

# Variation of Subsample Estimates of Selected Benthic Macroinvertebrate Mathematical Indices: Retrospective Analysis and Proposed Criteria

Russell Anthony Isaac<sup>1,\*</sup>, James Heltshe<sup>2</sup>

<sup>1</sup>Formerly with the Massachusetts Department of Environmental Protection, Boston, USA

<sup>2</sup>Formerly with the Computer Science and Statistics Department, University of Rhode Island, Kingston, USA

## Email address:

Risaac1@yahoo.com (Russell Anthony Isaac), jfh@cs.uri.edu (James Heltshe)

\*Corresponding author

## To cite this article:

Russell Anthony Isaac, James Heltshe. Variation of Subsample Estimates of Selected Benthic Macroinvertebrate Mathematical Indices: Retrospective Analysis and Proposed Criteria. *International Journal of Environmental Monitoring and Analysis*. Vol. 10, No. 5, 2022, pp. 127-139. doi: 10.11648/j.ijema.20221005.13

**Received:** August 24, 2022; **Accepted:** September 27, 2022; **Published:** October 17, 2022

---

**Abstract:** Rapid bioassessment protocols (RBP) have been used widely to assess and compare benthic macro invertebrate communities, often in the context of determining impacts from impairments to water quality. Given that a relatively small sample of 100 organisms often was used to calculate various biological metrics, the question of how frequently differences are inferred when in fact the subsamples are from the same population (i.e., Type 1 errors) is of interest. The analysis of 72 large (300-1760 organism) field samples uses the differentiation criteria recommended in the first edition of EPA's RBP 1989 guidance manual as a case example. A minimum of 100 subsamples each of 100 organisms was used to evaluate the uncertainty of metric estimates. Variability in estimates of Community Loss, Similarity (R-Ratio), Jaccard, Sorensen, Bray-Curtis Similarity indices, and Bray-Curtis Dissimilarity as well as Diversity and Evenness also are presented. Decision criteria for judging two samples are from different parent distributions are provided for each metric at  $\alpha = 0.15$  for Type 1 errors. The proposed decision criteria are based on pooling all of the estimates of a given metric using the entirety of the calculated values of that metric derived from all subsamples of the 72 field samples. The findings demonstrate the need to vet current and potential ecological numerical metrics, for variability when estimating their values from subsamples.

**Keywords:** Macroinvertebrate Indices, Ecological Indices, Community Loss Index, Type 1 Errors in Indices, Jaccard, Sorensen, Bray-Curtis Similarity Indices, Proposed Criteria

---

## 1. Introduction

*"...replicated determinations of a diversity or biotic index can be expected to vary by chance alone. Many variance formulae for diversity indices are for the sample variance and not the variance of the sampling distribution of the diversity index. This latter value is needed for statistical inference. Without replication and a knowledge of the sampling distribution of the index, statistical procedures cannot be applied to determine if observed trends or differences result because of sampling error, or if they are a reflection of true trends or differences in the community under study. The uncritical interpretation of trends in an index of community structure or condition is a major*

*shortcoming in their current application."* [1]

So, any mathematically based assessment, whether a mathematical index or model, should have its variability assessed to be useful. No model or index can be considered reliable without a sensitivity analysis. The present paper addresses the lack of sensitivity analyses for several mathematical indices often used to evaluate aquatic macroinvertebrate communities.

Benthic macroinvertebrates long have been considered an important if not definitive indicator of water quality. The main rationale for this consideration is the expectation that these organisms integrate variations in water quality and therefore reflect transient, adverse conditions that can be missed by sporadic sampling of chemical and physical

properties. In addition, these organisms constitute a fundamental element of a waterbody's ecology. Among other functions, these organisms form an important part of the food chain both as consumers and as prey as well as playing a critical role in the cycling of nutrients within lotic systems, functioning as transformers of both complex organic and inorganic materials. Over the years, many mathematical indices have been proposed to capture the status of the benthic community. Some of the early work on mathematical indices was done in the 1940s [2, 3] followed by several others (e.g., [4, 5]). Initially, much of the assessment examined the response of benthic organisms to organic carbon as measured by biochemical oxygen demand (BOD) and/or low dissolved oxygen (DO). This interest reflects the fact that many cities, towns and industries discharged domestic sewage and other oxygen consuming wastewater untreated or inadequately treated to the nearest waterway with resulting adverse impacts on water quality.

Today, as a result of correcting most of the DO problems caused by these point-source discharges of BOD, the concern has shifted to the water quality impairing forces of nutrient enrichment, toxic substances and eroded (clean) sediment. While not fully characterized, the impacts of these substances are thought to be reflected in the response of the benthic community to a range of stressors albeit in detail perhaps different from those caused by organic loads and low DO. While impairment can be detected based on the composition of the benthic community one cannot always identify the agent responsible for the impact. In spite of this challenge, assessments based on biological information remain very informative. In addition to efforts to characterize the response of benthic organisms to particular stressors, much effort has been devoted to sampling techniques and requirements. The United States Environmental Protection Agency (EPA) developed detailed guidance for collecting and evaluating samples of benthic organisms [6, 7]. Among the challenges that need to be addressed is the fact that identifications, especially to species, require specialists. In addition, the sheer number of organisms gathered in a standard sample can represent a significant demand on resources. EPA and others have addressed these issues by developing techniques which use classifications at levels above species and with limited numbers of organisms. While information is lost, these streamlined techniques make data gathering and evaluation more practical. Among the most widely used of these techniques was EPA's rapid bio-assessment protocol (RBP) which is based on selecting and identifying 100 organisms from a sample to either family level for one comparison or to genus or species for a second more robust assessment. In addition, the idea of using several indices (multi-metric) is now favored but continue to use some of the indices described here as part of the Integrated Biological Index (IBI), (e.g., [8]). Both EPA [9] and USGS [10] surveyed State biological monitoring programs. While the multi-metric analysis now is common, Similarity and Diversity remain components in many of the states' biological monitoring programs. One question that arises is

how reliable is this sampling protocol for characterizing the health or differences in benthic communities through various indices? More specifically, how often would two samples of 100 organisms each from the same population be considered different (i.e., a Type 1 error)? The variability of macroinvertebrate communities based on the number of organisms in subsamples was examined [11] in contrast to the actual field samples. A main conclusion was that "... in subsamples of 100-300 organisms, discriminatory power was low enough to mislead water resource decision makers." EPA's revised RBP guidance [7] provides some statistical assessment of certain metrics based on limited data. The present work was expanded by examining the effect of subsample size and species classification on Richness Ratio and Community Loss [12].

Analysis of macroinvertebrates continues to be a major tool in characterizing and evaluating a wide variety of aquatic systems in various parts of the world. This reality is illustrated by the diversity of a few selected reports [13-18]. However, the variability and hence reliability of the indices used rarely if ever is considered [1]. The objective of the present study is to explore and augment the information on the variability of selected bio-indices used to characterize benthic invertebrate communities and to propose consistent decision values for the metrics examined. Management decisions, often for expensive actions, benefit from appreciating the variability of the analysis being presented and may temper any conclusions and recommendations.

## 2. Method

Study design consisted of mathematically selecting 100 organisms, randomly from, and based on the distribution of organisms in a large field sample. One could construct theoretical populations (e.g., [19, 20]) simply as mathematical expressions to explore the variability of subsamples taken from a larger population. However, to gain a greater flavor of reality, numerically large (mostly 400 to over 1000 organisms) field samples from streams in the state of Vermont were used. The State of Vermont graciously shared 73 large field samples that were collected by or for it. The organisms were identified to the species level. In a few field samples, some organisms could not be identified at the species level. In those cases, the unidentified organisms were differentiated and assigned numbers in lieu of species classification. Of the 73 field samples available, one was eliminated because, while it had 22 species, it consisted of only 44 organisms. The remaining 72 field samples, representing the "universe" or population, were used to document the variability of each index based on at least 100 subsamples of 100 organisms each. Characteristics of the field samples evaluated are presented in Table 1.

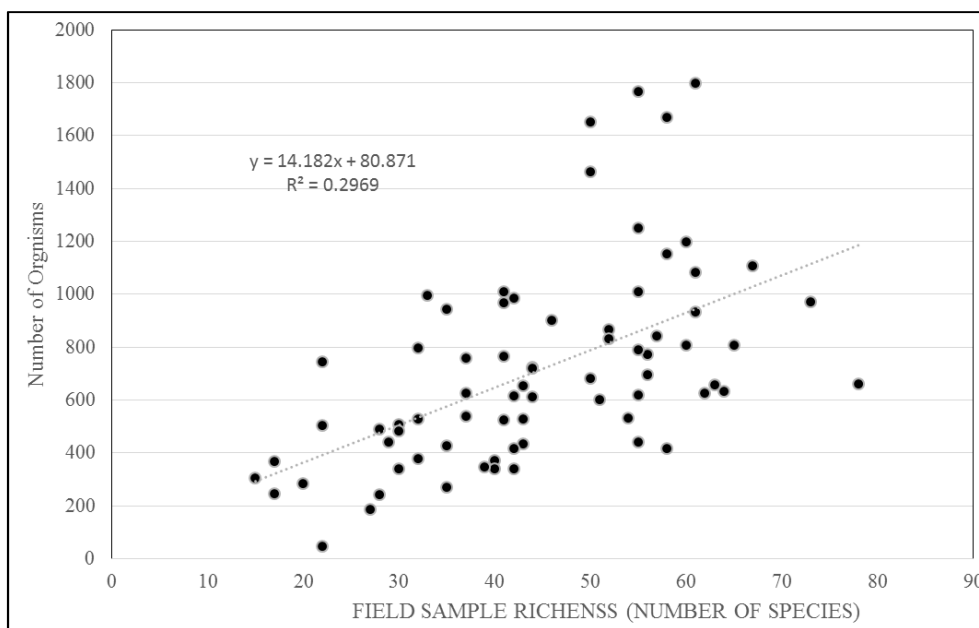
Typically, only one field sample from a site was chosen for this exercise even though some sites were sampled in more than one year. This restriction was to eliminate confounding the results of temporal variation. However,

when field samples were collected in duplicate at a location, the data were combined to provide the largest universe thereby capturing as many species as possible. Mathematical sampling of the field sample data was an automated "bootstrap" operation performed using an EXCEL® spread sheet. The standard approach was to generate 100 test subsamples of 100 organisms each and then to calculate species Richness Ratio (Similarity), Community Loss, Diversity and Evenness. In addition, values for Jaccard, Sorensen and Bray-Curtis indices were calculated.

**Table 1.** Statistics of the number of organisms in the 72 field samples.

N	72	
	ORANISMS	SPECIES
MAX	1796	78
MIN	185	15
MED	656	43
MEAN	721	45

It is expected that the number of species detected increases with the number of organisms in the sample. This qualitative understanding is given some quantitative estimate through Figure 1.



**Figure 1.** Number of organisms vs number of species in 72 field samples.

One hundred subsamples were generated and used to characterize the range of estimated values calculated for several commonly used mathematical indices. The indices commonly are used in the evaluation of aquatic benthic macroinvertebrate communities. Random numbers (EXCEL®: RANDBETWEEN) were selected and ranged from 1 to the total number of organisms in a given field sample. The random 100 numbers (i.e., organisms) were then sorted into bins (EXCEL®: FREQUENCY) equal to the number of species in the large field sample and based on the probability distribution of the organisms in the field sample. Each of the indices examined is defined in the following paragraphs and the steps to calculate its values using EXCEL® described. The size of subsamples (100 organisms) is specified in EPA's RBP guidance [6]. The use of the EPA guidance was to assess its performance and to follow its approach for other indices for consistency and for a retrospective analysis.

### 2.1. Richness Ratio (Similarity)

$$\text{Richness (R)} = \text{Number of Species} \quad (1)$$

$$\text{Richness Ratio (Similarity) of two samples} = \frac{S_1}{S_2} \quad (2)$$

$S_1$  is the subsample with the lower number of species and  $S_2$  is the subsample with the higher number of species.

The ratio is based on the number of species in two subsamples, which, in this case, have been generated from the same field sample. For this investigation, the subsample with the smaller number of species is considered the test sample and the one with the larger number is set as the reference sample. The maximum value of 1 therefore results when both subsamples have an equal number of species but not necessarily with each pair having the same number of organisms.

### 2.2. Bray-Curtis

$$\text{Bray - Curtis Similarity Index} = \frac{2\sum \text{MIN}(N_{12})}{(N_1 + N_2)} \quad (3)$$

$$\text{BC Dissimilarity} = 1 - \frac{2\sum \text{MIN}(N_{12})}{(N_1 + N_2)} \quad (4)$$

BC Dissimilarity is the Bray Curtis Dissimilarity Index.  $\sum \text{MIN}(N_{12})$  is the sum of the fewer of two counts of organisms when the same species occurs in both samples.

$N_1$  is the number of organisms in species 1.

$N_2$  is the number of organisms in species 2.

After generating two subsamples of 100 organisms each, one compares the number in each pair of the same species occurring in both subsamples and selects the minimum of the two entries as identified in EXCEL® by the statement

=IF(AND(AD104>0, AE104>0), MIN (AD104, AE104),0)

Where, for example, EXCEL® cells AD104 and AE104 are the first cells for the first set of common species in each of the two subsamples being compared.

The 100 organisms are distributed among some if not all of the species identified in the field sample according to the distribution determined through the field sample. The preceding function is used to generate a column of results equal to the number of species in the field sample. The minimum of the number of organisms is entered if the two entries for a given species both are greater than zero, or zero otherwise. The sum of this column represents the sum of the minima as specified in the Bray-Curtis Similarity equation. This sum is then divided by the sum of the total number of individuals in the two subsamples. Since, in this case, each subsample has 100 individuals, the total number of organisms is 200. The theoretical maximum value is 1 meaning the subsamples have exactly the same number of organisms for corresponding species and the number of organisms is the same in each subsample.

For the Bray-Curtis Dissimilarity metric, one simply subtracts the decimal resulting from the Bray-Curtis Similarity calculation from 1.

### 2.3. Community Loss

$$\text{Community Loss} = \frac{(d-a)}{e} \quad (5)$$

a = number of species common to both subsamples.

d = total number of species present in the subsample with the greater number of species and

e = total number of species present in the subsample with fewer species.

This index is generated from two 100-organism subsamples with the first being aligned over the second in the same spreadsheet column for convenience. The two subsamples will have the same number of organisms (100). The two grids are below the top two; they will be populated by examining the contents of each cell. If a cell has a non-zero number of organisms, a 1 will be entered in the first cell of grid 3, otherwise a 0 will be entered. The same is done for the first cell of subsample 2 and the result is entered in the first cell of grid 4. Each cell of grid 3, up to the number of species in the field sample, will be filled with a 0 or 1 by the statement:

=IF(AD104>0,1,0)

Where EXCEL® AD104 is the first cell in the first subsample. The same will be done for the rest of the cells in subsample 1. The same procedure will be done for subsample 2 with the results being entered into grid 4. Then the number

in the first cell of grid 3, which represents subsample 1, will be added to the first cell in subsample 2 and so on until the all the cells have been added resulting in values of 0, 1, or 2. This produces grid 5 with each column containing the same number of cells as there are species in the field sample. The number of 2s are counted in the column and represent the species that appear in both samples. This is done for 100 columns in the present investigation yielding 100 comparisons of paired subsamples.

### 2.4. Jaccard Index

$$J = S_c / (S_1 + S_2 - S_c) \quad (6)$$

J is the Jaccard Index.

$S_c$  is the number of species in common between the two samples.

$S_1$  is the number of species in Sample 1.

$S_2$  is the number of species in sample 2.

The Jaccard Index is calculated using the number of species common to two subsamples and the total number of species in each subsample reduced by the number in common. If Community Loss has been calculated already, then the information needed has been developed. Otherwise, one would follow the procedure for Community Loss calculations until the required data have been developed.

### 2.5. Sorensen Index

$$S = 2S_c / (S_1 + S_2) \quad (7)$$

S is the Sorensen Index and the other terms are as defined for the Jaccard Index.

The Jaccard and Sorensen Indices represent a similar approach. The same values of variables used in the Jaccard Index are used in the Sorensen Index.

### 2.6. Shannon-Wiener Diversity Index

Once the initial subsample is established and the 100 organisms, in this case, are distributed among the species, the decimal occurrence ( $p_i$ ) and  $-\ln(p_i)$  can be calculated. The product of the values of these two variables are entered for each species and the sum is the Shannon-Wiener Index.

$$H' = - \sum_{i=1}^s p_i \ln(p_i) \quad (8)$$

$H'$  is the diversity index.

$p_i$  = proportion of total number of organisms in the  $i^{\text{th}}$  species (taxon).

s = the number of species.

Note that Shannon and Wiener did not collaborate directly on developing the index. Shannon did credit Wiener's work as foundational to his own. Therefore, some prefer Shannon-Wiener Index to Shannon Index as discussed by [21].

### 2.7. Evenness

$$E(H') = [- \sum_{i=1}^s p_i \ln(p_i)] / \ln(s) \quad (9)$$

$E(H')$  is the Evenness metric.

$p_i$  = proportion of total number of organisms in the  $i^{\text{th}}$  species (taxa).

$s$  = the number of species (taxa).

The Evenness index is calculated by dividing the Shannon-Wiener Index by the natural logarithm (i.e.,  $\ln$ ) of the number of species appearing in the subsample being examined.

### 3. Results and Discussion

The results show high variability in estimates of the value of indices based on subsamples. This variability illustrates the concern voiced in [1].

#### 3.1. Similarity Indices

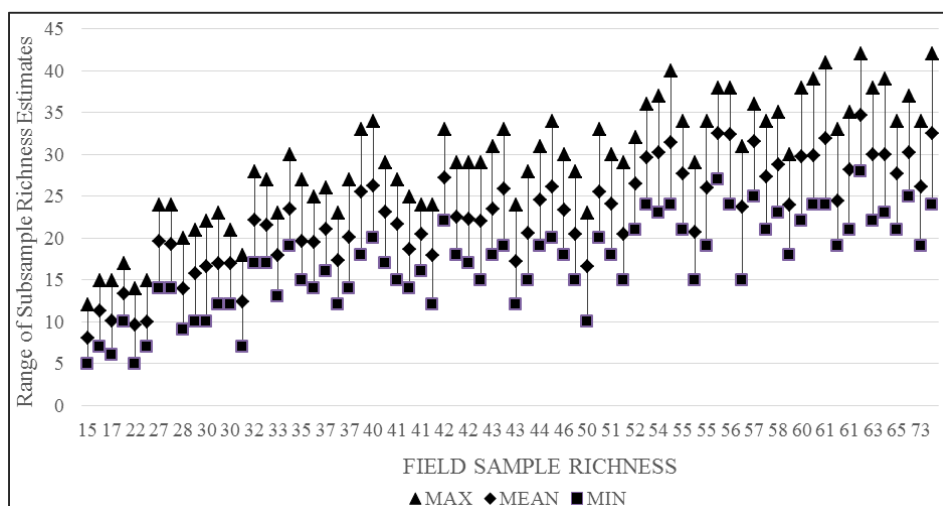
One question of interest is how the Richness of the "universal" field sample is reflected in that of the 100-organism sub samples. The range and mean of the Richness values for the 100 100-organism sub-samples are plotted in Figure 2.

The plots are ordered based on increasing richness of the

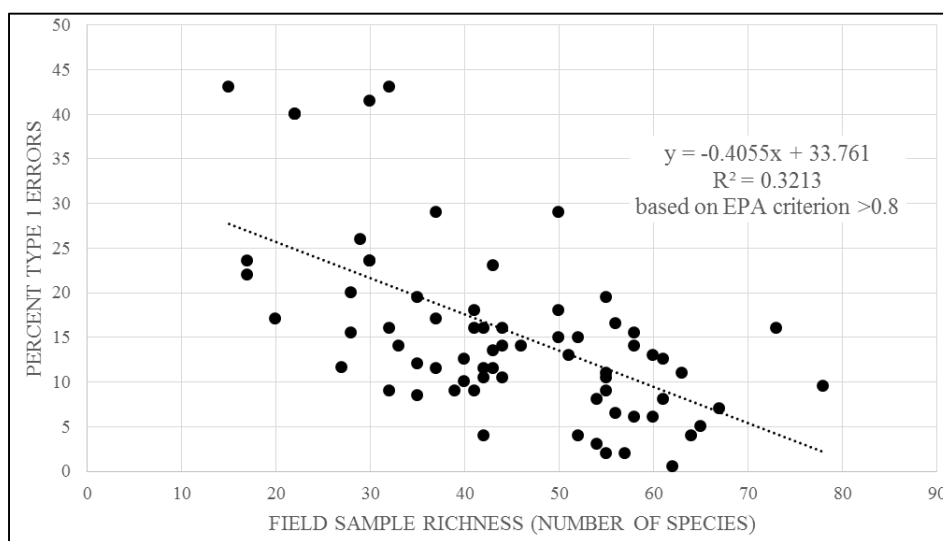
field sample (universe).

The number of species in a 100-organism sample is expected to decrease as the number of taxa increase, since rare species will be missed. There is a wide range of richness estimates for the 100 test samples as indicated in Figure 2. The subsample values increasingly underestimate the richness of the field sample as the field sample richness increases. Figure 2 indicates as many as 40% or of the species could be missed in the more species rich field samples.

The reality of the variability in estimates from subsample depicted in Figure 2 leads to the primary question of this retrospective analysis: how often would two-100 organism sub-samples from the same population (i.e., same field sample in this case) be considered different (Type 1 error) when using Similarity (Richness Ratio) between the two samples of  $\geq 0.8$  to infer both samples came from the same population. The maximum Similarity is 1 based on the definition presented previously. While the data are scattered, there is a modest inverse relationship between Type 1 error and richness of the field samples (Figure 3).

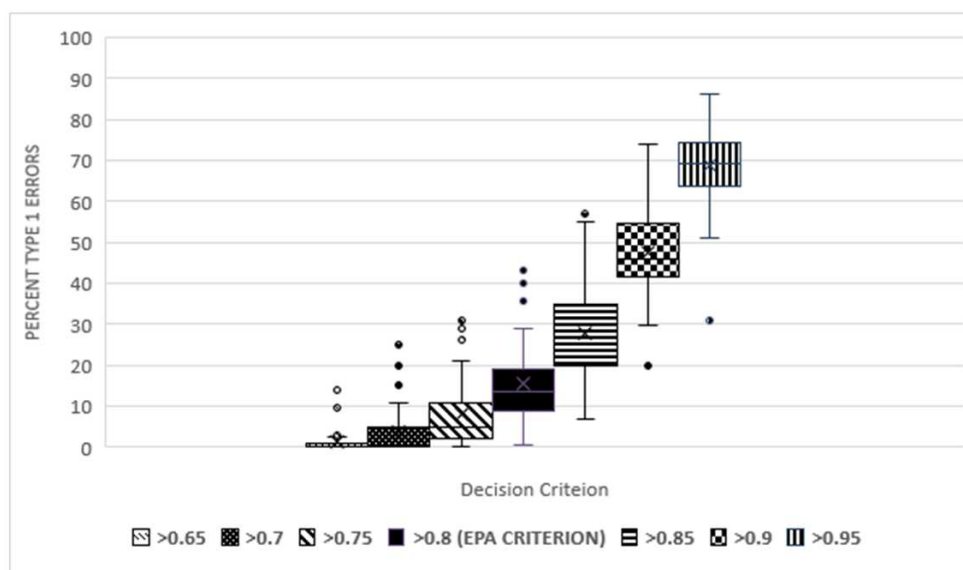


**Figure 2.** Range and mean of estimates for sample Richness based on 100 100-organism subsamples vs Richness (number of taxa) of 72 field samples.



**Figure 3.** Percent Type 1 errors versus Richness in 100 100-organism subsamples from each of 72 field samples.

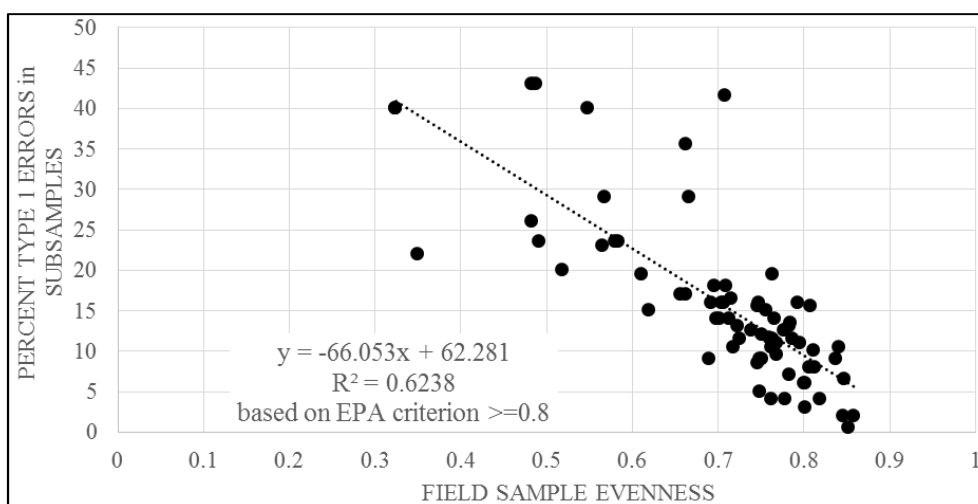
Type 1 errors can exceed 30% as noted in Figure 3. Finally, the frequency of Type 1 errors increases rapidly as the decision criterion is made more stringent by increasing it above 0.8 (Figure 4).



**Figure 4.** Type 1 error in estimates of Similarity (Richness Ratio) for 100-100 organisms subsamples from each of 72 field samples vs decision criterion (the higher, the more stringent and the greater potential for type 1 errors) using the criterion  $>0.8$ . Lower values increase the potential of Type 2 errors.

Increasing the criterion beyond  $\geq 0.8$  means that two samples have to have smaller differences for their ratio to be considered from the same population. For Similarity, the lower the criterion, the lower percentage of Type 1 errors. At the same time, lowering the criterion and reducing Type 1 errors may increase Type 2 errors (the conclusion that two

samples are not from different populations when they really are). This issue was not addressed in the present effort. Subsamples from field samples with higher values of Evenness tended to have higher estimates of Evenness and lower percentages of Type 1 errors (Figure 5).

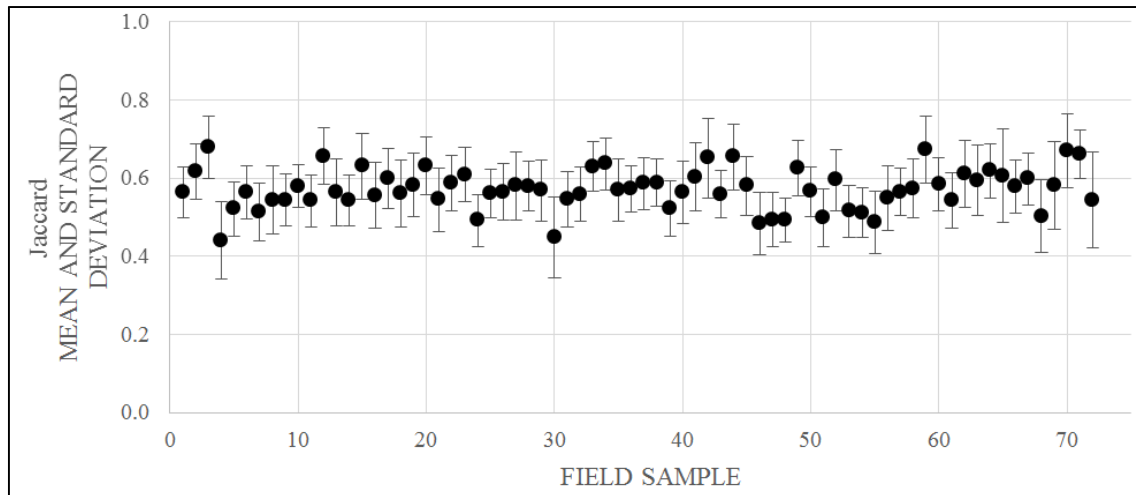


**Figure 5.** Percent of Type 1 errors in Similarity calculations based on 100 100-organisms subsamples from of 72 field samples. Values of  $<0.8$  constitute Type 1 errors.

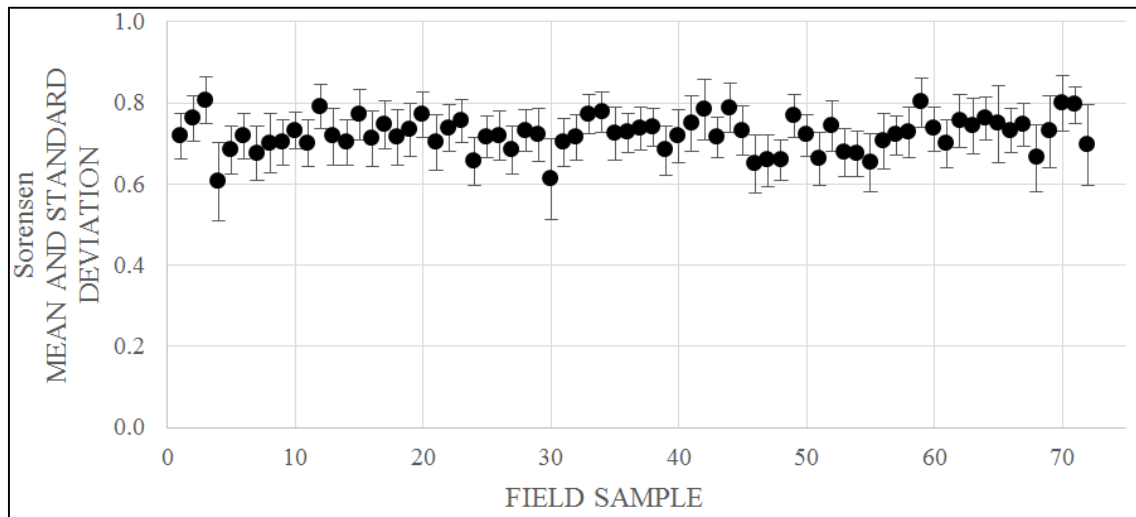
The Jaccard, Sorensen and Bray-Curtis similarity/dissimilarity indices do not have specific decision criteria. However, the variability of their estimates from subsamples is represented by their means and standard deviations which are shown in Figures 6-9.

The mean and standard deviation of the Richness Ratio is shown for comparison. All four similarity indices produce similar results. Given that the Jaccard and Sorensen Indices

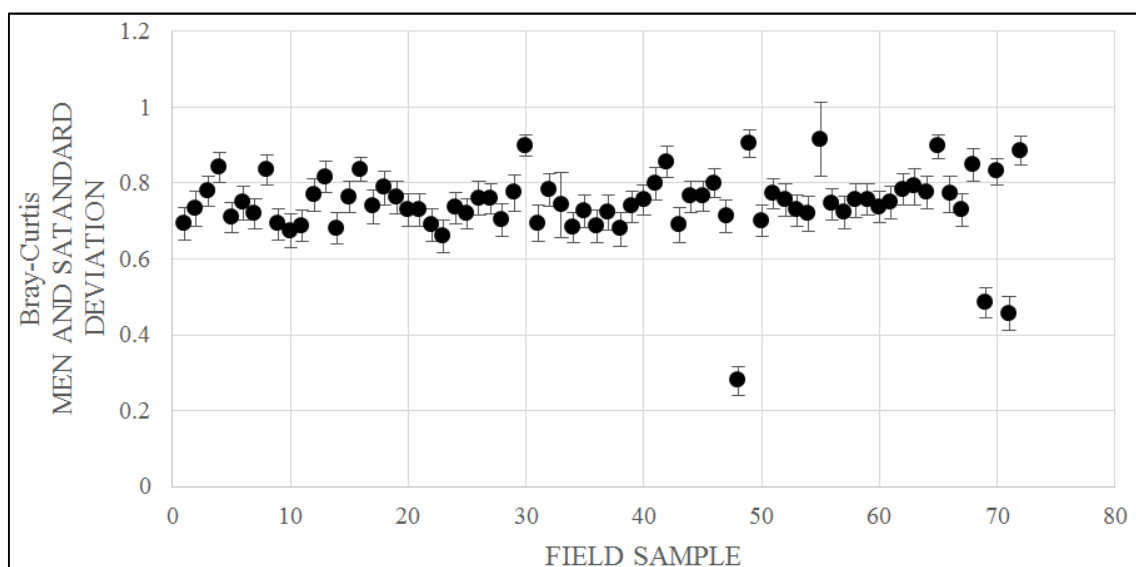
are mathematically related, the consistency between those two is expected. Bray-Curtis Dissimilarity involves calculating an estimate of similarity which is then subtracted from 1. So, the similarity portion can be compared to the other three indices explored here. The result is a similar pattern for all three similarity calculations (Figures 6-8). The Bray-Curtis Dissimilarity calculations are presented in Figure 9.



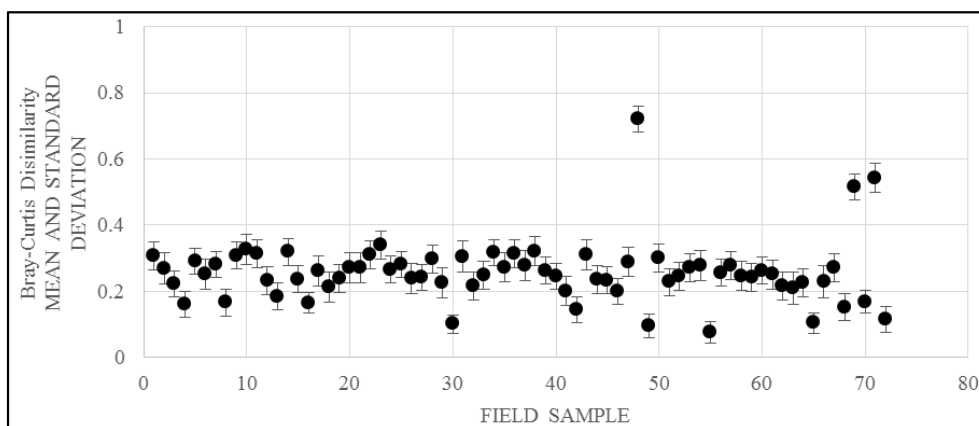
**Figure 6.** Mean and standard deviation of Jaccard Similarity Index based on 100 100-organism subsamples from 72 field samples.



**Figure 7.** Mean and standard deviation Sorensen Similarity Index based on 100 100-organism subsamples from 72 field samples.

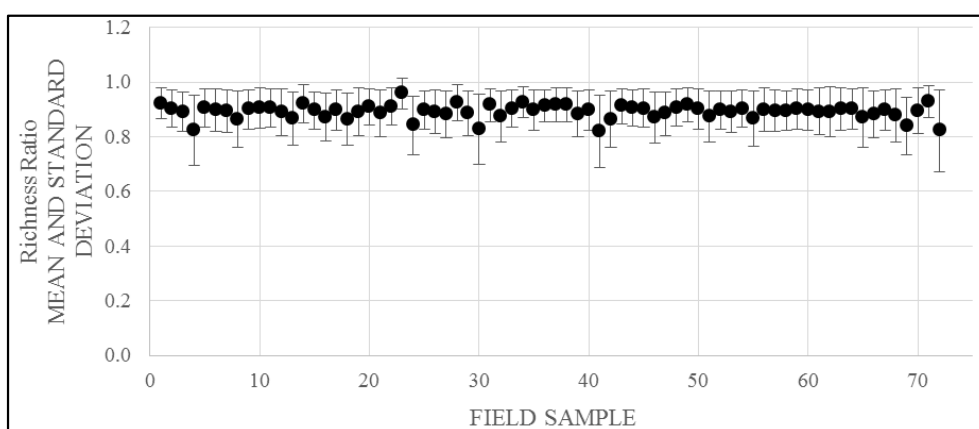


**Figure 8.** Mean and standard deviation of Bray-Curtis Similarity Index based on 100 100-organism subsamples from 72 field samples.



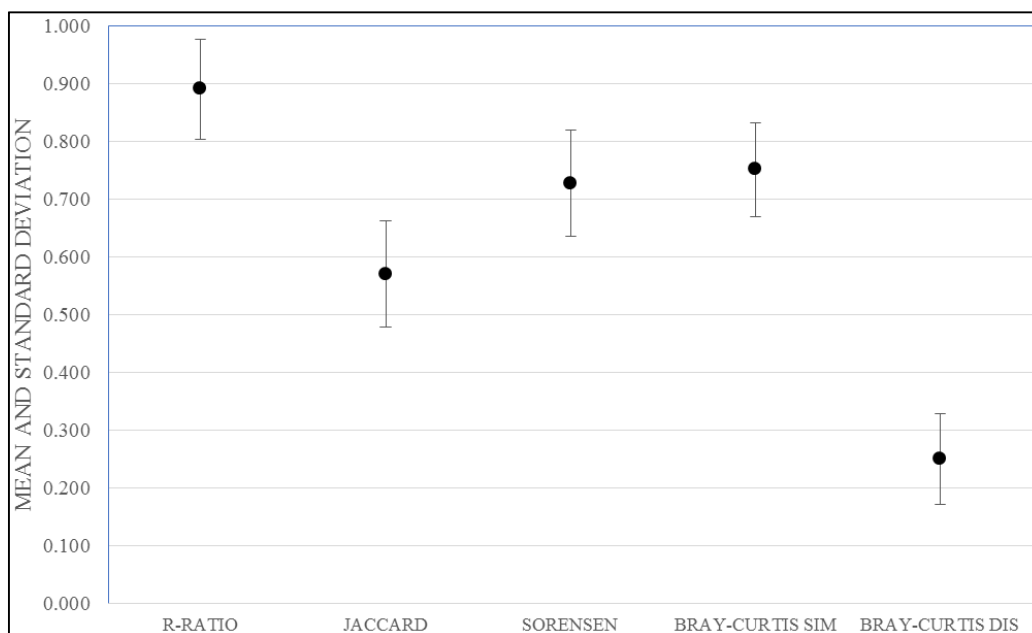
**Figure 9.** Mean and standard deviation of Bray-Curtis Dissimilarity Index based on 100 100-organism subsamples from 72 field samples.

The mean and standard deviation of the Richness Ratios (Figure 10) is shown for comparison.



**Figure 10.** Mean and standard deviation Richness Ratios based on 100 100-organism subsamples from 72 field samples.

One can obtain a more direct comparison by pooling each set of similarity calculations from subsamples. The mean and standard deviations are plotted in Figure 11.



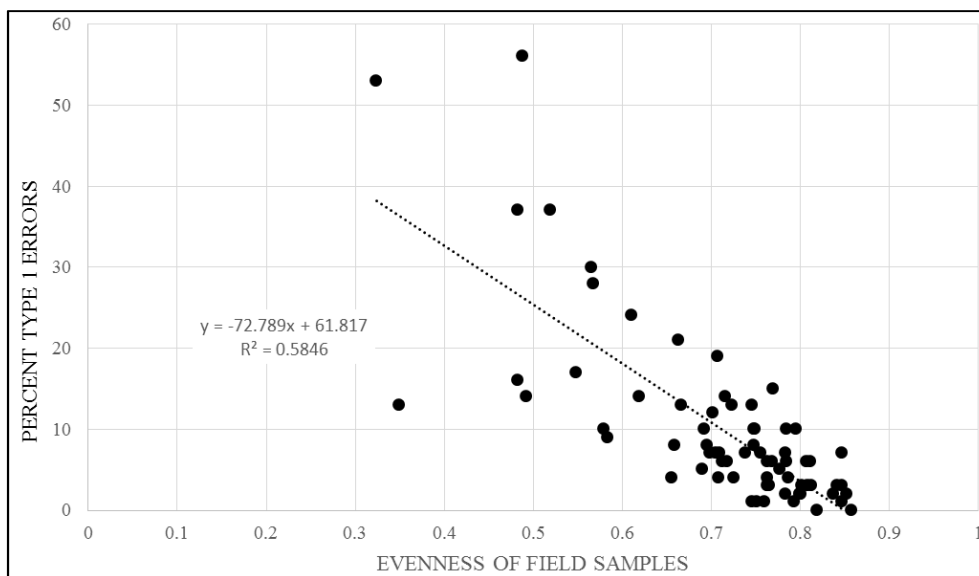
**Figure 11.** Mean and standard deviation of Similarity/Dissimilarity indices based on pooling all subsamples calculations from the entire 72 field samples  $N \geq 7500$ .



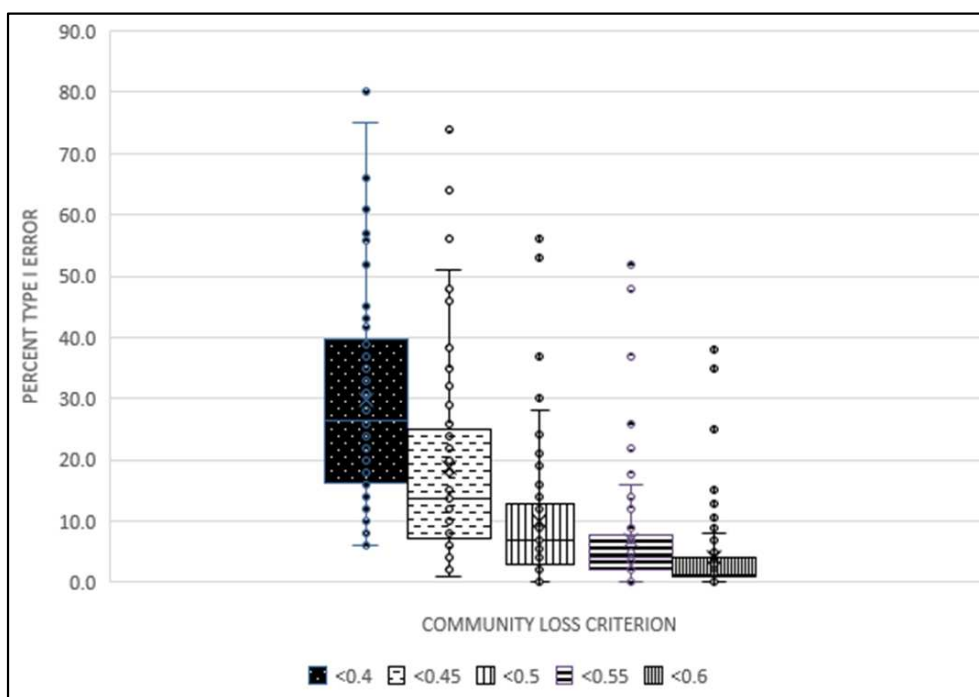
### 3.2. Community Loss

Community Loss would be zero for identical samples. The authors evaluated the test criterion of  $<0.5$  for the Community Loss criterion, as defined previously. A result of Community Loss  $< 0.5$  implies that the samples are not from different populations. Metric values  $> 0.5$  were considered to denote that the sub-samples came from different populations. Type 1 errors for Community Loss calculations for the 100 pairs of samples of 100 organisms each generally are below 25% although several values equaled or exceeded 30% (Figure 12).

The authors were unable to find any obvious differences between these five data sets and the other 67 that suggest reasons for the large percentage of Type 1 errors. The number of times one would conclude that samples from the same population were from different populations (i.e., the Type 1 error) was independent of the field sample richness ( $R^2 = 0.031$ , data not shown). For 11 of the 72 field samples, the Type 1 error for the calculated values of Community Loss exceeded 15% and Type 1 errors tended to be lower for field samples with higher Evenness (Figure 12).



**Figure 12.** Percent Type I errors for estimates of Community Loss from 100 subsamples from each of 72 field samples versus Evenness of field samples using the EPA criterion  $<0.5$ .



**Figure 13.** Community Loss: Type I error in 100 100-organism subsamples from 72 field samples versus value of decision criterion. EPA criterion  $<0.5$ . The lower the value, the more stringent the criterion and the potential for Type 1 errors increases.

The same was true for maximum and minimum estimates of Community Loss based on subsamples, but with lower  $R^2$  values (0.500, 0.423 respectively, data not shown). The frequency of Type 1 errors increases greatly as the decision criterion is reduced (i.e., made more stringent) below 0.5 (Figure 13).

In contrast to Similarity, Type 1 errors for Community Loss subsamples were independent of the number of species in the field sample (data not shown,  $R^2= 0.07$ ). However, a substantial reduction of Type 1 errors occurred with making the decision criterion less stringent than the one currently used (e.g., changing it from  $<0.5$  to  $<0.55$  reduces the Type 1 errors from 11 to 8 of the 72 field samples tested as noted in 3.4. Once again, Type 2 errors are likely to increase when the

criterion is increased for Community Loss.

3.3. Diversity and Evenness

Diversity and Evenness metrics share a computational component. There is no set criterion for interpreting calculations for either as there are for Community Loss and Similarity. Yet, the results of the subsamples from the field sample provide information that may be useful in selecting a criterion for each based on a reasonable knowledge of their variation presented here. The results of calculations based on subsamples and those calculated for the field samples are presented in Figures 14 and 15.

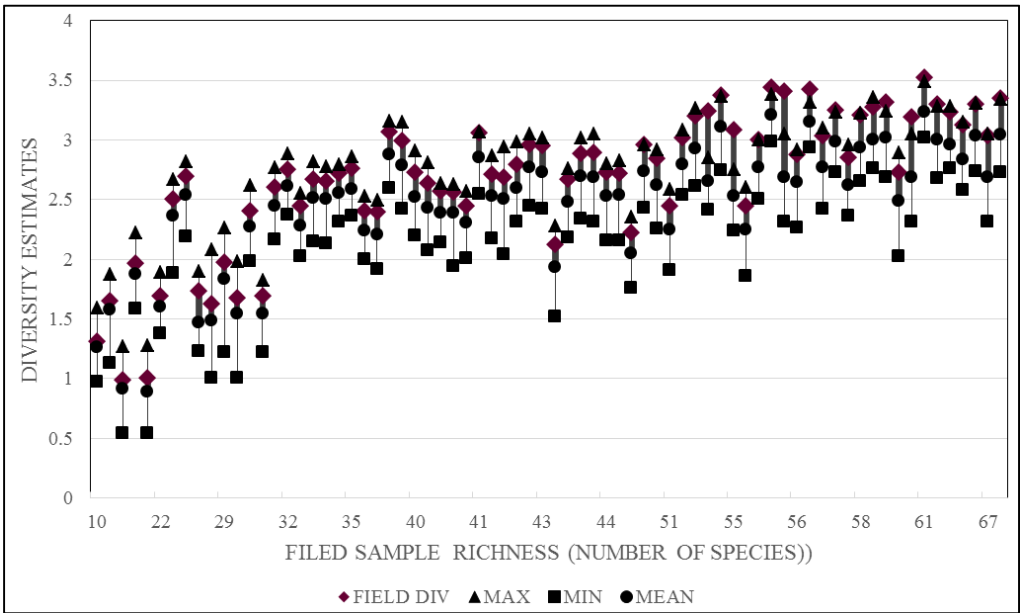


Figure 14. Range of diversity estimates calculated from 100 100-organism subsamples from 72 field samples vs Richness of the field samples.

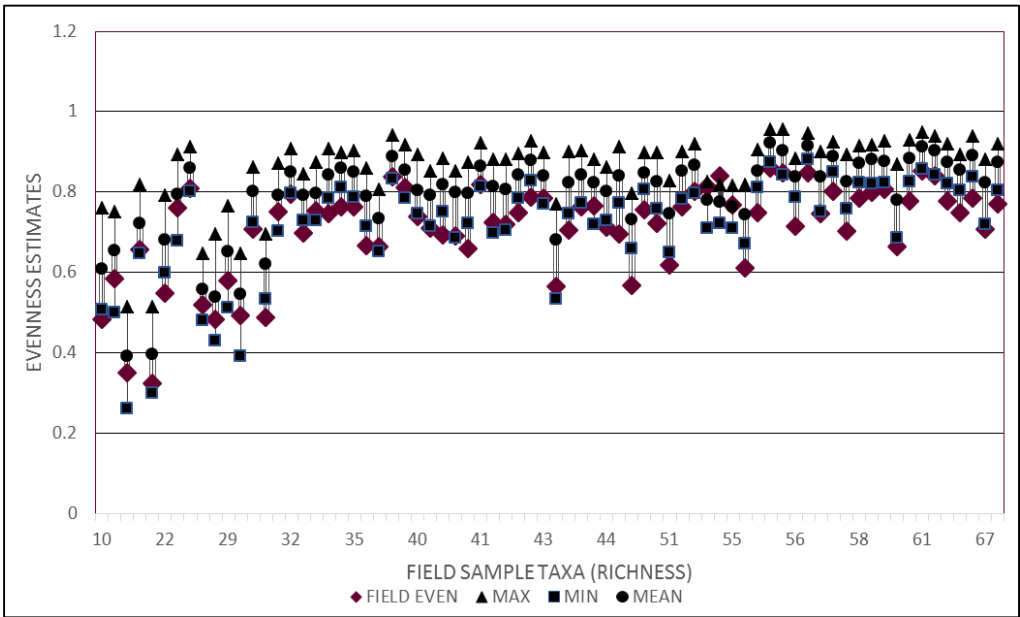
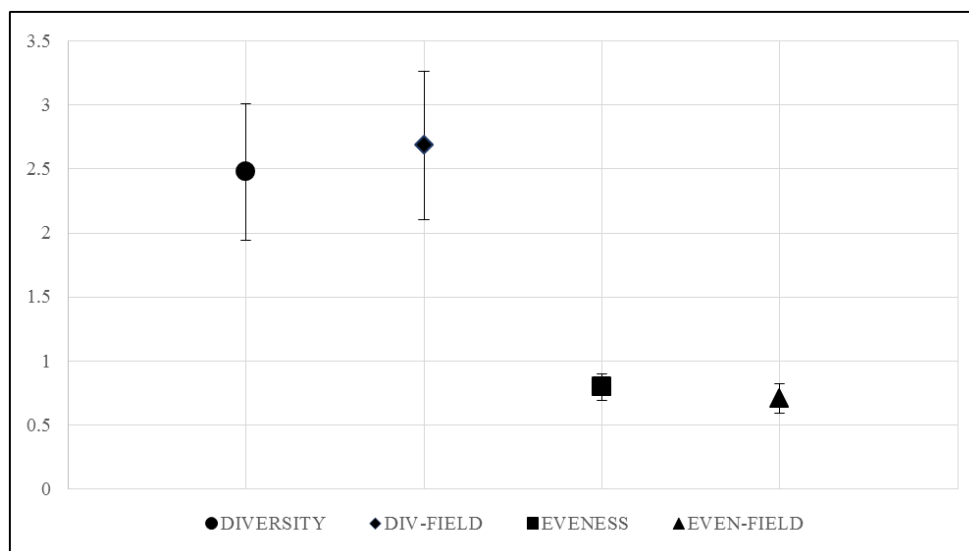


Figure 15. Range of Evenness estimates from 100 100-organism subsamples drawn from each of 72 field samples versus Richness (number of taxa) of the field samples.

Diversity of the field samples tended to be underestimated by the subsamples as the number of species increased (Figure 14). Evenness of the field samples tended to be overestimated by the random subsamples as the number of species in the field sample increased (Figure 15). The 100 calculated values

give some sense of the variability of the selected metric within a particular location. When all the values from subsamples are combined, the mean and standard deviation are presented in Figure 16.



**Figure 16.** Mean and standard deviation of Diversity and Evenness based on pooling all calculations of subsamples ( $N = 11900$ ) compared to combining values from 72 field samples.

Diversity and Evenness, like Richness, are characteristics of a given sample. To compare subsamples, it is proposed to follow the protocol used for Richness Ratios (R-Ratios) as follows.

$$\frac{MIN(DIV1,2)}{MAX(DIV1,2)} \quad (10)$$

Div1,2 are the Diversity values of two samples being compared.

$$\frac{MIN(EV1,2)}{MAX(EV1,2)} \quad (11)$$

EV1,2 are the Evenness values of two samples being compared.

The maximum value in each case, as for the R-Ratio, is 1.

The ratios are then compared to respective criteria

established based on pooled ratios using all subsamples from all field samples. The values of the criteria are set at the same frequency of occurrence as done for other metrics discussed in this paper (i.e., 15%).

Characterizing variability of estimates of an index's value from subsamples permits setting decision criteria so that there is consistency in confidence limits for those indices as presented in 3.4.

### 3.4. Proposed Criteria

The first edition of EPA's guidance for evaluating aquatic macro invertebrates included criteria only for the R-Ratio (Similarity) with  $>0.8$  for two subsamples indicating they are from the same population and Community Loss with  $<0.5$  for two subsamples indicating they are from the same population. The percentage of subsample comparisons not meeting those criteria are presented in Table 2.

**Table 2.** Percent Type 1 errors in subsamples of respective field samples based on using EPA criteria for 72 field samples sampled 100 times for 100 organisms mathematically and randomly with replacement. Note that the two EPA recommended criteria yield different estimates of Type 1 errors.

	Species Richness Ratio (Similarity) Criterion with those $<0.8$ assumed to be from a different population when they are not	Community Loss Criterion with those $>0.50$ assumed to be from a different population when they are not
Percent Type 1 Errors In 100 Subsamples	# Out of 72 field samples	# Out of 72 field samples
$>5\%$	60	20
$>10\%$	42	9
$>15\%$	22	4
$>20\%$	13	3

The decision criteria for Richness Ratio (Similarity) of  $>0.8$  and Community Loss of  $<0.5$  originally proposed by

EPA have different percentages of Type 1 errors. For example, using 15% for Type 1 errors as being acceptable, 22 of the 72 field samples exceed this target for Richness Ratio while 9 of the of the 72 exceeded the criterion for Community Loss (Table 2). The Richness Ratio criterion of >0.8 falls far short of the goal for no greater than 15% Type 1 errors and would have to be revised. The Community Loss criterion of <0.5 is more stringent than is the goal of no more than 15% Type 1 errors, but could be improved by revising the value as proposed later.

Based on the data presented, the authors are proposing criteria for all the indices evaluated in this research (Table 3).

**Table 3.** Proposed criteria for indices when comparing two samples at alpha= 0.15 or 0.85. Values with “<” preceding mean values less than this implies samples are different (alpha= 0.15), while values with “>” preceding mean samples with higher values are different (alpha= 0.85).

Index	Criteria for alpha=0.15	Based on N values
R-RATIO	<0.80	11,654
JACCARD	<0.48	7,695
SORENSEN	<0.65	7,695
BRAY-CURTIS	<0.68	11,828
DIVERSITY RATIO	<0.901	11,900
EVENNESS RATIO	<0.933	11,900
BRAY-CURTIS DISSIMILARITY	>0.32	11,828
COMMUNITY LOSS	>0.41	7,695

An alpha value of 0.15 or 0.85 as appropriate was chosen as reasonable for the inherent difficulties of field work. The proposed criteria place the two EPA values at the same confidence limit. It turns out that the >0.8 value for R-Ratio (Similarity) was close to the value chosen here.

3.5. Application of Criteria

The proposed criteria are derived from a large pool of data, so it is likely that more measurements will not change the values. However, since the data pool represents a set of field samples with a wide range of characteristics, any single metric is not robust enough to work for subsamples from every field sample. Therefore, it is recommended that a combination of these proposed criteria be used in a multi-metric application. This means testing each pair of samples by applying the values of each of the metrics listed in Table 3 using the proposed criterion for each metric. Either Bray-Curtis Similarity and Bray-Curtis Dissimilarity results are included but yield identical assessments and therefore are redundant. If one or more of the metrics results in the pair of samples being considered to be from the same population, it is judged that this is in fact true. Recall all pair comparisons are from the same population (i.e., field sample), so Type 1 are the only errors possible. All of the 72 field samples, one just barely, met at least one of the metrics.

4. Conclusions

Estimates are inherently uncertain. This is axiomatic. Important environmental decisions often also have major financial implications. For both reasons, it is imperative,

therefore, to have the best information available. This includes estimating the uncertainties associated with the data and information driving decisions. The work here quantifies the uncertainties some of the tools used in water quality evaluations and decisions through macroinvertebrates. The proposed decision criteria for the metrics examined places each of them at the same confidence level.

The work here can be expanded. Different ways of calculating Diversity and Evenness metrics might establish criteria more discriminating than are simple ratios. In addition, the size of the samples discussed here can be varied to assess its impact on the calculated values of these specific indices. For the samples analyzed here, [12] has provided a start for such an evaluation. On a broader, more general scale such impacts have been explored by [11]. Other future efforts should involve vetting the criteria proposed here using field samples composed of more than 100 organisms. Also, the proposed criteria could be compared to results derived from other analyses, mathematical or not, used to compare aquatic macroinvertebrate communities. Finally, large field samples could be used to evaluate Type II errors for the criteria proposed here to balance Type I and Type II errors explicitly for each of the proposed criteria.

The results here address a major concern about uncertainties in unvetted metrics voiced in [1].

Conflict of Interests

The authors declare that they have no competing interests.

Acknowledgements

The authors wish to thank the Vermont Department of Environmental Conservation and especially Steven Fiske for kindly sharing their data. Also, we appreciate Massachusetts Department of Environmental Protection personnel, especially Gerald M. Szal for his perceptive critique and recommendations, along with Christine Duerring’s and Alice Rojko’s aid in analyzing the data sets used in this paper and comments by Mark Mattson. The authors thank S. Lynch of the University of Rhode Island for additional calculations. The authors also appreciate several anonymous reviewers for their constructive critiques and useful recommendations.

References

[1] Norris, R. H. and A. Georges (1993) in Freshwater Biomonitoring and Benthic Macroinvertebrates D. M. Rosenberg and V. H. Resh (eds.), Chapman and Hall, Springer (US).

[2] Simpson, E. H. (1949) Measurement of diversity, Nature 163: 688.

[3] Shannon, C. E. and Weaver, W. The mathematical theory of communication, University of Illinois Press, Urbana 1949.

[4] Margalef, R. (1958) Information theory in ecology. General Systems 3, 36–71.

- [5] Cairns, J. Jr. and Dickson, K. L. (1971) A simple method for the biological assessment of the effects of waste discharges on aquatic swelling organisms. *J. Water Pollut. Control Fed.* 233 43: 755-772.
- [6] Plafkin, J. L. Barbour, M. T. Porter, K. D. Gross, S. K. and Hughes R. M. (1989) Rapid bioassessment protocols for use in streams and rivers: Benthic macroinvertebrates and fish. EPA/440/4-89-001. U. S. Environmental Protection Agency, Office of Water, Washington, DC.
- [7] Barbour, M. T. Gerritsen, J. Snyder, B. D. and Stribling, J. B. (1999) Rapid bioassessment protocols for use in streams and wadeable rivers: periphyton, benthic macroinvertebrates and fish, Second Edition. EPA 841-B-99-002. U. S. Environmental Protection Agency; Office of Water; Washington, D. C.
- [8] Karr, J. R. et al. (1986) Assessing biological integrity in running waters a method and its rationale, Illinois Natural History Survey Special Publication 5.
- [9] Davis, W. S., Snyder, B. D., Stribling J. B., and Stoughton, C. (1996) Summary of state biological assessment programs for streams and wadeable Waters. EPA 230-R-96-007. U. S. Environmental Protection Agency, Office of Policy, Planning and Evaluation; Washington, D. C.
- [10] Carter, J. L. and Resh, V. H. (2013) Analytical approaches used in stream benthic macroinvertebrate biomonitoring programs of State agencies in the United States: U. S. Geological Survey Open-File Report 2013-1129, 50 p.
- [11] Doberstein, G. P. Karr, J. R., and. Conquest, L. L. (2008) The effect of fixed-count sampling on macroinvertebrate biomonitoring in small streams, *Freshwater Biology*, 44, 355-371.
- [12] Lynch, S. (2002) Evaluation of several metrics of benthic macroinvertebrates. MS Thesis in Statistics, University of Rhode Island.
- [13] Aazami J., et al., (2015) Monitoring and assessment of water health quality in the Tajan River, Iran using physicochemical, fish and macroinvertebrates indices. *J Environ Health Sci Eng.* 13: 29. doi: 10.1186/s40201-015-0186-y. PMID: 25949817; PMCID: PMC4422490.
- [14] Dieu, T., et al., (2021) Invertebrate turnover along gradients of anthropogenic salinisation in rivers of two German regions, *Science of The Total Environment*, Volume 753, 141986 8624-2.
- [15] Huttuen K. L., et al. (2017) Habitat connectivity and in-stream vegetation control temporal variability of benthic invertebrate communities, *Sci Rep.* 7: 1448, 10.1038/s41598-017-00550-9.
- [16] Serrana J. M., et al., Ecological influence of sediment bypass tunnels on macroinvertebrates in dam-fragmented rivers by DNA metabarcoding, *Scientific Reports* 8: 10185 DOI: 10.1038/s41598-018, 2018.
- [17] Wang, L., et al., B. P. (2022) Species Diversity and Community Composition of Macroinvertebrates in Headwater Streams of Two Subtropical Neighboring Lowland Basins. *Diversity*, 14, 402. <https://doi.org/10.3390/d14050402>
- [18] Smith, B. J., et al., (2018) Comparison of aquatic invertebrate communities in near-shore areas with high or low boating activity, *Journal of Freshwater Ecology*, 34: 1, 189-198, DOI: date10.1080/02705060. 1556746.
- [19] Green, R. H. (1976) Some methods for hypothesis testing and analysis with biological monitoring data. in *Biological Monitoring of water and effluent quality*, ASTM STP 607, J. Cairns, Jr., Dickson, K. L., and Westlake, G. F. (eds.), pp. 200-211. American Society for Testing and Materials, West Conshohocken, PA.
- [20] Yong, C. and Epifanio, J. 2010 Quantifying the responses of macroinvertebrate assemblages to simulated stress: are more accurate similarity indices less useful? *Methods in Ecology and Evolution*, 1, 380–388 doi: 10.1111/j.2041-210X.2010.00040.x.
- [21] Spellerberg I. F. and Fedor, P. J. (2003) A tribute to Claude Shannon 1916-2001) and a plea for more rigorous use of species richness, species diversity and the Shannon-Wiener Index, *Global Ecology & Biogeography*, 12, 177-179.