Environmetrics, Volume 17, Issue 3 (May 2006), Pages 285-307. doi:10.1002/env.744

Full title: Stream Health Index for the Puget Sound Lowland **Short title:** SHIPSL

Authors: Grace Chiu*, Peter Guttorp

Affiliation: Department of Statistics, University of Washington

**E-mail: grace@stat.washington.edu*
Tel: 206-616-9262
Fax: 206-685-7419
Address: Box 354322, Seattle, WA 98195, U.S.A.

This research was made possible through a Postdoctoral Fellowship to G. Chiu co-funded by (1) the Pacific Institute for the Mathematical Sciences, and (2) Professor Richard Lockart and the Department of Statistics and Actuarial Science at Simon Fraser University, Burnaby, B.C., Canada; and partially through a Postdoctoral Fellowship to G. Chiu by the Natural Sciences and Engineering Research Council of Canada.

SUMMARY

Stream health is often measured by the multimetric *benthic index of biotic integrity* (B-IBI). For Puget Sound Lowland (PSL) streams, the B-IBI comprises ten metrics which quantify the well-being of benthic inhabitants of the stream. Each metric is converted to a score of 1, 3, or 5, where a higher value indicates a healthier stream with respect to the metric. Summing the metric scores yields the B-IBI. Stream health is then rated as *very poor, poor, fair, good*, or *excellent* according to the index value.

Different metrics may be measured on different scales. A scoring scheme is therefore required to standardize values across metrics. Conventional scoring for the B-IBI metrics requires subjective and space/time-dependent input on the cut-off points for discretizing each metric. In contrast, simple statistical standardization (centering and division by the standard deviation) appears to be more natural, is non-study-specific, and maps the metric space onto a continuous scale centered at 0. Comprising the same ten metrics of the PSL B-IBI, our *stream health index for the Puget Sound Lowland* (SHIPSL) is the sum of all statistically standardized scores. We use benthic taxonomy data on field samples from 1997 to develop a SHIPSL-based rating scheme of grades A to F for stream health. We also discuss an alternative version of SHIPSL defined using "gold-standard" values.

Bootstrap simulations are used to compare the performance of the two versions of SHIPSL to B-IBI in reflecting PSL stream health. Results show that both versions of SHIPSL are more efficient in measuring stream health. Without sacrificing information on underlying biological conditions, SHIPSL reduces bias and variability of the health index, and eliminates sensitivity of the rating to slight changes in the metric scoring scheme.

KEYWORDS: biomonitoring, biotic integrity, stream health, metric scoring, bias and variability reduction

1 INTRODUCTION

Monitoring the biological well-being of ecological systems over time is key to the preservation of our natural environment. The practice is often known as *biomonitoring*. It involves assessment of a non-human biological system for changes due to human activities. For the Puget Sound Lowland (PSL) of Washington state, monitoring stream health has been an integral part of local environmental research.

Karr (1981) first developed the multimetric *index of biotic integrity* (IBI) for assessing the health of Midwestern freshwater systems by quantifying biological conditions of fish communities. The term "biotic integrity" generally refers to the state of well-being of an ecological system that has undergone minimal human influence (Karr, 1999; Karr and Chu, 1999). The IBI was developed to gauge health relative to this reference point. Ever since, the IBI has been refined (Karr *et al.*, 1986), localized (Environmental Protection Agencies of coastal and inland states; other countries, e.g. Australia, Canada, France, Mexico), and adapted to examining other animal communities (e.g. benthic invertebrates (Kerans and Karr, 1994), birds (Canterbury *et al.*, 2000)). All such versions of the index retain the multimetric structure of the original IBI. Conventional methods of selecting metrics to compose the index combine scrutiny of a qualitative nature and simple statistical analyses to screen a large pool of candidate metrics before arriving at a final subset (Karr and Chu, 1999). For PSL streams, Karr (1998) adapted the *benthic IBI* (B-IBI) by Kerans and Karr (1994) to include ten metrics (Table 1), while Bunea *et al.* (1999) propose a screening procedure that is statistical and can be entirely automated, although extremely computer intensive.

Most PSL B-IBI metrics (e.g. total number of taxa) yield count data over a large range, while some give percentages over a continuous scale from 0 to 100. To combine health information provided by various metrics measured on different scales, Karr *et al.* (1986) devised the following metric scoring scheme:

- from the study region, identify reference sites, i.e. sites along streams that are least impacted by human activities;
- for a given metric:
  - rank all field samples from reference sites (often multiple samples from same sites) from worst to best health with respect to the metric;
  - if necessary, adjust this ranking for its relationship with stream size;
  - trisect the resulting ranking;
  - use this reference trisection to determine cutoff values on the metric scale for assigning a score of 1 (worst), 3, or 5 (best).

As the set of metrics is specific to a certain study region (Brinck, 2002), local scoring criteria are recalibrated only under unusual circumstances. For example, many PSL benthic species that were taxonomically identified as non-long-lived in the mid-1990's have recently been identified as long-lived. Thus, the scoring criteria for 1994 are different from those used for, say, 1997 for all ten metrics (Table 1). Once the scoring criteria are considered appropriate for a given year, each PSL stream site being studied receives a B-IBI value equal to the sum of its ten metric scores. Therefore, the PSL B-IBI is an even number between 10 and 50.

Devising a metric scoring scheme as above is, by design, subject to personal judgment and local policy preferences (Boulton, 1999; Lackey, 2003). This is particularly true in selecting metrics and reference sites (McCormick *et al.*, 2001) and trisecting stream rankings. While the former relies on the knowledge of and painstaking efforts by experts in the subject matter, the latter often involves fitting lines by eye through points on a graph (Karr *et al.*, 1986; Harris and Silveira, 1999; Karr and Chu, 1999; USDA-NRCS, 2003). Moreover, there has been concern over the arbitrariness of the score values, so that other discrete or continuous scales

have been proposed for the scoring (USDA-NRCS, 2003). Nevertheless, this process of calibrating scoring criteria continues to be widely practiced for IBI metrics.

For the PSL, Bunea *et al.* (1999) propose two statistically based "trisection" schemes that entirely remove subjectivity and appear to perform no worse than the traditional trisection method. However, it remains unclear what values are most sensible for metric scores. Our goal here is to introduce a new health index for PSL streams whose metric scoring mechanism eliminates subjectivity and painstaking efforts, and, at the same time, is statistically sensible. We will also discuss the desirable statistical properties of this index.

2 SHIPSL: A NEW STREAM HEALTH INDEX

Our new *Stream Health Index for the Puget Sound Lowland* (SHIPSL) comprises the same ten metrics as the PSL B-IBI (Table 1). (We do not attempt to discuss the appropriateness of these metrics in this article.)

Let there be a total of n sites in the study, and let y_{ijk} denote the i-th site's value for metric j in the k-th replicate sample, $i = 1, \ldots, n$, $j = 1, \ldots, 10$, and $k = 1, 2, 3$. (Each PSL stream site is commonly sampled three times per study.) Thus, the data can be arranged as

$$\mathbb{Y}_k = \begin{pmatrix} y_{1,1,k} & y_{1,2,k} & \cdots & y_{1,10,k} \\ y_{2,1,k} & y_{2,2,k} & \cdots & y_{2,10,k} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n,1,k} & y_{n,2,k} & \cdots & y_{n,10,k} \end{pmatrix}, \qquad k = 1, 2, 3,$$

and each row comprises the ten metric values for that site's k-th replicate sample. In the spirit of the B-IBI, we average the y_{ijk}'s over k, and denote the resulting data by $\mathbb{Y} = \left[\overline{y}_{ij}\right]$, where $\overline{y}_{ij} = (1/3)\sum_k y_{ijk}$. Note that for B-IBI, "# long-lived taxa" and "# intolerant taxa" are pooled instead of averaged over replicates — i.e. the metric value is that of a super-sample formed by combining all three

replicate samples — then scored as 1, 3, or 5. For SHIPSL, however, metrics are always averaged over replicates for the sake of consistency.

2.1 *Standardization: simple and natural*

A statistician's most natural metric scoring scheme is simple standardization. For example, metrics that are positively associated with stream health (all but "% tolerant taxa" and "% three most dominant taxa") are scored as

$$z_{ij} = j\text{-th metric score for } i\text{-th site} = \frac{\overline{y}_{ij} - \overline{\overline{y}}_j}{s_j} \tag{1}$$

where $\overline{\overline{y}}_j$ and s_j are the mean and sample standard deviation (SD), respectively, of column j of $\mathbb{Y}$. For the two metrics that are negatively associated with health, z_{ij} is the negative value of the standardized $\overline{y}_{ij}$. Similar to B-IBI, we take the i-th site's SHIPSL value to be the sum of it's metric z-scores and denote it by $w_i = \sum_j z_{ij}$.

In general, should the association between metric j (for some j) and stream size be deemed significant, raw metric values y_{ijk}'s may be adjusted via, say, regression, before being averaged over k then converted to z_{ij}'s (see Section 4). However, information from available sources (e.g. Morley, 2000; SalmonWeb) indicates that PSL streams considered in recent B-IBI studies fall in the "small" or "low stream order" category, and correction for stream size is apparently unnecessary.

Simple standardization is a statistically-based calibration method that is systematic and eliminates personal judgment. The same method can be equally applied in any geographical region and under any time frame regardless of the current / local protocol for taxonomic definitions. While spatial and temporal scales affect the metric score values through the mean and SD that appear in the scoring equation, the *scoring method* is entirely scale-independent. This is an advantage over calibration methods, such as those employed by different versions of the IBI, which require experts' input that may greatly differ over time and space. Moreover, metric z-scores are continuous and centered at 0, and similarly

for SHIPSL values, w_i's, by definition. One can expect most SHIPSL values to fall within ± 20 based on the properties of standardized scores. Therefore, the sign and magnitude of a metric z-score or SHIPSL value immediately provide intuitive information about the site's condition relative to all other sites being studied.

Note that the definition of SHIPSL eliminates the need for reference sites if one is interested in comparing sites within a single study, or if the same set of sites is monitored over time. When a value is standardized, it is automatically scored against the extreme values of the dataset. As least- and most-impacted sites are expected to produce extreme values for most metrics, SHIPSL z-scores should be no less effective than B-IBI metric scores in distinguishing between healthy and unhealthy sites, so long as both kinds of extremes are included in the study.

2.2 *Gold-standard SHIPSL for monitoring health over time*

For monitoring a single site i over subsequent years without collecting new data from other sites, the same field data on all sites from the current year may be used, with data on site i replaced by newly collected data. Metrics are then rescored over all sites and new SHIPSL values computed. This way, all sites from this year except site i act as "reference sites" in subsequent years.

A similar approach involves using "gold standard" values for the metric mean and SD in the scoring formula. Denoted by μ_j and σ_j, these pre-determined gold standards respectively replace $\overline{\overline{y}}_j$ and s_j of (1). The resulting metric score is

$$z_{ij}^{(g)} = \frac{\overline{y}_{ij} - \mu_j}{\sigma_j}\,. \tag{2}$$

The SHIPSL index thus defined is referred to as the *gold-standard SHIPSL* (GS-SHIPSL). GS-SHIPSL for site i is then $w_i^{(g)} = \sum_j z_{ij}^{(g)}$.

Much like the speed of light relative to which the world's fastest traveling objects are gauged, gold standard mean and SD provide pivot points upon which metric values are weighed. Such gold standards may be computed using, say, ob-

servations from a randomly chosen year made on "reference" sites which are randomly chosen from a nation-/continent-wide census-enumeration-type database. One may then reuse the same gold standards, or compute new values from resampled year-site combinations, neither of which involves a subjective definition of reference. (The idea of gold standards is not restricted to SHIPSL but possibly to other multimetric biological indices.) Section 3.6 below examines the performance of GS-SHIPSL. The reader may refer to Chiu and Guttorp, 2004 for a more detailed discussion of the properties of GS-SHIPSL. The non-gold-standard version of SHIPSL will be referred to as "ordinary SHIPSL" or simply "SHIPSL" in the remainder of this article.

2.3 *SHIPSL versus B-IBI: a case study*

To compare B-IBI's and SHIPSL's performance in reflecting stream health, we apply these indices to the 1997 and 1998 datasets taken from Morley, 2000. Rock Creek and Thornton Creek were sampled in both years. Respectively, they are among the least- and most-impacted streams of the PSL region (Karr, 1998). Therefore, we expect the SHIPSL metric scoring scheme to be reasonable here. As universal values of metric mean and SD are currently rare in practice, GS-SHIPSL is excluded from this comparison.

Figure 1 shows the distributions of B-IBI and SHIPSL values (upper two panels), as well as their scatterplots (lower panel). Although they reside on different scales, the two indices are highly correlated (0.965 and 0.979, respectively, for 1997 and 1998) and have similar distributional shapes. The scatterplots show that the high correlation is not solely driven by the streams on the extremes of the B-IBI and SHIPSL scales. Moreover, SHIPSL's continuous scale removes some of the clustering in B-IBI values, as is apparent in the tails of the histograms and the center of the scatterplots where many data points are vertically aligned. Altogether, SHIPSL is shown here to reflect most of the information carried by the

conventional B-IBI while being more naturally intuitive and much easier to be localized (see Section 2.1 above).

Two more scatterplots are shown in Figure 2. Each has an x-axis of "percentage of land cover (urbanized area) for the stream's basin." Morley (2000) uses this variable to measure the degree of impact on streams. The y-axis is "index value," although the SHIPSL plotted is rescaled to have a similar range as B-IBI on the plot. This way, the features of the indices' relationships with impact are overlaid and can be easily compared. Again, we see that both B-IBI and SHIPSL reflect similar information, and in particular, the dependence between stream health and human influence. However, the middle region of the lower plot (1998) shows a tighter clustering of SHIPSL values, i.e. SHIPSL yields a smaller disparity among mid-ranked streams. This phenomenon is clear for Little Bear Creek Basin (x=54), Laughing Jacobs Basin (x=59), and Swamp Creek Basin (x=70).

Streams on the extremes of the health scale can be easily rated by a knowledgeable naturalist without the use of any health index. Thus, the importance of a stream health index lies in how accurately and precisely it reflects the conditions of streams around the center of the health distribution. Besides 1998, does SHIPSL generally portray these mid-ranked streams as less variable than does B-IBI? If so, is SHIPSL not reliably reflecting the underlying state of nature, or is B-IBI too prone to chance variation, yielding unnecessary disparity between sites of similar health? We investigate this in Section 3 below.

3 EFFICIENCY AND SENSITIVITY: A BOOTSTRAP STUDY

Fore *et al.* (1994) conduct several bootstrap simulations to investigate statistical properties of the IBI. The desirable properties reported include approximate normality and constant variance, and efficiency in distinguishing sites (high power in

pairwise t-tests). A negative property reported is the index's biasedness. These results have since been extrapolated to other versions of IBI (Karr and Chu, 1999).

In this article, we employ bootstrapping in a different setup. Only one bootstrap simulation of 10,000 samples is obtained. Subsequently, all comparisons between B-IBI and SHIPSL are made based on the same bootstrap samples.

Our investigation is broken down as follows. The field data from which bootstrap samples are drawn are described in Section 3.1. In Section 3.2, biasedness of B-IBI is further investigated and contrasted against SHIPSL's unbiasedness. Additionally, the precision of the two indices is compared. In Section 3.3, the indices' efficiency in detecting pairwise differences between sites is examined. In Section 3.4, we investigate the performance of the B-IBI five-point classification scheme of stream health, and its sensitivity to field noise. A new SHIPSL-based six-point grading scheme is proposed and studied in Section 3.5. Finally, Section 3.6 summarizes some findings of Chiu and Guttorp (2004) about GS-SHIPSL.

3.1 *The data*

Our bootstrap study is based on the 1997 PSL data of Morley, 2000. Three replicate field samples were obtained from each of 18 sites (see Note (3) below). Benthic organisms were taxonomically identified. Thus, metric values can be computed for each replicate field sample and converted to B-IBI metric scores (1, 3, or 5) and SHIPSL z-scores. These scores are then summed accordingly to yield B-IBI and SHIPSL values. We refer to these as the *observed* index values.

Bootstrap samples are drawn from each replicate field sample to retain the structure of within-site variability. For a field replicate containing n benthic organisms, a bootstrap replicate is obtained by randomly sampling n individuals with replacement. Metric values, B-IBI and SHIPSL metric scores, and B-IBI and SHIPSL index values are subsequently computed from the bootstrap replicates. We obtain 10,000 sets of bootstrap replicates per site. The consistently

high correlations between bootstrap B-IBI and SHIPSL values (between 0.92 and 0.99, with mean 0.96) confirm the notion from Section 2.3 that SHIPSL preserves exceptionally well the biological information contained in B-IBI.

Notes. (1) Only metric scores and final B-IBI values are reported in Morley, 2000. We separately obtained the corresponding raw benthic taxonomy data that were required for the bootstrap analysis. (2) Our computation using these raw values yields a slight discrepancy in the *Plecoptera* metric score for site LB4 (1 instead of 3 as in Morley, 2000). (3) The raw data include different replicate sets for Thornton Creek (TH1): a set of only two replicates observed on an earlier date, and a set of three replicates observed on a later date. One can verify that Morley bases her TH1 values on the earlier set. Instead, we use the later set for our bootstrap study so that there are three replicates for all 18 sites.

3.2 *Bias in sample mean and standard deviation*

In their bootstrap study, Fore *et al.* (1994) show that IBI is significantly biased. They argue that it is likely due to the discrete three-point scale for metric scores. Our results confirm that the PSL B-IBI also exhibits this biasedness, as shown by the upper left histogram of Figure 3. It suggests that for a bootstrap field sample, the 18 sites have an expected mean B-IBI of around 26.95, which is 2.77 standard errors below the observed sample mean of 27.67. This underestimation is highly significant. In contrast, SHIPSL has zero mean by definition; hence, all bootstrap samples have mean SHIPSL values that are exactly 0 (Figure 3, lower left histogram). Therefore, using SHIPSL in place of B-IBI entirely eliminates bias of the sample mean.

We can also consider biasedness for individual sites. The mean of a given site's bootstrap B-IBI or SHIPSL distribution differs from the site's observed index value (Figure 4; Table 2) due to chance variation and possibly bias. Fore *et al.* note that site-wise bias for IBI is negatively correlated with index value. Using Table 2,

we see that this is also true for the PSL B-IBI ($r = -0.63$ with all sites, one-sided $p = 0.00$; $r = -0.41$ with most influential site removed, one-sided $p = 0.05$), but unlikely so for SHIPSL ($r = -0.23$, one-sided $p = 0.19$).

Statistical *accuracy* or unbiasedness is less relevant if bias can be efficiently estimated and corrected for accordingly. However, the overall effectiveness of a measure (after bias correction, if necessary) in reflecting underlying conditions depends heavily on the measure's *precision*: the higher its variability or uncertainty, the less precise it is. Figure 3, right panel shows the distributions of sampling variability (standard deviation) for B-IBI and SHIPSL bootstrap field samples, respectively. Note that the histogram's SD is approximately the standard error of the SD of a field sample.

The upper histogram's central value is 8.46, which is 1.72 standard errors below the actual SD of 9.06 for the observed field sample in 1997. This indicates that B-IBI's true variability is significantly underestimated (5% level), and hence, is more severe than what can be measured from a field sample. Ignoring this negative bias would be unwise. For instance, hypothesis testing may be used to compare the health conditions between sites (Fore *et al.*, 1994; Karr and Chu, 1999). As the technique involves division by SD, a minor difference in health may be mistaken to be significant due to a sample SD that is falsely small. Similar logic applies to confidence intervals (C.I.'s) for an underlying IBI value. In contrast, the underestimation is minimal for SHIPSL (the lower histogram's central value is less than half a standard error below the observed field sample SD), and could well be due to the artificial resampling environment of the bootstrap.

A proper comparison between B-IBI's and SHIPSL's precision involves correction for bias in the bootstrap SD's (BSD's) and transformation onto a common scale. The former merely requires adding a correction factor equal to the absolute difference between the mean of the BSD histogram and the observed field sample SD. However, it is unclear what is required for the latter. An *ad hoc* transfor-

mation similar to the rescaling for Figure 2 is to (i) divide the bootstrap B-IBI index values by the mean of the 10,000 BSD's, and similarly for SHIPSL; (ii) produce BSD histograms for these rescaled B-IBI and SHIPSL distributions (both are now centered at 1); and (iii) correct these histograms for bias. The resulting distributions are shown in Figure 5(a). We see that the intrinsic variability is 1.07 for B-IBI and 1.01 for SHIPSL. That is, B-IBI appears to be more variable than SHIPSL by a factor of 1.06 or an extra 6%. Figure 5(a) also shows a smaller spread for the SHIPSL BSD's, suggesting that the sample SHIPSL SD is a more reliable measure of variation.

In practice, variability or uncertainty in a measure translates to inefficiency or weak power in detecting the true state of nature, and insensitivity to true change. Section 3.3 below further investigates this issue.

Another way to perceive a measure's uncertainty is its sensitivity to random noise. Aside from the intrinsic variability of a measure, chance variation exists in field data, such as the differences among replicate samples taken from the same site. Therefore, if a field dataset exhibits extra noise, B-IBI can potentially show extra fluctuation that does not necessarily represent disparity in true health among sites. In this case, the health rating for a stream based on its B-IBI value may be less reliable. This aspect is further discussed later in Section 3.4.

3.3 *An ad hoc power analysis*

To determine how much more readily SHIPSL is in detecting a difference between sites than is B-IBI, we examine the two indices' empirical coverage of 95% bootstrap C.I.'s for pairwise differences in mean.

First, suppose we knew the *joint probability distribution* of B-IBI measurements for all 18 sites, and likewise for SHIPSL. Then, for both indices, we could develop an exact formula for a 95% C.I. for mean difference (i.e. difference in mean) between any two sites, and compute the actual power of rejecting the null

hypothesis of no difference when 0 is excluded from the C.I.

In reality, we approximate the marginal B-IBI and SHIPSL distributions by their respective bootstrap distributions (Figure 4). Moreover, as we do not have the theoretical formula for the C.I. for mean difference, we compute it based on the bootstrap distribution of difference in measurement between two sites.

Let δ denote the mean difference between two bootstrap distributions. Each distribution yields a single bootstrap C.I. which does not allow assessment of power given δ. Proper analysis of power as a function of δ would involve bootstrapping the bootstrap (resampling many times from a bootstrap distribution) which is extremely computer-intensive due to the large number of sites. Instead, we consider the power of detecting various non-zero δ's as a whole. For each of B-IBI and SHIPSL, there are a total of 153 pairwise comparisons over the 18 sites, and hence, 153 distinct non-zero values of δ. (The set of B-IBI-δ's is, of course, different from the set of SHIPSL-δ's. See Table 2.) Our *ad hoc* power analysis considers the coverage of the 153 bootstrap C.I.'s computed at an individual confidence level of 95% (Table 3). Coverages are 26.8% and 18.3%, respectively, for the B-IBI and SHIPSL C.I.'s. That is, SHIPSL reduces coverage by over 8%, or, the power to detect a non-zero difference between two sites is improved by almost 12% ($81.7/73.2-1$) when SHIPSL is used in place of B-IBI for these data. Furthermore, the off-diagonal entries indicate the index's tendency to detect a difference which the other index has missed, and SHIPSL is twice as efficient in this respect. Overall, improvement in power is likely due in part to a continuous scale for SHIPSL instead of an even-numbered B-IBI scale which makes it difficult to detect a mean difference of less than two. Weakened power due to scale compression is discussed by Blocksum (2003) and summarized by Chiu and Guttorp (2004).

3.4 *Stream health rating and its sensitivity to noise*

The widely used five-point rating of "very poor" to "excellent" is adapted for PSL

streams from the fish IBI-based classification by Karr *et al.* (1986). Due to B-IBI's weaker power, this rating scheme possibly underclassifies some pairs of sites with significantly different B-IBI values. This phenomenon is analogous to a Type II error in hypothesis testing.

While the goal in Section 3.3 was to examine how readily a C.I. could detect a difference in health between any two given sites, here, we instead consider the the set of health ratings for all 18 sites as a whole. Therefore, all 153 pairs of sites are compared simultaneously at a family-wise 5% significance level. Table 4 shows simultaneous 95% C.I.'s for the mean difference of those pairs in 1997 that are underclassified by the B-IBI rating scheme. (See **Remark 1** below for the computation of such C.I.'s.) Note that no sites are underclassified by the SHIPSL six-point grading scheme, which will be discussed in Section 3.5 below.

One may expect overclassification (analogous to a Type I error) to be not serious when underclassification is predominant. However, note the pairwise comparisons that yield different ratings based on this scheme, despite 0 being covered by the simultaneous 95% C.I.'s of mean differences (top half of Table 5).

Moreover, as noted in Section 3.2, this five-point classification is potentially highly sensitive to extra field noise. To this end, we make slight changes to all cut-points in the metric scoring criteria and investigate the resulting changes in health rating. Although not equivalent to adding extra noise to the field data, jittering the cut-points in the metric scoring criteria suggests how sensitive the B-IBI-based rating scheme could be to increased variability in field samples.

For each of the eight metrics whose replicate values are averaged before scoring, noise that is small relative to the original 1-3 and 3-5 cut-points is independently generated from a normal distribution, then added to the cut-points. For the two metrics (long-lived and intolerant taxa counts) whose values are obtained from pooled replicate field samples, similar noise — but from a Poisson distribution — is added to the corresponding cut-points. (See **Remark 2** below for more details.)

The resulting "jittered" scoring criteria are shown in Table 1.

Figure 6 shows examples that demonstrate the sensitivity of the five-point rating to these changes. Points outside the diagonal blocks indicate stream health ratings that are altered by the jittered scoring criteria. (Those sites whose bootstrap health ratings remain unaltered are JE1 (shown), BB2, BB3, BB4, LB1, LB3, and MI1 (not shown).) Note that each plot consists of 10,000 bootstrap values, many of which overlap due to the discrete scale for B-IBI. To unmask the extent of the alteration, we tabulate the off-diagonal elements in Table 6. The breakdown suggests that stream health rating may change as often as 50% of the time, and the change tends to move a site towards the middle category of "fair" health.

Remark 1

In previous sections, bootstrap C.I.'s at an individual 95% confidence level were obtained by taking the 2.5-th and 97.5-th percentiles of the bootstrap distributions. Another way to obtain bootstrap C.I.'s is via the quantiles of bootstrap Studentized values. Denote the bootstrap differences between sites i and j by δ^*_{ijk}, for $i, j = 1, \ldots, 18$, $i \neq j$, and $k = 1, \ldots, 10\,000$. The Studentized differences are $t_{ijk} = (\delta^*_{ijk} - \overline{\delta}^*_{ij})/s^*_{ij}$, where $\overline{\delta}^*_{ij}$ and s^*_{ij} are the mean and SD, respectively, of the bootstrap sample of δ^*_{ijk}'s. For fixed (i, j), let ℓ_{ij} and u_{ij} be the lower and upper $100(\alpha/2)$-th percentiles of the t_{ijk}'s. An individual $100(1-\alpha)\%$ bootstrap C.I. for the unknown δ_{ij} is then

$$[\overline{\delta}^*_{ij} - u_{ij}s^*_{ij}, \overline{\delta}^*_{ij} - \ell_{ij}s^*_{ij}]\,. \tag{3}$$

In this section, the quantiles ℓ_{ij}'s and u_{ij}'s need to be adjusted for a family-wise confidence level of 95%. We wish to have

$$\begin{aligned} 0.95 &\leq P\left\{\overline{\delta}^*_{ij} - u_{ij}s^*_{ij} \leq \delta_{ij} \leq \overline{\delta}^*_{ij} - \ell_{ij}s^*_{ij} \text{ for all } i \neq j\right\} \\ &= P\left\{\ell_{ij} \leq \frac{\overline{\delta}^*_{ij} - \delta_{ij}}{s^*_{ij}} \leq u_{ij} \text{ for all } i \neq j\right\} \end{aligned}$$

$$\leq \quad P\left\{\min_{i,j}\ell_{ij} \leq \frac{\overline{\delta}^*_{ij}-\delta_{ij}}{s^*_{ij}} \leq \max_{i,j} u_{ij} \ \text{ for all } i \neq j\right\}$$

$$= \quad P\left\{\min_{i,j}\frac{\overline{\delta}^*_{ij}-\delta_{ij}}{s^*_{ij}} \geq \min_{i,j}\ell_{ij} \ \text{ and } \ \max_{i,j}\frac{\overline{\delta}^*_{ij}-\delta_{ij}}{s^*_{ij}} \leq \max_{i,j} u_{ij}\right\}. \qquad (4)$$

Write $t_k^{(0)} = \min_{i,j} t_{ijk}$, $t_k^{(1)} = \max_{i,j} t_{ijk}$, $\ell = \min_{i,j} \ell_{ij}$, and $u = \max_{i,j} u_{ij}$. Then, the probability in (4) can be approximated by $\big(\#\{t_k^{(0)} \geq \ell\} + \#\{t_k^{(1)} \leq u\}\big)/10\,000$. That is, ℓ can be approximated by the 2.5-th percentile of the $t_k^{(0)}$'s, and u, by the 97.5-th percentile of the $t_k^{(1)}$'s.

The bootstrap distributions of t_{ijk}'s (not shown) exhibit heavy skewness and bias for B-IBI for some (i, j) pairs, but are unbiased and nearly normal for SHIPSL for all (i, j). Thus, for the former, the estimated quantiles should not be applied to (3) uniformly over all (i, j). Instead, we take the estimated SHIPSL quantiles of $\widehat{\ell} = -3.53658$ (for upper C.I. limit) and $\widehat{u} = 3.37374$ (for lower C.I. limit), and obtain the corresponding standard normal tail probabilities of 0.000203 and 0.000371. That is, simultaneous 95% C.I.'s for δ_{ij}'s (B-IBI or SHIPSL) have approximate individual confidence levels of 99.9426%. Subsequently, such C.I.'s are obtained by taking the 0.0371-st and 99.9797-th percentiles of the bootstrap δ^*_{ijk}'s.

Remark 2

Small random noise was independently generated for all 20 cutoff points in the B-IBI scoring criteria (Table 1). The distributions used were Normal$(0.25, 0.05^2)$ for *Ephemeroptera, Plecoptera, Trichoptera*, and clinger metrics; Normal$(0.5, 0.1^2)$ for total taxa; Normal$(1, 0.2^2)$ for the percent tolerant and percent dominant metrics; Normal$(0.1, 0.02^2)$ for percent predators; and Poisson(1) for long-lived and intolerant metrics. Generated values (all positive) were then multiplied by ± 1 at random.

3.5 *A new SHIPSL-based rating*

Based on the 18 bootstrap SHIPSL distributions for the 1997 data, we devise a

scale of six grade points: A, B, C, D, E, and F, in descending order of health (Figure 7). (We leave the physical interpretation of each grade point to the expertise of ecologists.) The cut-points for this scheme are chosen so as to minimize the overlap of bootstrap distributions between grades. With the exception of a few sites (possibly BB2 and JE1 for grades B and C, and BB1, LB2, and SW2 for grades C and D), slight random noise added to the original field samples would rarely cause the means of these SHIPSL distributions to fall into a different grade. Consequently, sensitivity of the rating scheme to extra field noise is reduced.

Another goal of this new scheme is to reduce possible underclassification of sites. Table 4 indicates that the six-point scale eliminates underclassification at a 5% family-wise significance level. However, the bottom half of Table 5 shows that the SHIPSL grading overclassifies some sites where the B-IBI five-point scale does not. This is likely due to a higher power in SHIPSL for detecting a true difference between sites (see Section 3.3).

Note that our grading scheme is developed solely based on the 1997 PSL data. Ideally, we would compare its performance in reflecting stream health across various years. However, except for the years 1994 and 1998, we have difficulty obtaining PSL data that are either published or accompanied by detailed documentation of the data's various attributes. As taxonomic identification of PSL benthic species has recently changed (see Section 1), putting the 1994 data together with later data poses a challenge: How would one assess the applicability of the same B-IBI or SHIPSL rating scheme across years over which metric scales have changed?

This issue is somewhat analogous to that in Brinck, 2002, between metrics identified by dataset-specific methods and the ten chosen B-IBI metrics. It is suggested that data-specific metric sets are inconsistent from year to year; hence, they may not properly reflect the underlying biological conditions being measured. In contrast, the B-IBI metrics are calibrated such that they are constant over time. From this, one may argue that the B-IBI scale is unaffected by changes in

the metric scales over time, whereas it may not be so for SHIPSL.

Here, we compare the SHIPSL and B-IBI ratings over 1994, 1997, and 1998 through Figure 8. The minimum SHIPSL value in 1994 is somewhat larger than that for the later years. This possibly suggests that the lower grades (E and F) of the 1997-based SHIPSL grading are not appropriate for 1994. However, the relative positions of modes, valleys, and inflection points among the B-IBI distributions are highly comparable to those among the SHIPSL distributions aside from the 1994 left tail. In addition, the "poor" and "fair" B-IBI categories closely resemble the "D" and "C" SHIPSL grades. Therefore, sites rated "very poor" by the B-IBI scheme apparently can be further broken down into SHIPSL grades "E" and "F". Note that the SHIPSL scheme seems to assign an "A" grade to slightly more sites than the B-IBI scheme would an "excellent" rating.

One may be tempted to conclude from Figure 8 that PSL streams have improved in overall health since 1994. However, these are cross-sectional data which cover different monitoring sites across different years. Sites along the same water channel are seldom sampled again at exactly the same geographical locations as previous years. Therefore, it is important to note that inference for health trend should not be drawn from distributions of stream health indices unless the same sites are involved over time.

3.6 *Performance of GS-SHIPSL*

As universal values are currently rare in practice, we treat $\{\overline{\overline{y}}_j, s_j\}$ of the original field sample from 1997 as $\{\mu_j, \sigma_j\}$ in the bootstrap version of (2). That is, for each of the 10,000 bootstrap samples,

$$z_{ij}^{(g)*} = \frac{\overline{y}_{ij}^{*} - \overline{\overline{y}}_j}{s_j}$$

is the (i, j)-th bootstrap GS-SHIPSL metric score, where "$*$" denotes a bootstrap value (as opposed to an observed value from the field, denoted without "$*$"). Below

is a summary of the findings for GS-SHIPSL based on Chiu and Guttorp, 2004.

Result 1 On the surface, GS-SHIPSL appears to behave similarly to B-IBI with respect to bias in sample mean and SD (see Figure 9, "before" curves) and to a negative correlation between site-wise bias and index value. Fore *et al.* (1994) attribute the latter phenomenon for B-IBI to heavy compression of extreme metric values into scores of 1 and 5. However, it does not explain the correlation for GS-SHIPSL, whose metric scores are allowed to move freely in either direction of an unbounded, continuous scale. A closer look at the bootstrap resampling mechanism reveals severe negative bias in bootstrap mean and SD for all seven taxa richness (count) metrics, but not for percentage metrics (Table 7, "count" columns). One reason for this bias is that a taxa richness metric tallies the presence of many different taxa, so that a taxon that is absent in an observed field sample is always absent in a bootstrap resample. Consequently, the bootstrap values of a taxa richness metric can never exceed the observed field value, and their range is severely reduced. In the context of resampling from field samples, this limitation of the bootstrap heavily distorts the randomness that occurs across samples in practice.

Result 2 When bias adjustments are made to the bootstrap distribution for each of the seven taxa richness metrics, GS-SHIPSL behaves more closely to SHIPSL than B-IBI (see Figure 9, "after" curves). In particular, biases in sample mean and SD are insignificant for GS-SHIPSL (one-sided $p \geq 0.4$ for both), and so is the negative correlation between site-wise bias and index value (one-sided $p = 0.2$). For B-IBI, although bias in sample mean is also insignificant here (one-sided $p = 0.4$), bias in sample SD is borderline significant at a 10% level, and the negative correlation remain highly significant (one-sided $p = 0.1$). Furthermore, (a) intrinsic variability of B-IBI remains higher than both versions of SHIPSL, (b) intrinsic variability is comparable between both versions of SHIPSL, and (c) the error

in sample SD is smaller for GS-SHIPSL than B-IBI (and smallest for ordinary SHIPSL; see Figure 5(b)). Thus, when artificial biases due to the bootstrap mechanism are corrected for, general conclusions from Sections 3.2 and 3.3 (except for the bias in B-IBI's sample mean) can be generalized to include either version of SHIPSL. Consequently, assuming that current protocols for collecting and identifying benthic organisms from the field produce metrics which well reflect underlying population conditions, both versions of SHIPSL have statistical properties that are (1) highly comparable, and (2) generally more desirable than those exhibited by the B-IBI.

4 DISCUSSION

Biomonitoring largely involves the tracking of human-induced environmental degradation over time. When subjectivity plays a major role in the monitoring scheme, scientific integrity of any conclusion drawn from such studies may be sacrificed. For example, ideas of what values of certain variables indicate a "healthy" ecosystem could be influenced by local policy preferences, thereby differing across geographical regions. Thus, one should be cautious about the current popular use of ecological health indices such as the IBI in devising public policy. Due to a lack of protocols that unify expert opinions across nation (continent), existing methodologies for gauging ecological health perhaps should be modified to reduce protocol-dependence. The SHIPSL scoring scheme is one such modification, as it removes personal judgment from metric calibration.

For PSL streams, B-IBI values reportedly vary little from one year to the next (Karr, 1998; Brinck, 2002); yet our findings in Section 3.2 indicate otherwise. As a measure, B-IBI seems volatile: it underestimates stream health, and the underlying amount of uncertainty (variability) in this measure is much higher than what is portrayed by its values observed from field samples. The latter remains

true even after bootstrap bias in metric mean and SD have been removed. Such a property poses reliability questions about the SD values reported in, say, Fore *et al.*, 1994 and Karr and Chu, 1999. As the index varies widely over repeated (bootstrap) sampling from the same site in the same year, a likely scenario is for the stream's underlying health to have greatly degraded in one year, and yet the large uncertainty of the index produces very similar values for consecutive years. Altogether, these undesirable properties make it questionable to rely heavily on the B-IBI as a "report-card" measure when monitoring stream health over time.

We propose SHIPSL as an alternative. It employs standardization as a metric scoring scheme. Standardization is a great improvement to the use of the discrete 1-3-5 scale or the continuous [0,1] or [0,10] scales, because it maps metric values to an unbounded real line centered at 0, immediately providing an intuitive interpretation of the metric z-scores or SHIPSL values. While no standard protocol is available for the selection of IBI reference sites (USDA-NRCS, 2003), SHIPSL reference sites and recalibration of its scoring mechanism are unnecessary regardless of time (when assessing the same set of sites) and space of the study. When a single site is to be longitudinally gauged against some constant "reference," we propose the use of GS-SHIPSL which employs non-subjective gold standards in its metric scoring scheme. Besides the selection of appropriate metrics, virtually no human input is required in the development of a (GS-)SHIPSL-like index for any geographical region.

We provide the following guideline for developing a localized version of (GS-)SHIPSL for streams outside of the PSL:

1. Arrive at a final set of metrics that are deemed efficient indicators of ecological health.

2. Ensure that least- and most-impacted sites are included in the study (ordinary SHIPSL), or define census-based gold-standard mean and SD for each

metric (GS-SHIPSL).

3. Collect the same number of replicate samples from all streams being studied.

4. Determine raw metric values from field samples.

5. Identify metrics that are associated with stream size or other obvious external factors, then produce new metric values corrected for the association via some form of regression (e.g. generalized linear models). Denote the final set of metric values by $\mathbb{Y}^*$.

6. As outlined in Section 2, average the values in $\mathbb{Y}^*$ over replicate samples, convert them into metric z-scores using sample metric mean and SD (ordinary SHIPSL) or gold-standard counterparts (GS-SHIPL), then sum them to obtain health index values w's.

The advantages of either version of SHIPSL go far beyond ease of intuitive interpretation and localization. Our results indicate that SHIPSL is comparable to B-IBI in the information it carries about underlying biological conditions. Statistically, SHIPSL appears to be more reliable than B-IBI in measuring stream health, as it removes bias and is less prone to chance variation, thus more efficiently reflecting the true state of health among streams. Most importantly, SHIPSL achieves all of these without adding any technical requirement to conventional biomonitoring.

However, two issues remain unresolved. Firstly, we have seen that bootstrap index values can be distorted by bootstrap taxa richness which can never exceed observed taxa richness from field samples. While conclusions for GS-SHIPSL are drawn based on bias-corrected bootstrap distributions of taxa richness metrics, it is unclear how one may verify that these corrections are appropriate for actual field sampling practices. Concerns over bias due to the bootstrap are further addressed by Chiu and Guttorp (2004), who propose the use of percentage-valued

counterparts to replace conventional integer-valued taxa richness metrics. They find that percentage richness is much less prone to bias in sample mean and SD (Table 7), and is more highly correlated with urbanization. However, a health index that employs percentage richness is found to remain sensitive to the few low-abundance metrics, for which observed field counts of zero are treated as structural zeros in the bootstrap. In practice, repeated sampling of benthic organisms may be similarly affected. For instance, consider a taxon unobserved in a field sample replicate. Although future replicates are not restricted to yield zero counts as would bootstrap resamples, how may one determine, based on the observed zero, whether this taxon is indeed present at the sampled site? If it is not, then there is no variability or bias whatsoever in the frequency count. If it is but organisms are not abundant, it may take many replicates before a positive frequency is observed, and bias in a taxa richness metric involving this taxon is almost undoubtedly negative among a small to moderate number of replicates. Therefore, conditions gauged by such a metric are possibly irreproducible in a handful of field samples. It appears that the practicality of metrics involving low abundance taxa as biological indicators is questionable.

Secondly, current biomonitoring practices remain highly geographically dependent. For example, the ten SHIPSL metrics may not be informative if used in a health index for, say, eastern United States. Thus, any effort for enhancing the index's universality would be made in vain unless a common "language" for describing and quantifying health is available to different geographical regions.

ACKNOWLEDGMENTS

We thank the referee and the staff at the U.S. Environmental Protection Agency (EPA) Western Ecology Division (WED) for providing valuable comments and suggestions. We also thank Susannah Iltis (Columbia Basin Research, University of Washington) and Sarah Morley (Northwest Fisheries Science Center) for various benthic taxonomy datasets, and Professor James Karr (Aquatic and Fishery Sciences, University of Washington) for clarifications and reference material.

REFERENCES

Blocksum KA. 2003. A performance comparison of metric scoring methods for a multimetric index for mid-Atlantic highlands streams. *Environmental Management* **31**: 670–682.

Boulton AJ. 1999. An overview of river health assessment: philosophies, practice, problems and prognosis. *Freshwater Biology* **41**: 469–479.

Brinck KW. 2002. *Comparing Methods for Inferring Site Biological Condition from a Sample of Site Biota.* M.S. thesis: University of Washington.

Bunea F, Guttorp P, Richardson T. 1999. *Ecological Indices and Graphical Modeling of Factors Influencing Benthic Populations in Streams.* NRCSE Technical Report Series No. 036.

Canterbury GE, Martin TE, Petit DR, Petit LJ, Bradford DF. 2000. Bird communities and habitat as ecological indicators of forest condition in regional monitoring. *Conservation Biology* **14**: 544–558.

Chiu G, Guttorp P. 2004. *New Developments involving the Stream Health Index for the Puget Sound Lowland.* NRCSE Technical Report Series No. 079.

Fore LS, Karr JR, Conquest LL. 1994. Statistical properties of an index of biological integrity used to evaluate water resources. *Canadian Journal of Fisheries and Aquatic Sciences* **51**: 1077–1087.

Harris JH, Silveira R. 1999. Large-scale assessments of river health using an index of biotic integrity with low-diversity fish communities. *Freshwater Biology* **41**: 235–252.

Karr JR. 1981. Assessment of biotic integrity using fish communities. *Fisheries* **6**: 21–27.

Karr JR. 1998. Rivers as sentinels: Using the biology of rivers to guide landscape management. In *River Ecology and Management: Lessons from the Pacific Coastal Ecosystems,* Naiman RJ, Bilby RE (eds). Springer, New York; 502–528.

Karr JR. 1999. Defining and measuring river health. *Freshwater Biology* **41**: 221–234.

Karr JR, Chu EW. 1999. *Restoring Life in Running Waters.* Island Press, Washington.

Karr JR, Fausch KD, Angermeier PL, Yant PR, Schlosser IJ. 1986. *Assessing Biological Integrity in Running Waters – Method and Its Rationale.* Illinois Natural history Survey: Special Publication 5.

Kerans BL, Karr JR. 1994. A benthic index of biotic integrity (B-IBI) for rivers of the Tennessee Valley. *Ecological Applications* **4**: 768–785.

Lackey, RT. 2003. Appropriate use of ecosystem health and normative science in ecological policy. In *Managing for Healthy Ecosystems,* Rapport DJ, Lasley WL, Rolston DE, Nielsen NO, Qualset CO, Damania AB (eds). Lewis, Boca Raton; 175–186.

McCormick FH, Hughes RM, Kaufmann PR, Peck DV, Stoddard JL. 2001. Development of an index of biotic integrity for the Mid-Atlantic Highlands region. *Transactions of the American Fisheries Soceity* **130**: 857–877.

Morley, SA. 2000. *Effects of Urbanization on the Biological Integrity of Puget Sound Lowland Streams: Restoration with a Biological Focus.* M.S. thesis: University of Washington.

SalmonWeb. `http://www.salmonweb.org`.

USDA-NRCS. 2003. *Fish Assemblages as Indicators of the Biological Condition of Streams and Watersheds,* Ecological Sciences 190, Technical Note 16. Wetland Science Institute: United States Department of Agriculture, Natural Resources Conservation Service, Washington DC.

metric	*averaged / pooled*	*1994*			*1997*			*1997 "jittered"*		
		1	**3**	**5**	**1**	**3**	**5**	**1**	**3**	**5**
total # taxa (#Tx)	averaged	$[0, 10)$	$[10, 20)$	≥ 20	$[0, 14)$	$[14, 28)$	≥ 28	$[0, 13.53)$	$[13.53, 27.50)$	≥ 27.50
# *Ephemeroptera* taxa (#EphTx)	averaged	$[0, 3)$	$[3, 5.5)$	≥ 5.5	$[0, 3.5)$	$[3.5, 7)$	≥ 7	$[0, 3.26)$	$[3.26, 7.22)$	≥ 7.22
# *Plecoptera* taxa (#PleTx)	averaged	$[0, 3)$	$[3, 5.5)$	≥ 5.5	$[0, 2.7)$	$[2.7, 5.3)$	≥ 5.3	$[0, 2.46)$	$[2.46, 5.07)$	≥ 5.07
# *Tricoptera* taxa (#TriTx)	averaged	$[0, 2)$	$[2, 4.5)$	≥ 4.5	$[0, 2.7)$	$[2.7, 5.3)$	≥ 5.3	$[0, 2.87)$	$[2.87, 4.99)$	≥ 4.99
# long-lived taxa (#LLTx)	pooled	$[0, 0.5)$	$[0.5, 2)$	≥ 2	$[0, 4)$	$[4, 8)$	≥ 8	$[0, 3)$	$[3, 7)$	≥ 7
# intolerant taxa (#IntolTx)	pooled	$[0, 0.5)$	$[0.5, 2)$	≥ 2	$[0, 2)$	$[2, 4)$	≥ 4	$[0, 3)$		≥ 3
% tolerant individuals (%Tol)	averaged	>50	$(20, 50]$	$[0, 20]$	>44	$(27, 44]$	$[0, 27]$	>45.16	$(25.86, 45.16]$	$[0, 25.86]$
% predatory individuals (%Pred)	averaged	$[0, 5)$	$[5, 10)$	≥ 10	$[0, 4.5)$	$[4.5, 9)$	≥ 9	$[0, 4.61)$	$[4.61, 9.11)$	≥ 9.11
# clinger taxa (#ClingTx)	averaged	$[0, 8)$	$[8, 15)$	≥ 15	$[0, 8)$	$[8, 16)$	≥ 16	$[0, 7.79)$	$[7.79, 15.70)$	≥ 15.70
% individuals in 3 most dominant taxa (%Dom3)	averaged	>75	$(50, 75]$	$[0, 50]$	>75	$(55, 75]$	$[0, 55]$	>76.22	$(54.16, 76.22]$	$[0, 54.16]$

Table 1: B-IBI metric scoring criteria for the 1994 and 1997 PSL data, taken from SalmonWeb and Morley, 2000, respectively. The "jittered" criteria contain independently generated random noise added to the 1997 cutoff points.

site	*B-IBI*					*SHIPSL*				
	observed	rating	bootstrap			observed	rating	bootstrap		
			mean	bias	95% C.I.			mean	bias	95% C.I.
BB1	32	fair	30.52	-1.48	[28 , 32]	0.57	C	0.53	-0.04	[-0.96 , 1.86]
BB2	36	fair	34.74	-1.26	[32 , 36]	5.97	B	5.51	-0.46	[3.62 , 7.28]
BB3	32	fair	32.08	0.08	[30 , 34]	3.84	C	3.61	-0.23	[2.13 , 4.98]
BB4	34	fair	33.02	-0.98	[30 , 34]	2.93	C	3.13	0.20	[1.80 , 4.33]
BB5	28	fair	27.80	-0.20	[26 , 30]	3.70	C	3.83	0.13	[2.50 , 5.04]
BS1	26	poor	25.19	-0.81	[22 , 26]	2.00	C	1.72	-0.28	[0.14 , 3.16]
JE1	32	fair	32.02	0.02	[32 , 32]	6.19	B	5.55	-0.64	[3.60 , 7.29]
LB1	36	fair	33.79	-2.21	[32 , 36]	6.58	B	6.71	0.13	[4.75 , 8.41]
LB2	28	fair	27.96	-0.04	[28 , 28]	0.60	C	0.77	0.17	[-0.55 , 1.93]
LB3	22	poor	22.55	0.55	[22 , 24]	-3.72	D	-3.36	0.36	[-4.52 , -2.33]
LB4	16	v. poor	14.18	-1.82	[12 , 18]	-9.48	E	-9.35	0.13	[-10.54 , -8.26]
MA1	24	poor	23.93	-0.07	[22 , 24]	-2.45	D	-3.38	-0.93	[-5.22 , -1.65]
MI1	12	v. poor	11.61	-0.39	[10 , 14]	-16.80	F	-16.41	0.39	[-17.43 , -15.40]
RO1	48	excellent	44.24	-3.76	[40 , 48]	15.72	A	15.32	-0.40	[13.21 , 17.07]
SW1	28	fair	27.19	-0.81	[26 , 28]	1.71	C	2.49	0.78	[1.45 , 3.40]
SW2	26	poor	25.62	-0.38	[24 , 26]	-1.01	D	-0.98	0.03	[-2.18 , 0.13]
SW3	28	fair	27.75	-0.25	[26 , 28]	1.82	C	2.62	0.80	[1.42 , 3.64]
TH1	10	v. poor	10.93	0.93	[10 , 14]	-18.16	F	-18.29	-0.13	[-20.20 , -16.13]

Table 2: Observed B-IBI and SHIPSL values for 1997 PSL field samples, and their corresponding ratings and bootstrap means, biases, and 95% C.I.'s.

SHIPSL C.I.'s	*B-IBI C.I.'s*		
	0 excluded	0 included	Total
0 excluded	100 (65.4%)	25 (16.3%)	125 (81.7%)
0 included	12 (7.8 %)	16 (10.5%)	28 (18.3%)
Total	112 (73.2%)	41 (26.8%)	153 (100.0%)

Table 3: Coverage of 95% bootstrap C.I.'s for pairwise difference in mean between two sites for the 1997 PSL data. The null hypothesis of no difference is rejected at an individual 5% significance level when 0 is excluded from the C.I.

sites	*B-IBI*		*SHIPSL*	
	95% C.I. for mean diff	*rating*	*95% C.I. for mean diff*	*rating*
JE1−BB5	[2 , 8]	**fair**	[-2.37 , 5.61]	**B, C**
BB2−BB5	[2 , 10]	**fair**	[-2.51 , 5.77]	**B, C**
BB3−LB2	[2 , 8]	**fair**	[-0.51 , 6.46]	C
JE1−LB2	[2 , 6]	**fair**	[0.87 , 8.77]	B, C
BB2−LB2	[2 , 10]	**fair**	[0.71 , 8.62]	B, C
LB1−LB2	[2 , 10]	**fair**	[1.87 , 9.55]	B, C
BB3−SW1	[2 , 8]	**fair**	[-1.88 , 3.86]	C
JE1−SW1	[2 , 8]	**fair**	[-0.72 , 6.59]	**B, C**
BB2−SW1	[2 , 10]	**fair**	[-0.59 , 6.34]	**B, C**
LB1−SW1	[2 , 10]	**fair**	[0.25 , 7.67]	B, C
BB3−SW3	[2 , 8]	**fair**	[-2.20 , 4.11]	C
JE1−SW3	[2 , 8]	**fair**	[-1.01 , 6.39]	**B, C**
BB2−SW3	[2 , 12]	**fair**	[-0.93 , 6.39]	**B, C**

Table 4: PSL sites from 1997 that are underclassified by the B-IBI-based five-point rating at a simultaneous 5% significance level, and their corresponding SHIPSL-based grades. Bold-faced ratings disagree with the rejection or non-rejection of H_0 : *no mean difference* based on the corresponding family-wise 95% bootstrap C.I's.

sites	*B-IBI*		*SHIPSL*	
	95% C.I. for mean diff	*rating*	*95% C.I. for mean diff*	*rating*
SW1−MA1	[0 , 8]	**fair, poor**	[2.63 , 9.78]	C, D
SW3−MA1	[0 , 8]	**fair, poor**	[2.49 , 9.70]	C, D
BB5−MA1	[0 , 8]	**fair, poor**	[3.44 , 11.32]	C, D
SW1−BS1	[-2 , 6]	**fair, poor**	[-2.31 , 4.12]	C
SW3−BS1	[-2 , 6]	**fair, poor**	[-2.17 , 4.05]	C
BB5−BS1	[-2 , 8]	**fair, poor**	[-1.34 , 5.82]	C
LB2−BS1	[0 , 6]	**fair, poor**	[-4.19 , 2.74]	C
BB1−BS1	[0 , 10]	**fair, poor**	[-4.79 , 2.61]	C
SW1−SW2	[0 , 4]	**fair, poor**	[0.96 , 6.29]	C, D
SW3−SW2	[-1 , 4]	**fair, poor**	[0.92 , 6.41]	C, D
BB5−SW2	[0 , 6]	**fair, poor**	[1.67 , 7.85]	C, D
LB2−SW2	[0 , 6]	**fair, poor**	[-1.11 , 4.67]	**C, D**
BB1−SW2	[0 , 8]	**fair, poor**	[-1.75 , 4.54]	**C, D**
SW2−BS1	[-4 , 4]	poor	[-6.09 , 0.75]	**D, C**
JE1−BS1	[4 , 12]	fair, poor	[-0.05 , 8.22]	**B, C**
BB2−BS1	[4 , 14]	fair, poor	[-0.67 , 8.28]	**B, C**
JE1−SW1	[2 , 8]	**fair**	[-0.72 , 6.59]	**B, C**
BB2−SW1	[2 , 10]	**fair**	[-0.59 , 6.34]	**B, C**
JE1−SW3	[2 , 8]	**fair**	[-1.01 , 6.39]	**B, C**
BB2−SW3	[2 , 12]	**fair**	[-0.93 , 6.39]	**B, C**
JE1−BB5	[2 , 8]	**fair**	[-2.37 , 5.61]	**B, C**
LB1−BB5	[0 , 10]	fair	[-1.14 , 6.52]	**B, C**
BB2−BB5	[2 , 10]	**fair**	[-2.51 , 5.77]	**B, C**
BB3−JE1	[-4 , 4]	fair	[-5.70 , 2.44]	**B, C**
BB4−JE1	[-4 , 4]	fair	[-6.37 , 1.53]	**B, C**
LB1−BB3	[-4 , 6]	fair	[-0.93 , 6.99]	**B, C**
BB2−BB3	[-2 , 6]	fair	[-2.26 , 6.31]	**B, C**
LB1−BB4	[-4 , 6]	fair	[-0.33 , 7.50]	**B, C**
BB2−BB4	[-4 , 8]	fair	[-1.44 , 5.93]	**B, C**

Table 5: Overclassification of sites for the 1997 PSL data. Bold-faced ratings disagree with the rejection or non-rejection of H_0 : *no mean difference* based on the corresponding simultaneous 95% bootstrap C.I's.

site	*v. poor*	*poor*	*fair*	*good*	*excellent*	*absolute change*
TH1	$-$ <1	+ <1				<1
LB4	−43	+43				43
BB1		− <1	+ <1			<1
		+ <1	− <1			
MA1		− <1	+ <1			<1
LB2		−2	+2			2
BB5		−13	+13			13
SW3		−7	+7			7
		+ <1	− <1			
SW2		−32	+32			32
SW1		−39	+39			39
BS1		−52	+52			52
		+ <1	− <1			
RO1			− <1	+ <1		38
			+ <1	− <1		
				−16	+16	
				+22	−22	

Table 6: Approximate percentages of bootstrap samples differently classified due to jittered cutoffs in the metric scoring. A "+" indicates a gain in occurrence of the rating, whereas a "−" indicates a loss. No change of classification is observed for sites BB2, BB3, BB4, JE1, LB1, LB3, and MI1.

taxa richness metric		count richness			percentage richness		
		observed	*bootstrap bias*	*p*	*observed*	*bootstrap bias*	*p*
#Tx	mean	23.26	**−2.23**	0.00			
	SD	6.65	**−0.32**	0.01			
EphTx	mean	4.74	**−0.28**	0.00	19.93	**1.09**	0.00
	SD	1.76	**−0.12**	0.00	4.31	−0.06	0.42
PleTx	mean	3.89	**−0.37**	0.00	15.58	−0.06	0.42
	SD	1.66	**−0.12**	0.01	5.78	0.20	0.27
TriTx	mean	4.00	**−0.39**	0.00	16.88	−0.18	0.34
	SD	1.37	−0.06	0.19	2.83	0.51	0.19
LLTx	mean	3.02	**−0.37**	0.00	11.94	−0.41	0.08
	SD	1.47	**−0.15**	0.00	5.38	−0.14	0.26
IntolTx	mean	0.20	**−0.05**	0.02	0.71	−0.14	0.07
	SD	0.49	**−0.10**	0.04	1.61	−0.22	0.14
ClingTx	mean	12.26	**−1.19**	0.00	50.35	−0.31	0.27
	SD	4.46	**−0.32**	0.00	11.55	0.49	0.32

non-taxa richness metric		*observed*	*bootstrap bias*	*p*
%Pred	mean	5.34	0.00	0.50
	SD	3.76	0.00	0.33
%Tol	mean	74.49	0.00	0.50
	SD	14.24	0.00	0.46
%Dom3	mean	38.21	−0.01	0.06
	SD	11.36	0.00	0.20

Table 7: **Bias in mean and SD of SHIPSL metrics.** One-sided p-values are based on (1) normal approximations to the bootstrap distributions of metric mean and SD, and (2) unrounded observed and bootstrap values. Numbers in bold are significant biases at a 5% level.

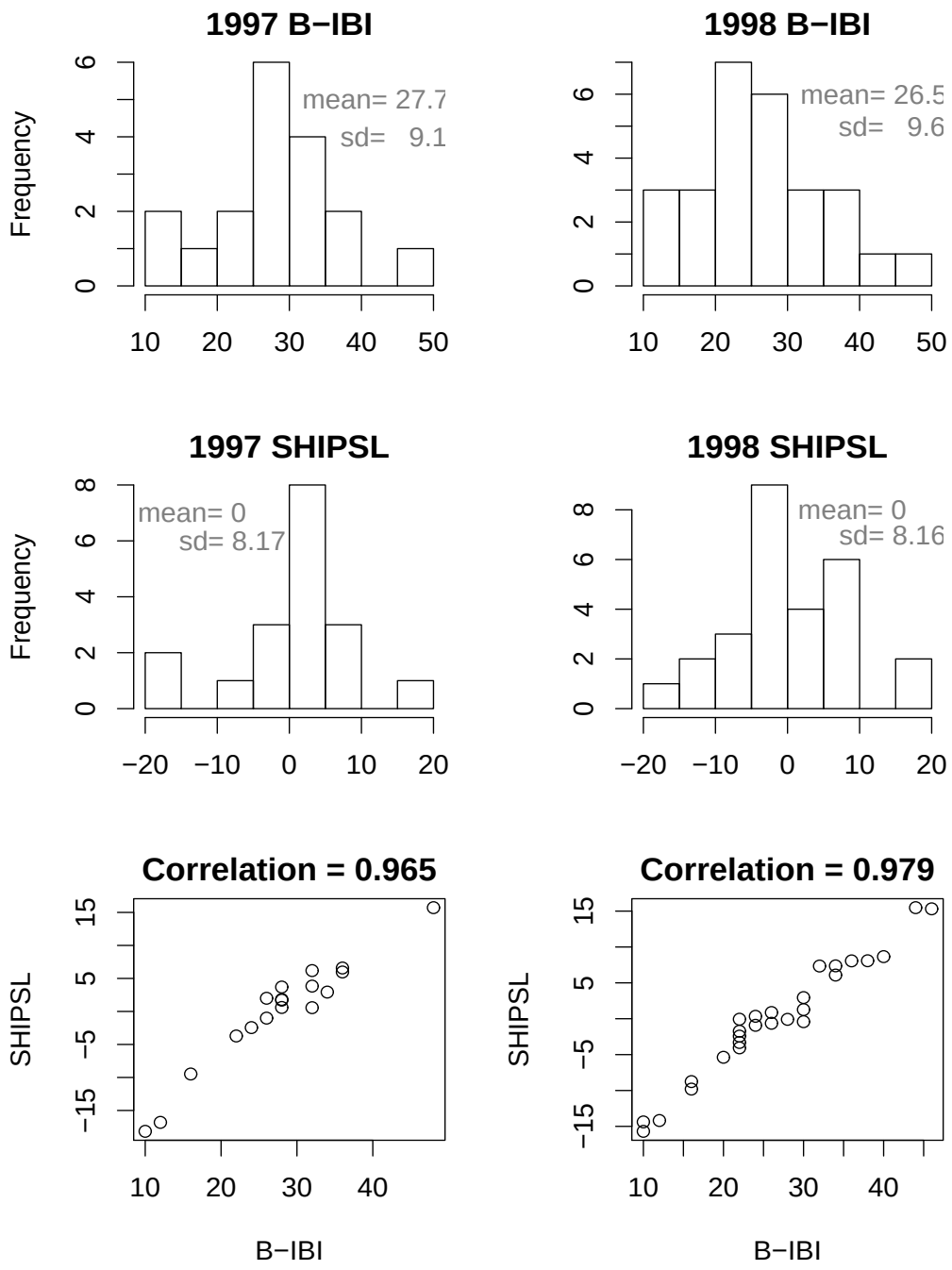

Figure 1: B-IBI and SHIPSL distributions for years 1997 and 1998, and the dependence between the two indices.

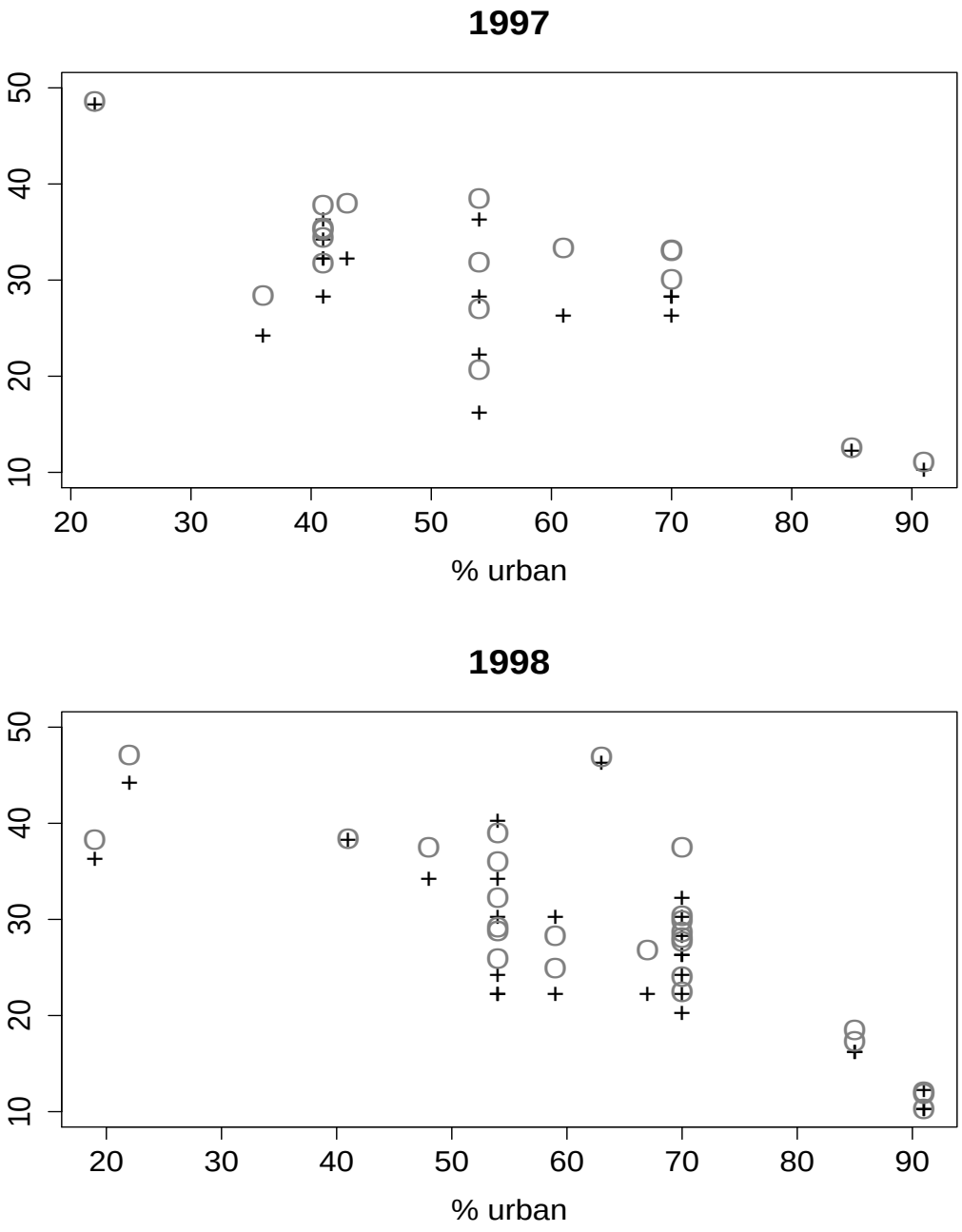

Figure 2: B-IBI (+) and SHIPSL (o), respectively, plotted against the percentage of urbanized land (a measure of impact due to human activities). Here, SHIPSL values are rescaled to have an approximate range of the B-IBI values observed.

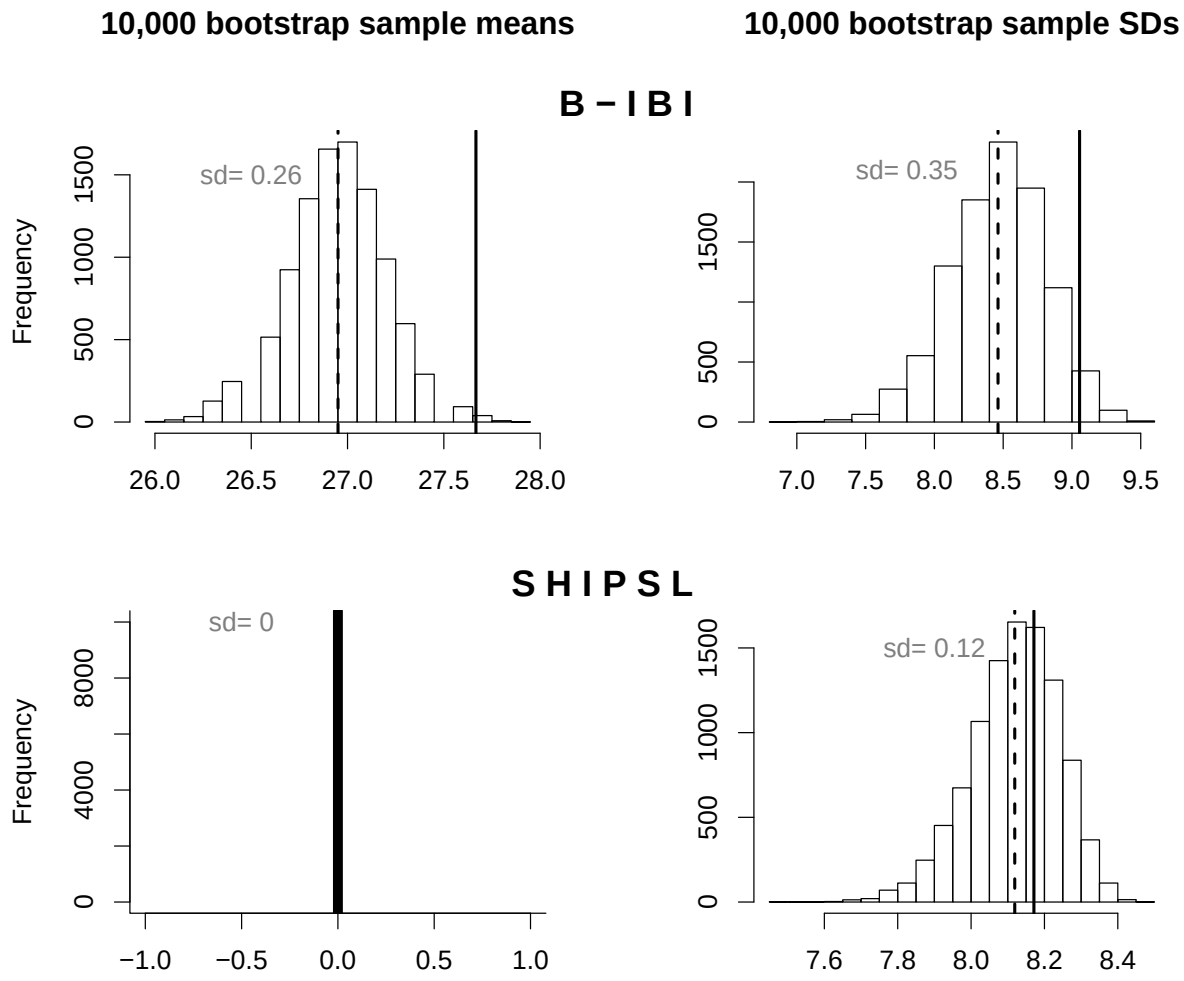

Figure 3: Distributions of sample statistics (mean and SD) for bootstrap field samples of the 18 PSL sites of 1997. The histogram's mean is marked by a dotted line, next to which is the histogram's SD (approximate standard error of the sample statistic). A solid line marks the corresponding statistic of the observed field sample (whose numerical value is given in Figure 1).

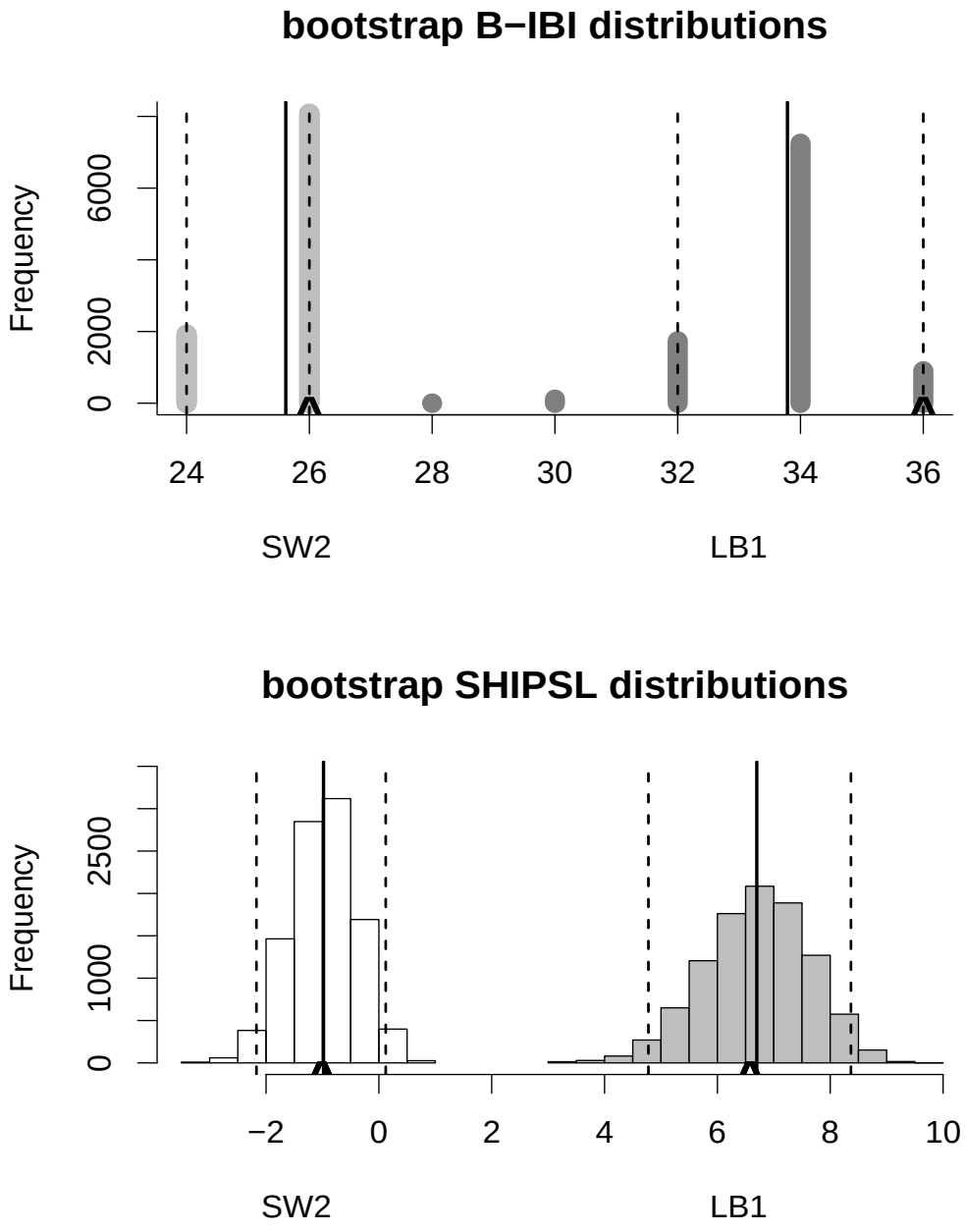

Figure 4: Examples of site-wise distributions of 10,000 bootstrap B-IBI and SHIPSL values. Observed index values are marked by a " ^ " along the x-axes. A solid line marks the mean of the bootstrap distribution. Dotted lines delimit a 95% bootstrap C.I. for the underlying true index value.

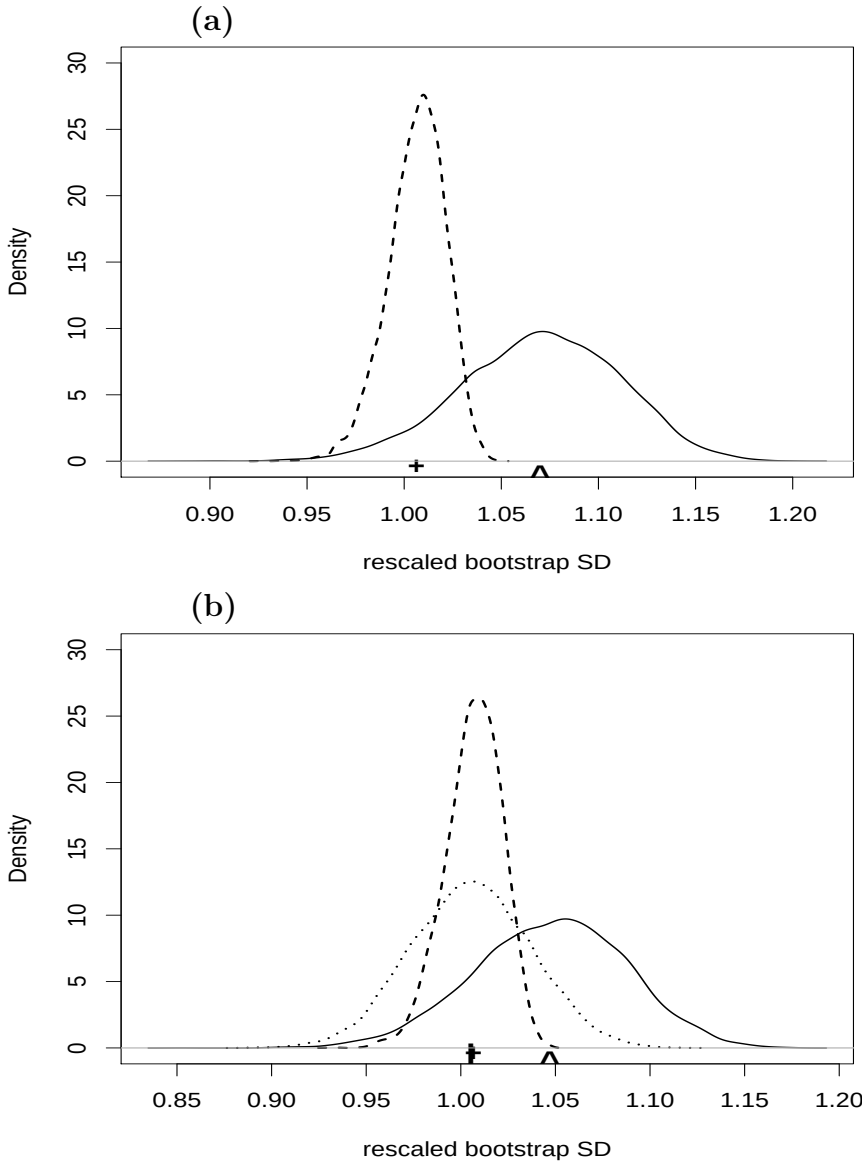

Figure 5: Bootstrap distributions for sample SD of B-IBI (—), SHIPSL (- - -), and GS-SHIPSL (. . .), before ((**a**)) and after ((**b**)) biases in count-valued taxa richness metrics have been corrected for. The means of the respective distributions are marked by " ^ ", "+", and "|". Both panels show resulting index values that have been rescaled to allow sensible comparison between the indices' precision.

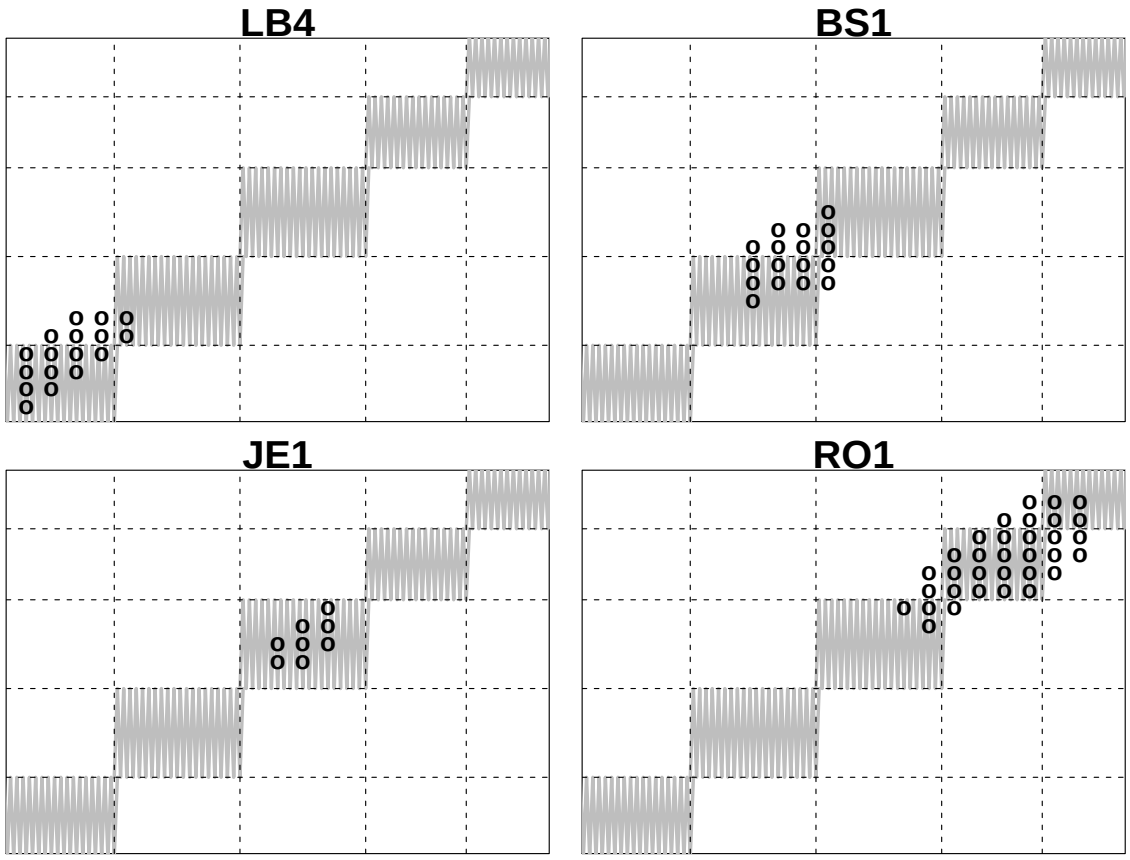

Figure 6: Examples of scatterplots for bootstrap B-IBI values *before* (x-axis) and *after* (y-axis) jittering the metric scoring criteria. Dotted lines delimit the five classes of stream health. Points within the shaded blocks are from instances where the health rating is unchanged.

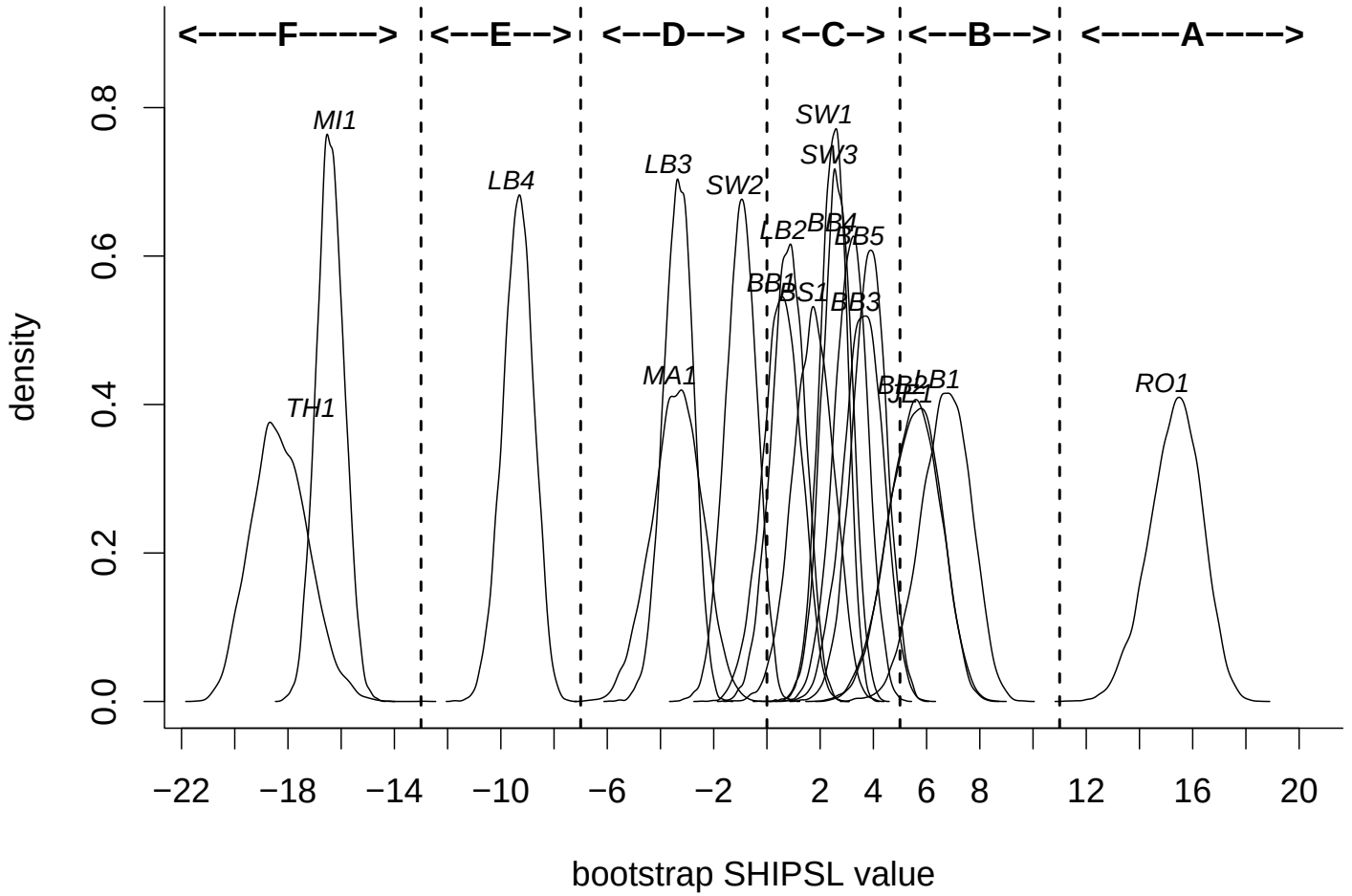

Figure 7: Bootstrap SHIPSL distributions (smoothed) and our six-point grading scale for stream health. Note that the BB2 and JE1 distributions almost entirely coincide.

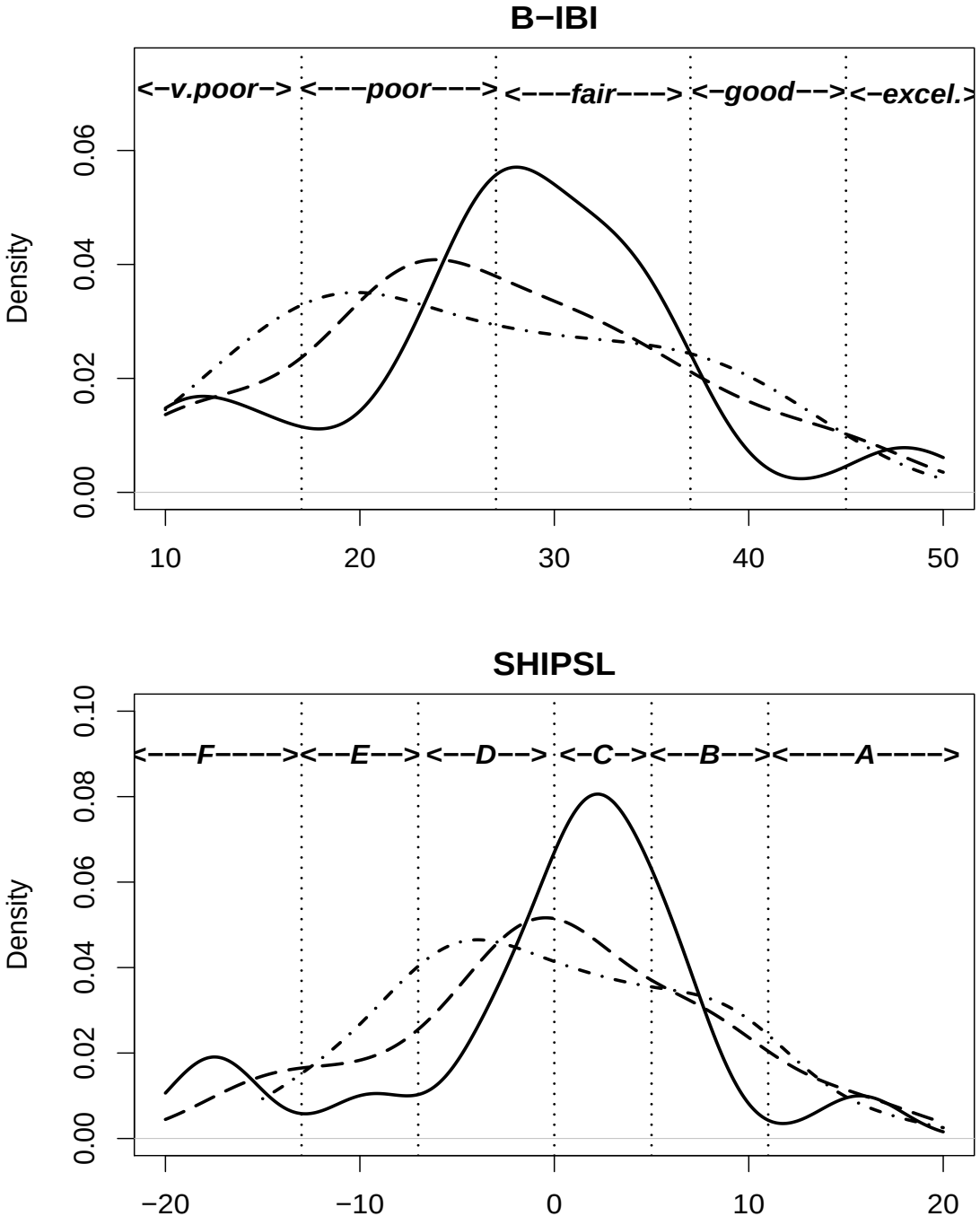

Figure 8: B-IBI and SHIPSL distributions for 1994 (_ . _ . _), 1997 (——), and 1998 (- - -), and the corresponding stream health rating schemes.

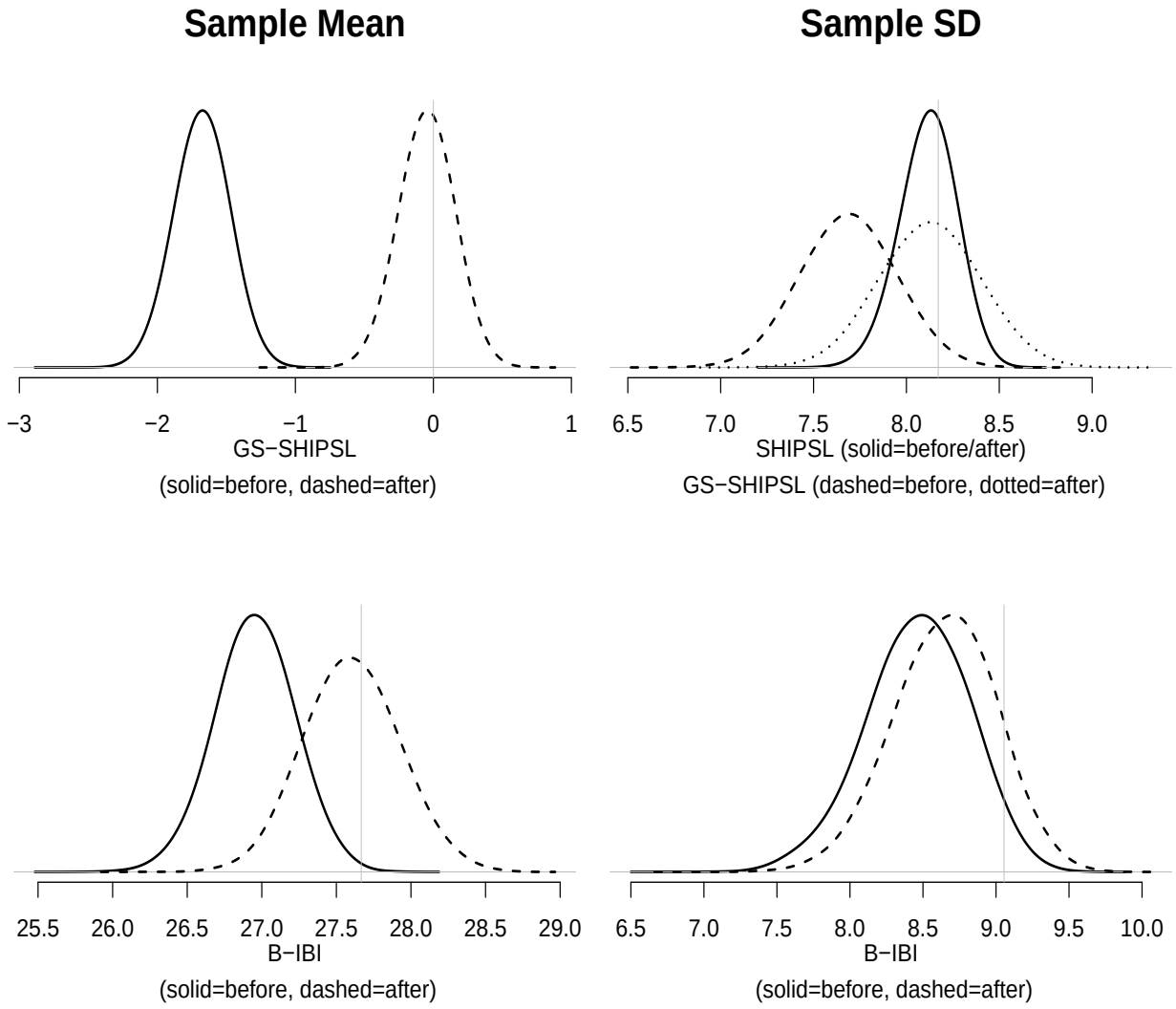

Figure 9: **Effect of bias correction for count-valued taxa richness metrics on bootstrap sample mean and SD.** Displayed are bootstrap distributions of sampled mean (left) and sampled SD (right) before and after bias correction. Vertical lines in gray denote values observed from 1997 field samples. Sample mean for ordinary SHIPSL is 0 by definition, and is excluded from this figure. Note that bias correction has virtually no effect on the sample SD for ordinary SHIPSL.