5.2 Metric selection

Metrics were evaluated for the following:

- Sensitivity
  - How well does the metric distinguish between reference and stressed sites?
  - What is the relationship between the metric and the disturbance variables?
    - Direction of response
    - Strength/significance
- Redundancy
- Representation across metric categories (richness, composition, evenness, tolerance, functional attribute, habit, thermal preference, and life cycle)
- Precision

The discrimination efficiency (DE) and *Z*-score were the primary performance statistics used to determine metric sensitivity. DE was calculated as the percentage of metric scores in stressed sites that were worse than the worst quartile of those in the reference sites. For metrics with a pattern of decreasing value with increasing environmental stress, DE is the percentage of stressed values below the 25th percentile of reference site values. For metrics that increase with increasing stress, DE is the percentage of stressed sites that have values higher than the 75th percentile of reference values. DE can be visualized on box plots of reference and stressed metric or index values with the inter-quartile range plotted as the box (Figure 10). Higher DE denotes a more frequent correct association of metric values with site conditions. DE values ≤25% show no discriminatory ability in one direction. Metrics with DE values ≥50% were generally considered for inclusion in the index. However, metric selection was usually dependent on relative DE values within a metric category.

The *Z*-score was calculated as the difference between mean reference and stressed metric or index values divided by the standard deviation of reference values. The *Z*-score is similar to Cohen's D (Cohen 1992) and gives a combined measure of index sensitivity and precision. There is no absolute *Z*-score value that indicates adequate metric performance, but among metrics or indices, higher *Z*-scores suggest better separation of reference and stressed values. Cohen proposed that *Z* values ≥ 0.80 indicated a "large" effect.

The DE and *Z*-scores summarize the difference in distributions at critical potential threshold levels and incorporate the precision of the reference distribution. They were used in favor of a t-test or signal to noise (S:N) ratio. The DE is an estimate of the percentage of correct impaired assessments and can be interpreted for management applications. While the *t*-test has been used elsewhere (Stoddard et al. 2008), we are not testing a hypothesis about the difference between reference and stressed sites. The *Z*-score and S:N ratio are similar measures of responsiveness as a function of variability.

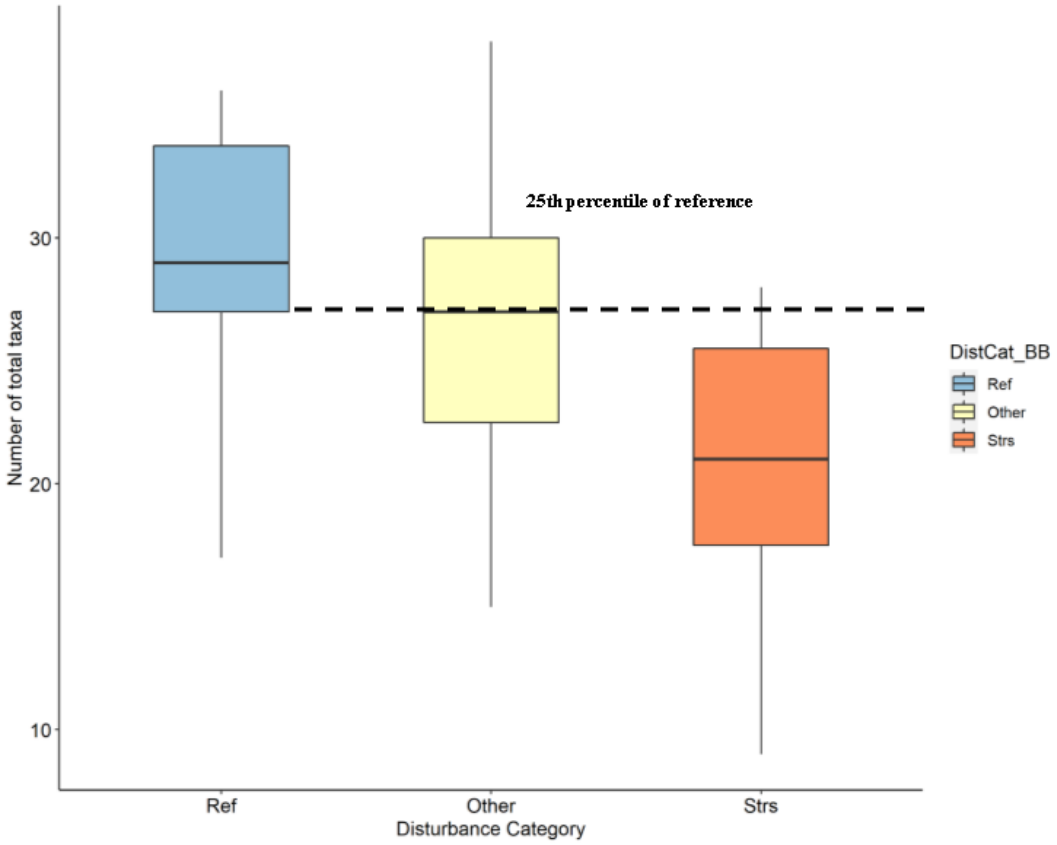

*Figure 10. Discrimination efficiency (DE). In this example, which uses the total number of taxa (a metric that decreases with stress), the 25th percentile of the reference distribution is used as the standard (and we calculate what percent of stressed sites were below that threshold; for example, if 15 out of 20 stressed sites have # total taxa metric values below the threshold (in this case, 27), the DE would equal 75%; if metric values for all 20 of the stressed sites were < 27, the DE would equal 100%). If it were a metric that increased with stress, we would have used the 75th percentile of the reference distribution as the standard (and calculated what percent of stressed sites were above that threshold). The formula is: DE = a/b*100, where a = number of a priori stressed sites identified as being below the degradation threshold (in this example, 25th percentile of the reference site distribution) and b = total number of stressed sites. The higher the DE, the better (the more frequent the correct association of metric values with site conditions).*

Table 8 contains a list of the metrics that had the best performance (with high DE and *Z*-scores) within each metric category and were selected to be tested in the index compilations. The list of candidate metrics was further culled by identifying redundant metrics (metrics that represent similar taxa or traits) and removing the poorer performing metrics. Finally, the remaining metrics and those being considered in the SNEP IBI project were favored since having the same IBI for both projects would simplify application across the region. In the MA/SNEP dataset, the best performing metrics had DE of 100%. Each metric category was represented by at least one metric with DE > 50%. Spearman correlation analyses were performed on all pairwise combinations of candidate metrics (Table 9). Metric pairs with Spearman $|r| \geq 0.85$ were considered redundant and were not both used in any index alternative. Metrics correlated at Spearman $|r| \geq 0.75$ were evaluated for possible exclusion.