12/28/2012

Evaluating Anomalous Results from the Bay BIBI Assessment (2012)
Key Points of Discussion

Participants: Roberto Llanso (Versar), Tish Robertson (VADEQ), and Matt Stover (MDE) talked on 12/12/12.

Summary of Problem: VA and MD noticed in the final BIBI assessment worksheet that several Bay segments were classified as being "not impaired" and yet had very low mean IBI scores and a high percentage of sites with BIBI scores that were below 3.0[1], the generally accepted threshold for biological impairment. In one particular case, the Corrotoman River assessment had a mean BIBI of 1.97, with 9 out of 17 sites being equal to or less than 2.0. VA, MD, and Versar held a conference call to determine what might have caused these counterintuitive results and whether they might have occurred in error.

- Partial Explanation**:** Several errors were found in the dataset used for the 2012 Integrated Report assessment. Specifically, Roberto, Michael Lane, and Bud Rodi found that there were segment designation discrepancies, incorrect BIBIs and incorrect salinities used. Of these, the incorrect salinities and resulting BIBIs were the most serious error types. It was determined that a transcription error caused these inaccuracies. This transcription error has since been resolved and all data entered into CIMS should now be accurate. (The only remaining exception, according to Bud Rodi, are some of the Elizabeth River Monitoring Project stations (with 'Z' in their name).) Since both Maryland and Virginia have submitted their respective 2012 Integrated Reports to EPA for approval, the affected assessment results will remain in the report until the next (2014) round of Integrated Reports assessments. (Though these errors likely caused some inaccuracies in the final statistics, they were not the sole reason for some of the counterintuitive results.)

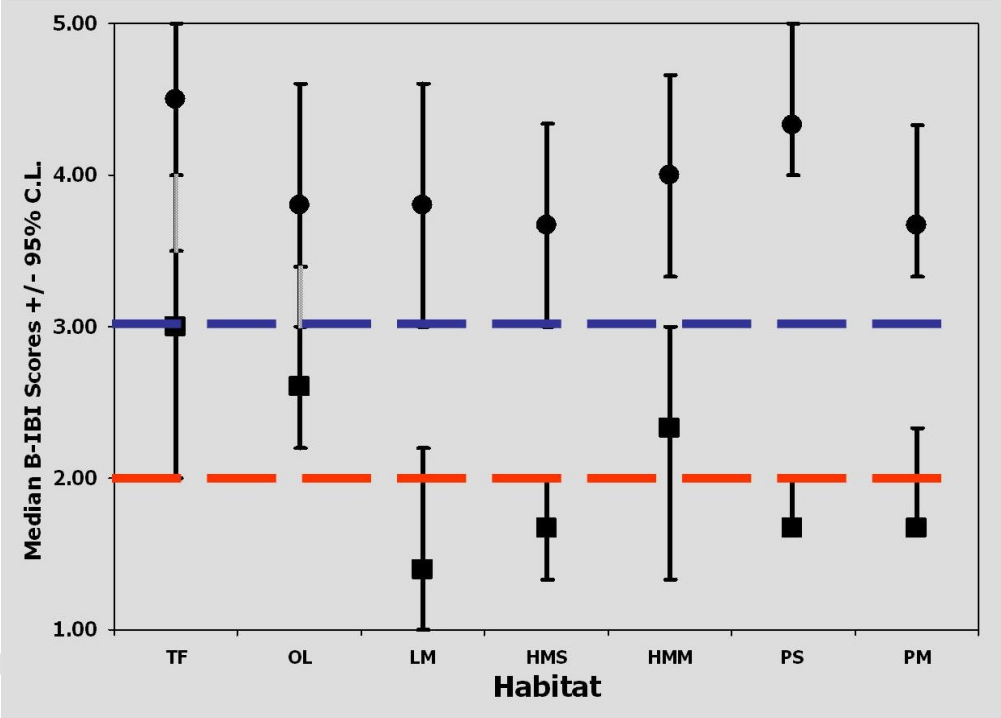
- Review of Assessment Methodology and Explanation: Roberto provided a synopsis of how the Chesapeake Bay Benthic Assessment Methodology works emphasizing the following tenets:

    - The methodology uses the null hypothesis that a sampled Chesapeake Bay segment is not impaired.
    - The methodology was created to limit Type I error, i.e., classifying a non-impaired segment as 'impaired'. This was done so as to reduce the likelihood that states would develop TMDLs and use precious water quality improvement resources on a segment that was not truly impaired. As a result, the methodology errs on the side of classifying borderline impaired segments as unimpaired, especially in those scenarios with a low sample size and high degree of variability.

[1] (In some cases like the Corrotoman, it had a high percentage of sites below 2.0.)

- Bootstrapping has the largest impact on those reference thresholds for tidal fresh and oligohaline waters, where the range of IBI scores for the reference degraded sites overlaps the range of IBI scores for the reference non-degraded sites. (See slide below from Roberto's presentation.) For these habitat types, since variability is high with such limited reference datasets, bootstrapping is used to improve confidence in the impairment thresholds used for classifying these sites. It should be noted that when the BIBI was originally developed, the reference dataset was fairly small and therefore, the thresholds for impairment could be better defined by incorporating newer, larger reference data.

Table 1: Range of IBI scores for both the reference degraded and reference non-degraded sites in each habitat type (Llanso 2007).



- A BIBI of 3.0 is not the standard impairment threshold for all benthic sites sampled within the Bay. During the CAP call, we incorrectly stated that the Llanso et al. (2003) and Alden et al. (2003) papers provide the impairment and non-impairment thresholds for each of the seven habitat types. Since then, however, we discovered that the thresholds used in the impaired waters assessment vary from run to run based on a random resample of the reference data for the good sites. Then, the smaller of either the 5th percentile of the good sites or the maximum value of the poor sites is used to determine the status of a sample. Where the overlap is large (freshwater, oligohaline, low mesohaline, and high mesohaline mud), the 5th percentile of the good sites is likely the threshold in each run. In the other habitats, the max value for the poor sites is likely the threshold.

- In some cases the thresholds for classifying a site as degraded or as non-degraded are quite far apart. This leads to a wide range of BIBI scores in the Intermediate range. When a site has a score in this intermediate range, it does not get classified as degraded and therefore does not get counted in the percent area degraded statistic.

- Roberto noted that the reference dataset used for BIBI development was a smaller data set than that available to Alden et al (2002). As a result, the range of the B-IBI values for the freshwater and oligohaline habitats is very wide (e.g., 1-5). So, in particular, the impaired waters assessment would benefit most from more reference benthic sampling in these habitats. However, even other habitat types lacked very robust reference datasets which likely contributed to the improper classification of the Corrotoman and other segments.

- Roberto also noted that segments such as the Corrotoman River (CRRMHa) that had counterintuitive results, i.e., low mean BIBIs and yet were classified as not impaired, for the most part, had low sample sizes. Many of these segments also had a fairly high proportion of sites sampled in habitats with smaller reference datasets. For this reason, many of samples in these segments could not be assessed as degraded with the level of certainty that the method requires and therefore did not result in an impairment listing. The group concluded that the low sample sizes coupled with the uncertainty associated with the impairment thresholds (due to the small reference dataset) is what likely led to these unusual results.

- Potential Solutions: The group discussed several potential solutions to this issue including creating some add-on rules to the assessment methodology to deal with scenarios such as the Corrotoman. The ideas presented here are not exhaustive.
  - Direct future sampling efforts to under-sampled segments and segments with high levels of variability. For these segments, the Bay partners would direct random sampling within individual segments maintaining the ability to characterize the larger bay segments if necessary. For instance, we could potentially forgo sampling in segments such as POTMH (MD) for the next few years. POTMH has been heavily sampled in the past and has a high degree of impairment which is not likely to change in the near future. Thus there would be little lost if we did not reassess this segment in the coming years.
  - Tish proposed the following rule using the coefficient of variation: **IF IMPAIRED = NO AND MEAN_BIBI < 2.95 AND CofVar > 25%**, then there is enough uncertainty to warrant an "insufficient info/Category 3" determination.
  - **IF IMPAIRED = NO AND MEAN_BIBI≤2.43 And Upper Confidence Limit – Lower Confidence Limit≥0.5**, then there is enough uncertainty

to warrant a Category 3 (insufficient information) determination. See footnote[2]

- **IF IMPAIRED = NO AND THE PERCENTAGE OF SITES≤2.0** (or some other relevant threshold(s)) **IS > X, THEN ASSESSMENT= INSUFFICIENT/IMPAIRED.** In this rule, if the percentage of sites below the degraded threshold(s) exceeds the management goal then determination of not impaired can be overridden with a determination of insufficient information or even impaired. (This is currently the closest approximation to what Maryland uses for its non-tidal 8-digit watersheds.)

Note: These add on rules have not yet been tested.

- Future Considerations: Roberto mentioned that a proposal had been submitted to the Chesapeake Bay Program to recalibrate the BIBI using a more robust reference dataset with a focus on the habitat types with high levels of variability (tidal freshwaters and oligohaline). Both Tish and Matt are supportive of this idea recognizing the need to update this assessment tool in light of lessons learned. In addition, to improve states' assessment coverage in under-sampled segments, MD and VA support the idea of apportioning more sites within segments that are currently in Category 3 of the Integrated Report or which have high degrees of variability.

[2] The threshold 2.42 is merely presented as an example. It was derived by taking the arithmetic mean of the impairment thresholds found in Table 2 from Llanso et al. 2003. A more relevant threshold(s) could be developed based on the habitats found within a segment