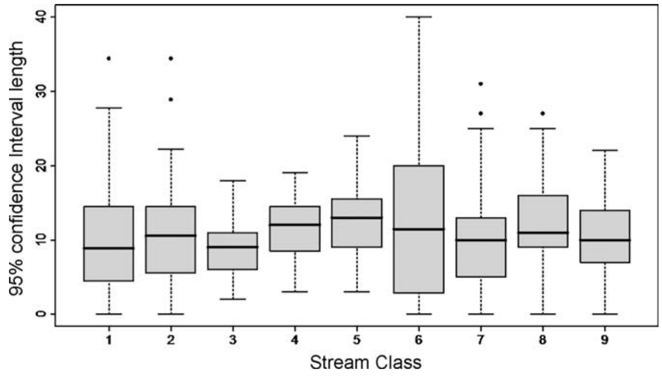

Fig. 8. Distribution of IBI confidence interval lengths (95% percentile) within each of the nine stream classes. Boxes represent first and third quartiles, black lines are medians, whiskers are 1.5× the Interquartile Range, circles are outliers. A Dunnett–Tukey–Kramer multiple comparison test indicated that confidence intervals for class 5 and class 8 IBIs were significantly greater than those for class 3 IBIs.

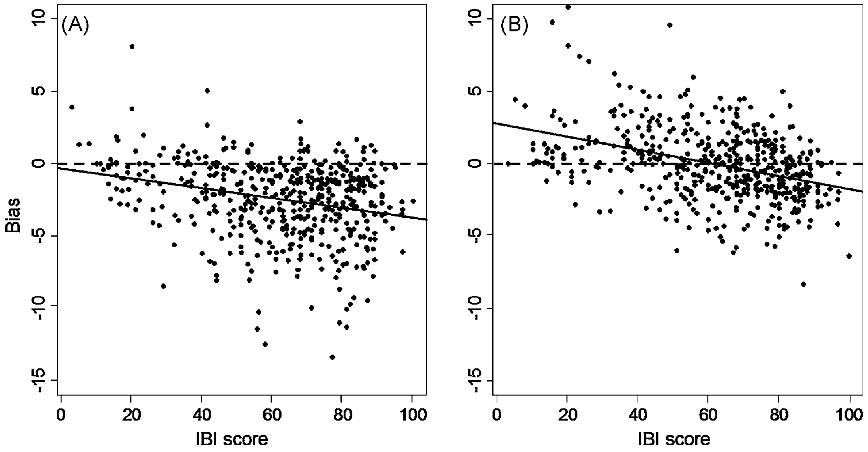

Fig. 9. Bias (mean bootstrap IBI–original IBI) for all stream sites when IBIs are calculated using (A) the original discontinuous scoring system used by the Minnesota Pollution Control Agency; and (B) a continuous scoring system using a linear curve drawn between the midpoints of metric values. Dashed lines represent zero bias expected if bootstrap IBI scores matched original IBI scores; solid lines are statistically significant linear regressions of bias as a function of IBI score.

range of 40 points. In contrast, when converted to a 100 point scale, confidence interval lengths for a set of IBIs from Ohio ranged from 0 to 25, with few sites exhibiting confidence intervals greater than 15 (Fore et al., 1994). Despite the high variability of some of the Minnesota IBIs, however, only one in 10 replicate IBI scores for a given site, on average, indicated a different impairment outcome than the original score. Moreover, random sampling variability was not sufficient to change a site's status from 'unimpaired' to 'impaired' or vice versa in over 99% of stream site visits included in this study (although an IBI score could change from 'unimpaired' or 'impaired' to 'potentially impaired', or vice versa). Not surprisingly, the number of replicate samples that produced different outcomes relative to the original sample increased for sites with IBI scores close to the impairment threshold. Taken together, however, we suggest that the effects of random sampling error on IBI score are not likely to change the impairment status of a stream site in most cases.

When random sampling variability did change the impairment outcome, we suggest that type I error (underestimating stream health) was more common than type II error (underestimating stream impairment). In this study, type I error would occur if bootstrap samples from sites with original IBI scores above the confidence region (i.e., unimpaired sites) yielded scores below or within this region. Type II error, on the other hand, corresponds to the probability that IBI scores based on bootstrap replicate samples fell above the confidence region around the impairment threshold, when the original IBI score fell below or within the impairment confidence region. The increased prevalence of type I relative to type II errors suggests that impairment decisions based on IBI scores are conservative in terms of protecting stream health; that is, by using IBI scores to determine site impairment, management agencies are more likely to list unimpaired sites as impaired or potentially impaired than they are to fail to list impaired sites. If a goal among resource managers is to protect water resources before they become severely degraded, this conservative approach may be appropriate.

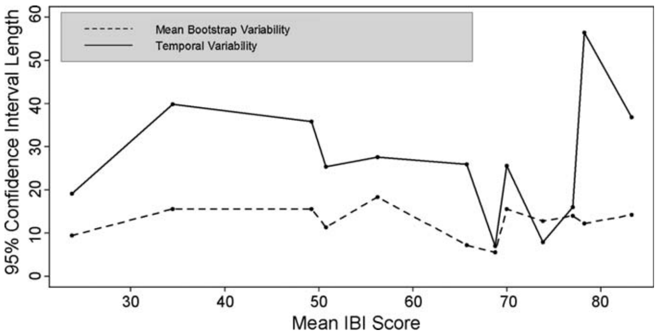

Fig. 10. Comparison of temporal variability with estimates of bootstrap variability for 12 sites in the St. Croix basin sampled over four consecutive years. Bootstrap variability was calculated as the mean of the four bootstrap confidence interval lengths generated for each site.

The greater number of type I errors is also indicative of the negative bias among IBI scores calculated from bootstrap replicate samples. Although we found that mean bias was relatively small (approximately 2–3 points), in some cases mean IBI scores derived from bootstrap replicates were 13–14 points lower than the original scores. At the same time, the negative correlation between bias and IBI score indicates that only the quality of less-impacted sites tended to be underestimated, whereas the quality of highly degraded sites tends to be more accurately conveyed by bootstrap replicate samples. Fore et al. (1994) argued that this bias occurs in part because of the discontinuous scoring system typically used in many IBIs. They further argued that scoring metrics on a continuous scale would help to reduce the effects of this bias, since small changes in samples would have a smaller effect on metric and overall IBI scores.

We designed a continuous scoring system for the metrics included in the fish IBIs, and assessed whether the revised IBI scores determined for bootstrap replicate samples exhibited reduced bias. Indeed, when metrics were scored using linear piecewise continuous curves, mean bootstrap IBI scores did not differ significantly from original IBI scores. The advantages of scoring metrics on a continuous score have long been argued (Minns et al., 1994), and our analysis appears to justify the adoption by management agencies of continuous scoring methods for new IBIs.

Many states try to avoid misdiagnosis of site impairment by collecting more than one biological sample from streams with IBI scores near the impairment threshold (Fore, personal communication, 2008). Alternatively, as in Minnesota, a weight of evidence approach may be used in which additional land use, habitat, or water chemistry data are required to list a stream site as impaired if the IBI score falls within the confidence region around the impairment threshold (MPCA, 2007). The size of this confidence region (±9 and 13 points in the St. Croix and Upper Mississippi River basins, respectively) appears sufficient to account for variability due to random sampling error, which we have estimated in this study as approximately ±5 or 6 points.

4.2. Covariates of IBI sensitivity

A number of factors were implicated in this study as possible drivers of variability among IBI scores. For all sites, bootstrap variability was significantly correlated with the total number of fish collected in the original field sample. This relationship is intuitive, given that capturing a larger number of fish provides more information about each of the component metrics of the IBI score, and thus reduces uncertainty. Based on our analysis, we found that field samples containing at least 160 fish could be interpreted with a reasonable degree of confidence. Obtaining a sample this large for all sites would likely require increasing either the standard length over which a stream reach is sampled, or increasing the sampling intensity (i.e., conducting multiple sampling passes of the same stream reach). However, although this sample size is smaller than the minimum of 400 fish recommended by Fore et al. (1994), achieving collections of 160 fish may not be realistic for highly degraded sites. Moreover, increasing the number of fish in a sample may affect IBI score, because there appears to be a significant positive relationship between these two variables.

IBI sensitivity was also related to the number of rare taxa in a sample. This relationship likely stems from the susceptibility of singletons to random sampling error; when creating replicate samples using the bootstrapping algorithm, species that occur as single individuals have only a 63% chance, approximately, of being selected in any given replicate sample. As a result, these species will not be consistently represented in all 1000 bootstrap replicate samples. In addition, 91% (88 of 97) of the species that occurred as singletons were used to score one or more richness-based metrics (i.e., metrics that are based on the number of species in a given category). Thus, the inconsistent occurrence of these species in replicate samples likely introduced variability into richness metric scores, and ultimately, into overall IBI scores.

Whether species that occurred as singletons were actually rare in the environment, or just distributed more patchily and thus difficult to sample proportionately, cannot be determined from this dataset alone. However, regardless of their true abundance in the field, our analysis indicates that the failure to accurately account for the presence of these taxa may affect the resulting IBI score. Moreover, the fact that IBI confidence interval length was not related to community evenness suggests that IBIs sensitivity will be driven more by the presence or absence of these rare taxa *per se*, rather than a skewed community distribution. This observation contradicts the contention, made by several authors, that the failure to capture all species present at a site does not significantly affect IBI score (Steedman, 1988; Fore et al., 1994; Reynolds and Herlihy, 2003), and suggests that managers may want to carefully weigh trade-offs between minimizing sampling effort and obtaining a sample that accurately portrays maximum species richness.

4.3. IBI performance and metric sensitivity

A detailed analysis of the contribution of each individual metric to overall IBI sensitivity is beyond the scope of this study. However, because the IBIs developed for each stream class contained a slightly different combination of metrics, we used differences in IBI sensitivity among stream classes to evaluate whether some combinations of metrics were more sensitive to random sampling error than others. Few significant differences in confidence interval length were found among the nine different stream classes, suggesting that the use of different combinations of metrics for IBIs in each stream class did not result in dramatic differences in IBI sensitivity to sampling error. An exception was the IBI for class 5 streams (large river sites in the NCH region of the St. Croix), which was more sensitive to random sampling error than the IBI for streams in class 3 (moderate-sized sites in the St. Croix). Even in this case, however, we speculate that the difference in IBI sensitivity may have stemmed more from the effects of stream size than from differences in sensitivity among component metrics. We base this conclusion on the fact that IBIs for these two classes actually used the same set of 10 metrics, as well as the same scoring categories for each metric. The only difference between these IBIs was that, for two of the 10 component metrics (total no. of species, and no. of sensitive species), the scoring requirements for class 5 streams were more stringent. This difference in scoring stringency is indicative of the greater species richness expected in large, undisturbed river systems in the NCH region of the St. Croix basin relative to the moderate-sized stream class, but would not be sufficient to cause increased sensitivity in the class 5 IBIs. However, samples from class 5 streams contained considerably more singletons than streams from class 3, on average. Absence of these species in some bootstrap samples would introduce greater variability in richness metrics and hence in overall IBI score. It is perhaps not surprising that singletons were more frequently collected from larger stream classes. As stream size increases, biomonitoring crews are often unable to sample the entire channel and are instead constrained to weave among available habitat types, which in turn may have prevented the collection of patchily distributed organisms in great abundance.

The higher sensitivity among class 8 IBIs relative to class 3 IBIs presents a complex problem because these two IBIs were comprised of two different sets of metrics. One possible explanation may lie in the fact that the class 8 IBI incorporated a greater number of metrics that were scored using only three scoring categories (0, 5, 10) compared to the class 3 IBI. Because the addition or loss of a species can cause the metric score to change by a five point interval, metrics scored in this way could exhibit increased sensitivity to random sampling error. However, additional analysis is necessary to identify the contribution of individual metrics to differences in overall IBI sensitivity in this case.

4.4. Random sampling error vs. variability over time

In previous analyses of IBI variability at the same sites over time, more highly degraded sites were often found to exhibit more variability than less-impacted ones (Steedman, 1988; DeShon, 1994; Niemela and Feist, 2000). Greater variability among degraded sites is typically viewed as the result of changes in the biological community related to the effects of ongoing anthropogenic disturbances, or to a compromised ability among the resident biota of these sites to maintain equilibrium following such disturbances. IBI scores may also vary to a greater extent at more degraded sites because the IBI as a quantitative tool is less capable of producing precise results at these sites (Fore et al., 1994). Although our analysis appeared to indicate that IBIs are less precise at higher quality sites, this trend is likely driven by the extremely low variability of some highly degraded sites. In both basins, several sites with low IBI scores (<30) had confidence interval lengths of 0, indicating that 95% of bootstrap replicate samples generated for these sites received the same score. Samples from these sites typically consisted of only a few species of fish, with most metrics in the resulting IBI receiving the lowest possible scores. For these sites, random differences in sample composition were not sufficient to raise metric values and thus introduce variability in bootstrap samples.

Interestingly, however, two high quality sites in this study also exhibited high levels of temporal variability. Both of these sites were located on the St. Croix River and have large drainage basins, and thus corroborate a previous study by Niemela and Feist (2000), who found that river sites in Minnesota have highly variable IBI scores over time. They suggested that this increased variability was likely due either to an inability to capture a representative fish sample in larger rivers, or to the greater array of disturbances acting on larger watersheds compared to smaller watersheds over time. While some large river sites in the St. Croix appeared more vulnerable to random sampling error than smaller sites (see above discussion), simulated sampling error alone did not account for the large variation at these two sites over time. Our analyses suggest, as argued by Jacobson (2000), that further research may be needed to develop theoretical understanding and biocriteria for larger river systems.

Temporal variance also exceeded that attributed to random sampling error for poor to moderate quality sites (IBI scores < 70). This higher variance indicated that additional factors other than random sampling may be driving changes in the biological samples collected at these sites over time. Presumably, these additional factors are related to anthropogenic stress. For the remaining high quality sites (IBI scores > 70, excluding the two large river sites), temporal variance and random sampling error were similar. In these sites, which presumably experience relatively little anthropogenic disturbance, biological communities appeared relatively stable over time, and temporal variance was relatively small. We suggest that variability in biological samples from these sites is as likely due to random sampling errors as temporal changes in biological communities.

4.5. Limitations of a bootstrap approach

There are a number of limitations associated with using the bootstrap method to quantify variability associated with IBI scores. First, the amount of variation across bootstrap replicates is constrained by the bootstrap algorithm, which samples with replacement from the original sample until the number of fish in the original sample is reached. As a result, every bootstrap replicate will have the same number of specimens as the original sample from which it is derived. We found that fish abundance was significantly correlated with IBI score; thus, restricting the potential for total abundance to vary may result in a tendency to underestimate the total amount of random sampling variability associated with IBI scores. A similar problem occurs with the total number of species in a sample; the bootstrap algorithm does not substitute new species into replicate samples beyond those found in the original sample.

Moreover, using a bootstrapping approach to calculate IBI scores assumes that the original sample used to create bootstrap replicates represents a random sample of all individuals and species present at a site, and that individuals and species are represented in the sample in proportion to their true abundance at the stream site. In reality, the original sampling event was unlikely to have met these criteria in full. Rare taxa, for example, are more likely to be unrepresented in fish samples, especially in larger river sites where by necessity electrofishing crews are unable to pass through all available fish habitat. In addition, because some fish taxa may be easier to collect than others, their relative abundance in a fish sample may not correspond to their true relative abundance in the stream's fish community. One way that such bias may occur is if electrofishing fails to capture smaller, more cryptic species or age classes (Peterson et al., 2004).

5. Conclusions

In the last several decades, the volume of literature dedicated to the development of new IBIs and related indices has grown at a vigorous rate. However, despite the sustained attention to index development, as well as the increased application of IBIs to water quality management decisions, questions remain regarding the strengths and limitations of these indices. In particular, the precision and sensitivity of IBIs have never been thoroughly addressed. Because IBIs do not use a uniform set of metrics or metric scoring criteria across various geographic regions, we cannot assume that all indices will perform with uniform sensitivity to random sampling error. Indeed, our analyses suggest that Minnesota fish IBIs are more sensitive to random sampling error than a set of previously studied IBIs from Ohio.

Moreover, if our understanding of IBI variability is to keep pace with trends in water policy, we need to directly examine how such variability relates to stream impairment thresholds. Here, we demonstrated that, while IBI sensitivity to random sampling error was somewhat larger than expected, this variability was not sufficient to change the impairment status in the majority of replicate samples for stream sites in Minnesota. For sites with IBI scores near the impairment threshold, random sampling variability is more likely to affect status determination, and more than one field sample may be needed to verify status.

We also highlighted new relationships between IBI variability and aspects of fish samples including abundance and the number of rare species, and we proposed suggestions for the number of fish upon which IBI scores should be based if the effects of random sampling errors are to be minimized. Lastly, we have demonstrated that a continuous scoring approach for the component metrics of the IBI can reduce bias in overall IBI score, relative to the discontinuous scoring methods that remain the standard approach in IBI development today.

Acknowledgements

The authors thank David Wright and the Minnesota Department of Natural Resources for funding this research, and Scott Niemela, Joel Chirhart, and Mike Feist and the Minnesota Pollution Control Agency for providing biomonitoring data, IBI scoring criteria, and assistance in developing and carrying out project objectives. We also thank Leska Fore, Scott Niemela, and two anonymous reviewers for highly constructive reviews of this manuscript. The use of trade, product, industry or firm names or products or software or models, whether commercially available or not, is for informative purposes only and does not constitute an endorsement by the U.S. Government or the U.S. Geological Survey.

References

Allan, J.D., Castillo, M.M., 2008. Stream Ecology: Structure and Function of Running Waters. Springer, Dordrecht, The Netherlands, pp. 18–20.

Blocksom, K.A., 2003. A performance comparison of metric scoring methods for a multimetric index for Mid-Atlantic highland streams. Environmental Management 31, 670–682.

Borja, A., Bricker, S.B., Dauer, D.M., Demetriades, N.T., Ferreira, J.G., Forbes, A.T., Hutchings, P., Xiaoping, J., Kenchington, R., Marques, J.C., Zhu, C., 2008. Overview of integrative tools and methods in assessing ecological integrity in estuarine and coastal systems worldwide. Marine Pollution Bulletin 56, 1519–1537.

Carlisle, D.M., Clements, W.H., 1999. Sensitivity and variability of metrics used in biological assessments of running waters. Environmental Toxicology and Chemistry 18, 285–291.

Carpenter, J., Bithell, J., 2000. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statistics in Medicine 19, 1141–1164.

Chernick, M.R., 1999. Bootstrap Methods: A Practioner's Guide. John Wiley and Sons, Inc., New York, NY.

de Boor, C., 1978. A Practical Guide to Splines. Springer-Verlag, New York, NY.

DeShon, J.E., 1994. Development and application of the invertebrate community index (ICI). In: Davis, W.S., Simon, T.P. (Eds.), Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making. Lewis Publishers, Boca Raton, FL, pp. 217–243.

Dunnett, C.W., 1980. Multiple comparisons in the unequal variance case. Journal of the American Statistical Association 75, 796–800.

Efron, B., 2003. Second thoughts on the bootstrap. Statistical Science 18, 135–140.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Annals of Statistics 7, 1–26.

EPA, 2002. Summary of Biological Assessment Programs and Biocriteria Development for States, Tribes, Territories, and Interstate Commissions: Streams and Wadeable Rivers. EPA 822-R-02-048. U.S. Environmental Protection Agency, Office of Environmental Information and Office of Water, Washington, DC.

Fore, L.S., Karr, J.R., Conquest, L.L., 1994. Statistical properties of an index of biological integrity used to evaluate water resources. Canadian Journal of Fisheries and Aquatic Sciences 51, 1077–1087.

Furse, M.T., Hering, D., Brabec, K., Buffagni, A., Sandin, L., Verdonschot, P.F.M., 2006. The ecological status of European rivers: evaluation and intercalibration of assessment methods. Hydrobiologia 566, 1–2.

Jacobson, P.T., 2000. Evaluation of multi-metric bioassessment as an approach for assessing impacts of entrainment and impingement under Section 316(b) of the Clean Water Act. Environmental Science and Policy 3, S107–S115.

Karr, J.R., 1981. Assessment of biotic integrity using fish communities. Fisheries 6, 21–27.

Karr, J.R., Chu, E.W., 1999. Restoring Life in Running Waters: Better Biological Monitoring. Island Press, Washington, DC.

Karr, J.R., Yoder, C.O., 2004. Biological assessment and criteria improve total maximum daily load decision making. Journal of Environmental Engineering 130, 594–604.

Lau, M.K., 2009. DTK: Dunnett–Tukey–Kramer pairwise multiple comparison test adjusted for unequal variances and unequal sample sizes. In: R Package Version 2.1, .

Manly, B.F.J., 2007. Randomization, Bootstrap and Monte Carlo Methods in Biology, 3rd edition. Chapman & Hall/CRC, Boca Raton, FL.

Marchant, R., Norris, R.H., Milligan, A., 2006. Evaluation and application of methods for biological assessment of streams: summary of papers. Hydrobiologia 572, 1–7.

Minns, C.K., Cairns, V.W., Randall, R.G., Moore, J.E., 1994. An index of biological integrity (IBI) for fish assemblages in the littoral zone of Great Lakes' Areas of Concern. Canadian Journal of Fisheries and Aquatic Science 51, 1804–1822.

MPCA, 2007. Guidance Manual for Assessing the Quality of Minnesota Surface Waters for the Determination of Impairment: 305(b) Report and 303(d) List. Minnesota Pollution Control Agency, Environmental Outcomes Division, St. Paul, MN, St. Paul, Minnesota.

MPCA, 2008. Final MPCA 2008 TMDL List. Minnesota Pollution Control Agency, Environmental Outcomes Division, St. Paul, MN, St. Paul, MN. , http://www.pca.state.mn.us/water/tmdl/tmdl-303dlist.html.

Muggeo, V.M.R., 2003. Estimating regression models with unknown break-points. Statistics in Medicine 22, 3055–3071.

Muggeo, V.M.R., 2008. Segmented: Segmented relationships in regression models. In: R Package Version 0.2-4, .

Niemela, S., Feist, M., 2000. Index of Biotic Integrity (IBI) Guidance for Coolwater Rivers and Streams of the St. Croix River Basin in Minnesota. Minnesota Pollution Control Agency, Biological Monitoring Program, St. Paul, MN.

Niemela, S., Feist, M., 2002. Index of Biotic Integrity (IBI) Guidance for Coolwater Rivers and Streams of the Upper Mississippi River Basin in Minnesota. Minnesota Pollution Control Agency, Biological Monitoring Program, St. Paul, MN.

Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Wagner, H., 2009. Vegan: community ecology package. Ordination methods, diversity analysis and other functions for community and vegetation ecologists. In: R Package Version 1.15-3, .

Peterson, J.T., Thurow, R.F., Guzevich, J.W., 2004. An evaluation of multipass electrofishing for estimating abundance of stream-dwelling salmonids. Transactions of the American Fisheries Society 133, 462–475.

Pinto, R., Patricio, J., Baeta, A., Fath, B.D., Neto, J.M., Marques, J.C., 2009. Review and evaluation of estuarine biotic indices to assess benthic condition. Ecological Indicators 9, 1–25.

Reynolds, L., Herlihy, A.T., 2003. Electrofishing effort requirements for assessing species richness and biotic integrity in Western Oregon streams. North American Journal of Fisheries Management 23, 450–461.

Simon, T.P., Lyons, J., 1995. Application of the Index of Biotic Integrity to evaluate water resource integrity in freshwater ecosystems. In: Davis, W.S., Simon, T.P. (Eds.), Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making. Lewis Publishers, Boca Raton, FL, pp. 245–262.

Steedman, R.J., 1988. Modification and assessment of an index of biotic integrity to quantify stream quality in southern Ontario. Canadian Journal of Fisheries and Aquatic Sciences 45, 492–501.

Stoddard, J.L., Herlihy, A.T., Peck, D.V., Hughes, R.M., Whittier, T.R., Tarquinio, E., 2008. A process for creating multimetric indices for large-scale aquatic surveys. Journal of the North American Benthological Society 27, 878–891.

Wang, L.Z., Brenden, T., Seelbach, P., Cooper, A., Allan, D., Clark, R., Wiley, M., 2008. Landscape based identification of human disturbance gradients and reference conditions for Michigan streams. Environmental Monitoring and Assessment 141, 1–17.

Yagow, G., Wilson, B., Srivastava, P., Obropta, C.C., 2006. Use of biological indicators in TMDL assessment and implementation. Transactions of the American Society of Agricultural and Biological Engineers 49, 1023–1032.