UNIVERSITY OF AMSTERDAM

Fifty Shades of Grey. Variability in Metric-based Assessment of Surface Waters using Macroinvertebrates

H.E. Vlek

Summary

Since the beginning of the 20th century, a wide variety of methods have been developed for the biological assessment of surface waters. Macroinvertebrates are a commonly applied taxonomic group for assessing water quality. Since the introduction of the European Water Framework Directive (WFD) in 2000, every member state is obligated to assess the effects of human activities on the ecological quality of all water bodies and indicate the level of confidence and precision of the results provided by the monitoring programs in their river basin management plans (European Commission, 2000). Currently, the statistical properties associated with aquatic monitoring programs are often unknown. Therefore, the overall objective of this thesis is to quantify the variability and accuracy associated with biological assessment based on macroinvertebrates in order to guide (1) the process of metric selection in the development of biological assessment systems and (2) the process of standardizing sampling and sample processing.

At the time the WFD was published, the biological assessment system(s) applied in the Netherlands did not meet the criteria for biological assessment systems set by the WFD. **Chapter 2** describes the development of a macroinvertebrate-based WFD compliant biological assessment system for fast and slow running streams in the Netherlands. A large dataset of 949 samples collected by water authorities from different regions in the Netherlands was used to construct a multimetric index. All sites received an ecological quality (post-) classification ranging from 1 (bad status) to 4 (good status) based on biotic and abiotic variables using a combination of multivariate analysis and expert judgment. More than 100 hundred metrics were tested for both stream types to examine their power to discriminate between streams of different ecological quality. Finally, 10 metrics were selected for the assessment of slow running streams and 11 metrics for the assessment of fast running streams. The individual metrics were combined into a multimetric index. Calibration showed that 67% of the samples from slow running streams and 65% of the samples from fast running streams were classified in agreement with their post-classification. In total, only 8% of the samples differed more than one quality class from the post-classification. The multimetric index was validated with ‘new’ data collected from 82 sites. Validation showed that 54% of the streams were classified correctly.

In order to standardize the biological assessment of surface waters in Europe, a standardized method for sampling, sorting, and identifying benthic macroinvertebrates in running waters was developed during the AQEM project. The AQEM method has proved to be relatively time-consuming. **Chapter 3** explores the consequences of reducing sample size on the variability, accuracy, and costs of bioassessment results. Macroinvertebrate samples were collected from six different streams: four streams located in the Netherlands and two in Slovakia. Twenty sampling units were collected from one or two dominant habitats in each stream using a pond net (25 x 25 cm) over a length of approximately 25 cm per sampling unit. The effect of increasing sample size on variability and accuracy was examined for six metrics and the multimetric index developed in Chapter 2 for the assessment of Dutch slow running streams. The accuracy of metric results increased and variability decreased with increasing sample size. In addition, accuracy and variability varied depending on the habitat and metric. The AQEM sampling method prescribes a multihabitat sample of 5 m. The results suggest that a sample size of less than 5 m is adequate to attain a coefficient of variation (CV) and mean relative deviation (MRD) of 10% or less for the metrics Average Score Per Taxon (ASPT), Saprobic Index, and the percentage of individuals with a preference for the akal, littoral, and psammal (type Aka+Lit+Psa (%)). The metrics number of taxa, number of individuals, and EPT-taxa (%) required a multihabitat sample size of more than 5 m to attain a CV and MRD of ≤ 10%. For the metrics number of individuals and number of taxa, a multihabitat sample size of 5 m is not adequate to attain a CV and MRD of ≤ 20%. The accuracy of the multimetric index for Dutch slow running streams can be increased from ≤ 20% to ≤ 10% by increasing labor time by 2 hours. Considering this

low increase in cost and the possible implications of incorrectly assessing the results, striving for this ≤ 10% accuracy is recommended. To achieve an accuracy of ≤ 10%, a multihabitat sample of the four habitats studied in the Netherlands requires a sample size of 2.5 m and a labor time of 26 hours (excluding identification of Oligochaeta and Diptera) or 38 hours (including identification of Oligochaeta and Diptera).

To reduce the costs of surveillance monitoring, water managers often collect only one sample a year. A lack of standardization of the sampling period (season) introduces a source of variation in bioassessment results. In **Chapter 4**, the monthly variation in the composition of the macroinvertebrate community is examined, including the effect this has on variations in metric values. For this purpose, two replicate samples were collected every other month for one year from a fourth order calcareous stream in the western Carpathian Mountains of central Europe, the Stupavský potok brook. Any single replicate contained, on average, 42% of the total number of taxa collected during this study. Multivariate analysis of the macroinvertebrate communities clearly separated the samples into three groups: (1) April samples, (2) June and August samples, and (3) October, December, and February samples. Thirty-one of 76 metrics showed significant ($p<0.05$, $\alpha=0.05$) differences between months. The majority of metrics exhibiting significant differences between months were quantitative metrics (i.e., metrics based on the relative abundance of a particular taxonomic group). The CV of most qualitative metrics did not exceed 20%. However, the highest CV values (above 40%) were found in most cases for the quantitative metrics. Thus, when using quantitative metrics, it is important to recognize that the season in which samples are collected can, and often will, have a strong influence on the results. In terms of individual metrics, differences between months strongly depend on the metric being evaluated. This makes it difficult to recommend a preferred sampling month or season. For metrics with high seasonal variation, the best solution is to always sample during the same month or to take into account seasonal variation when setting class boundaries for assessment purposes.

Another aspect of sampling and sample processing, which may influence bioassessment results in terms of variability, accuracy, and cost, is the choice of whether or not to use a preservative before sorting macroinvertebrate samples (i.e., dead specimens vs. living specimens). In **Chapter 5**, preserved and unpreserved samples collected from three lowland streams in the Netherlands were compared using identical sample processing protocols. Significantly different numbers of Ephemeroptera individuals and Hydracarina taxa and individuals were collected from preserved samples compared to unpreserved samples. In assessments based on these individual metrics, sample processing will need to be standardized. In streams with Ephemeroptera, the preservation of samples is necessary to optimize the number of Ephemeroptera individuals collected. In streams that contain Hydracarina, the preservation of samples will result in an underestimation of the number of Hydracarina taxa and individuals. A difference in ecological quality between preserved and unpreserved samples was observed in only one case, indicating that assessing small Dutch lowland streams does not require standardization of sample preservation in the sample processing protocol. We did not detect significant differences in sample processing costs between preserved and unpreserved samples.

Since the introduction of the Habitat Directive and the WFD, water authorities are obliged to monitor changes in conservation value/ecological quality on larger spatial scales (as opposed to site scale) and, indicate the level of confidence and precision of the results provided by the monitoring programs in their river basin management plans (European Commission, 2000). To increase insight into the statistical properties associated with aquatic monitoring programs, the spatial and temporal variability of taxonomic richness metrics were quantified in **Chapter 6**. We collected macroinvertebrate samples from 25 meso-eutrophic drainage ditches located in the Wieden natural preserve in the Netherlands and selected seven taxonomic richness metrics for the evaluation of spatial and temporal variability. The results from this study clearly indicated that, in general, it is easier to detect changes in a drainage ditch network based on metrics than on individual species. The required monitoring effort for rare species automatically implies that data collected by water authorities in biomonitoring programs developed to meet the requirements of the WFD will not meet the requirements of conservation managers. When interested in an individual species, sampling methods will have to be adjusted to the specific species in order to increase the frequency of collection. Irrespective of the metric applied, a large effort will be required to detect changes within the drainage ditches of the Wieden due to high spatial variability. Therefore, we need to explore the possibilities of applying alternative, more cost-effective methods for sampling and sample processing in biomonitoring programs.

This thesis shows that the variability in metric values applied in biological assessment is often high. Also, the variability in metric values varies between stream types, season (sampling period), and the sampling and sample processing method, making it difficult to give 'universal' advice on metrics to be included in biological assessment systems and optimal choices regarding the standardization of sampling and sample processing. However, high variability is not solely an issue of biology. Although the variation in biological data can be high, the temporal and spatial variation in physical and chemical variables can also be high (Veeningen, 1982). We should face the issue of high variability by gaining a better understanding of ecosystem functioning and unraveling cause-effect mechanisms, as well as by developing more cost-effective sampling and sample processing methods. A short-term solution to reduce variability and improve the performance of currently applied assessment systems in the Netherlands would be the implementation of quality assurance and quality control procedures, which have been successful in the United Kingdom. Apart from training personnel in sampling and sorting and performing audits of identification and sorting, additional standardization of the sampling and sample processing protocol is required, especially in terms of sorting effort. In the long run, water managers need to consider applying probability sampling to draw statistically sound conclusions at water body/national level. Combining probability sampling with a relatively cheap sampling and sample processing method to assess ecological status ('Quick Scan' method) will result in more cost-effective monitoring programs.