Fifty shades of grey: Variability in metric-based assessment of surface waters using macroinvertebrates
Keizer-Vlek, H.E.

Citation for published version (APA):
Chapter 1

1 General introduction

The Rode Beek part of natural preserve the Meirweg. *Photo: Piet Verdonschot.*
1 General introduction

History of biological assessment based on macroinvertebrates

Since the beginning of the 20th century, a wide variety of methods have been developed for biological assessment of surface waters. Macroinvertebrates are a commonly applied group of organisms for assessing water quality (e.g., Hawkes, 1979; Hellawell, 1986; Bailey et al., 2001; Hering et al., 2006). Many authors have stressed the advantages of using macroinvertebrates compared to other groups for biological monitoring and assessment purposes (e.g., Hellawell, 1986; Metcalfe, 1989). First, their intermediate life span makes it possible for them to exhibit a relatively quick response to stress (compared to macrophytes) while simultaneously reflecting ‘past’ environmental conditions (compared to algae). Second, their relatively sedentary lifestyle makes them representative of local conditions. Third, because of the heterogeneity of the macroinvertebrate community, the community will likely respond to a wide range of stressors. As such, macroinvertebrates are able to exhibit an integrative response to a combination of stressors.

With their Saprobien system, Kolkwitz & Marsson (1909) were the first in Europe to introduce the concept of organisms as indicators of environmental conditions. The Saprobien system was developed to detect organic pollution. Since its introduction the Saprobien system has been extended and revised by numerous European ecologists (Liebmann, 1951; Sládeček, 1965). In Germany, the Netherlands, and the Czech Republic the focus was mainly on improving the Saprobien system, but in countries such as Belgium, France, and the United Kingdom, ‘score systems’ focusing on the detection of general degradation were developed. Score systems such as the Trent Biotic Index (Woodiwiss, 1964) and the Indice Biotique (Tuffery & Verneaux, 1968) were developed in the 1960s following the introduction of the first diversity indices in the 1940s. Later, multivariate approaches, such as RIVPACS (Wright et al., 1984) in the UK and EKOO (Verdonschot, 1990) in the Netherlands, were introduced.

Developments comparable to those in Europe took place in the United States. In the 1980s, a multimetric index for fish was introduced in the United States (Karr, 1981). This was an approach to assessment not generally known in Europe. A multimetric index consists of a combination of several metrics, each providing different ecological information about the observed community and acting as an overall indicator of the biological integrity of a water resource. The
strength of the multimetric index is its ability to integrate information from individual, population, community, and ecosystem levels (Karr & Chu, 1999). A multimetric index provides detection capability over a broad range of stressors, creating a more complete picture of the ecosystem than single biological indicators (Intergovernmental Task Force on Monitoring Water Quality, 1993). Throughout this thesis the word metric is used to refer to any measure that can be calculated based on a sample from the macroinvertebrate community (e.g., the percentage of rheophilic species, Average Score Per Taxon, and German Saprobic Index) and a multimetric index is defined as the combination of two or more metrics to obtain a final assessment.

Rosenberg & Resh (1993) listed seven different approaches for assessing streams by using macroinvertebrates: richness measures, enumerations, diversity indices, similarity indices, biotic indices, functional feeding group measures, and the multimetric approach. In the Netherlands, only biotic indices focusing on the detection of organic pollution have been applied widely, and multivariate approaches have been developed (Verdonschot & Nijboer, 2000).

The first biotic indices applied in the Netherlands were those developed by Kolkwitz & Marsson (1909) and Sládeček (1973). These were already existing saprobic indices developed to detect organic pollution in Mid-European streams. It soon became clear that Dutch streams often possess distinctive features that require a different approach to assessment. For example, the current velocity in most Dutch streams is considerably lower than that of streams in other more mountainous European countries. These experiences initiated the development of a Dutch assessment system for organic pollution in lowland streams (Moller Pillot, 1971). The K135-index (Tolkamp & Gardeniers, 1971) was based on the Moller Pillot classification (Moller Pillot, 1971) and used for decades.

The biotic indices discussed above are generally limited to a single impact factor, namely organic pollution. The disadvantage of an index reflecting a single aspect of the stream is that it may fail to reveal the effects of other or combined impact factors (Forte et al., 1994; Barbour et al., 1996). This problem was overcome by the introduction of EBEOSWA (ecological assessment of running waters) (Stichting Toegepast Onderzoek Waterbeheer, 1992), a system for the biological assessment of Dutch streams. EBEOSWA assesses more than one impact factor; as such it can be qualified as a multimetric index. The system considers metrics related to stream velocity, saproby, trophy, functional feeding groups, and substrate. The disadvantages of the system are separate scores for each metric instead of one final classification for a location and not determining the ecological status of a water
body by comparing the actual status of a body with near-natural reference conditions. Furthermore, EBEOSWA is based on data collected in the 1980s. These data comprised mainly impacted sites, and collection and identification was not performed in a standardized manner. Also, EBEOSWA has never been validated or subjected to peer review.

The European Water Framework Directive (WFD) has led to a demand for a ‘new’ Dutch assessment system. With the implementation of the WFD, every EU member state is obligated to assess the effects of human activities on the ecological quality of all water bodies. The criteria set by the WFD for the assessment of streams are (European Commission, 2000):

- the use of different biological water quality elements: benthic invertebrate fauna, macrophytes and phytobenthos, phytoplankton, fish fauna;
- the ecological status of a water body is determined by comparing the composition of the biological community in the investigated body with near-natural reference conditions;
- it is based on a stream-type specific approach;
- the final classification of water bodies ranges from 5 (high status) to 1 (bad status).

One of the objectives of this thesis is to develop and test a multimetric index for Dutch streams based on macroinvertebrates that meets the criteria of the WFD. Chapter 2 describes the development and validation of this multimetric index.*

Variation and accuracy in biological monitoring

Before the biological condition can be assessed at a site, samples from the macroinvertebrate community present at the site will have to be collected and processed. The collection and processing of macroinvertebrate samples consists of a sequence of steps (Fig. 1.1). Each step in this sampling and sample processing chain represents choices that have to be made, such as “Do we sample all habitats?” and “Do we identify to genus or species level?” Depending on the choice, the actual composition and condition of the macroinvertebrate community may be misinterpreted (Diamond et al., 1996). The choice will influence the final result, the taxa list, including the number of individuals per taxon. Because biological assessment is based on this taxa list, results can vary based on the choices made during sampling and sample processing. Nijboer (2006) focused on the effects of choices made during data analysis on the results of an ecological typology or assessment system for

* Since the introduction of the multimetric index described in Chapter 2 a WFD-compliant bioassessment system has been developed that can be applied to most types of Dutch surface waters: the ‘KRW maatlatten’ (Van der Molen et al., 2012).
surface water. In this thesis, the focus is on the effects of choices made during the steps of sampling and sample processing.

Biological monitoring usually has two purposes: (1) to estimate variables of interest at a site and (2) to make comparisons among sites or time intervals. Variables of interest in biological monitoring are primarily metric values (e.g., the number of taxa, Average Score Per Taxon, Saprobic Index) and ecological quality classes resulting from biological assessment systems. Metric values and ecological quality classes are calculated based on the macroinvertebrate community composition. Various methods have been developed to collect macroinvertebrates from streams and to process macroinvertebrate samples. These sampling and sample processing methods can vary in terms of sampled area, mesh size of sampling gear, sampled habitats, intensity of sorting, and taxonomic resolution of identification, among other parameters. The methodology that is applied influences the accuracy and variability of bioassessment results (expressed as metric values and/or ecological quality classes) (e.g., Barbour & Gerritsen, 1996; Diamond et al., 1996; Haase et al., 2004). Also, each method can be selective for certain species or groups of species that vary in their exposure and sensitivity to anthropogenic stress (Barton & Metcalfe-Smith, 1992).
Accuracy and variability are both important aspects of bioassessment. Variability refers to the extent to which data points in a statistical distribution or data set diverge from the average or mean value. Accuracy refers to the closeness of a measurement to its true value (Norris et al., 1992). Therefore, differences in accuracy between methods may result in different bioassessment results. Differences in accuracy depend on the spatial and temporal scale at which the true value is defined - a method might be accurate at representing the organisms present in a sample, but less accurate at representing the biota at a site. Variability is important when making comparisons because the validity of conclusions depends on data variability (Norris et al., 1992); higher variability increases the probability of incorrect bioassessment results. An increase in accuracy or a reduction in variability is not always possible because the associated costs are often high. However, when assessing ecological quality for biological monitoring purposes, catching all organisms or taxa present at a site is not necessary (Barbour & Gerritsen, 1996). The standardization of sampling is required, though, for valid comparisons among sites and points in time (Courtemanch, 1996; Vinson & Hawkins, 1996). Thus, the question to focus on is which steps of sampling and sample processing need to be standardized. When two methods are equally variable and provide comparable bioassessment results, standardization is not necessary. Extensive evidence indicates that at least two steps in the sampling and sample processing chain require standardization when metrics based on taxa richness are considered: the sampled area and the effort spent sorting samples. For example, several studies have shown that the number of taxa collected from a sample increases asymptotically with an increase in sampled area and/or sorting effort (e.g., May, 1975; Verdonschot, 1990; Colwell & Coddington, 1995; Vinson & Hawkins, 1996).

In addition to accuracy and variability, cost plays an important role in decision-making related to method standardization. The cost of collecting and processing macroinvertebrate samples is high and can depend strongly on the sampling technique used (e.g., Barbour & Gerritsen, 1996; Metzeling et al., 2003; Vlek et al., 2006). Higher variability and lower accuracy increases the risk of incorrect assessment results. In the case that ecological quality at a site is incorrectly assessed as less than good, water managers will unnecessarily take costly restoration measures to reach a good ecological quality by 2015 (European Commission, 2000). From this point of view, the consequences of poor decision-making due to low accuracy and/or high variability potentially outweighs the savings associated with a less time consuming sampling and sample processing method (Doberstein, 2000).
Information on variability and accuracy is not only important in relation to the standardization of sampling and sample processing methods, but this information can also play an important role in deciding which metrics to incorporate in a biological assessment system. Metrics that exhibit relatively high variability will have more problems discerning signal (sensitivity to anthropogenic stress) from noise (variability).

Since the introduction of the WFD, water authorities have been obliged to monitor changes in ecological quality on larger spatial scales as opposed to site scale and to 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 meet these requirements, the statistical power of the monitoring programs should be analyzed. The statistical properties associated with freshwater monitoring programs are often unknown. Power analysis (assessing the ability of a program to accurately detect change) could help avoid unnecessary expenditures for monitoring programs that cannot provide meaningful results or that lead to overspending. The statistical power of monitoring programs depends, in part, on the variability of biological assessment results.

Given the importance of accuracy, variability, and cost in the decision-making process, one of the main objectives of this thesis is, to gain insight into the variability/accuracy of individual metrics 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. Three different steps from the sampling and sample processing chain that can influence the variability/accuracy of assessment results were studied: sample area (Chapter 3), sampling period (Chapter 4), and the use of preservative before sorting samples (i.e., dead specimens vs. living specimens) (Chapter 5).

In Chapter 3 the implications of a change in (physical) sample size, or sample area, on the variability and accuracy of metric values, bioassessment results, and costs is studied. 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 (AQEM consortium, 2002). The AQEM method is relatively time-consuming. Thus, the study described in Chapter 3 explores the consequences of reducing the sample size in regards to cost and bioassessment results. In Chapter 4 the effect of seasonal variation in macroinvertebrate community composition on metric values is studied. National monitoring protocols are available in many European countries (e.g., Spain, Sweden, Slovakia, Germany, The Netherlands). All these protocols dictate when to collect macroinvertebrate samples, but in most cases scientific evidence for the
indicated time period is lacking. Chapter 5 deals with whether significant differences exist in the metric values, bioassessment results, and costs of sample processing between preserved (i.e., sorting dead specimens) and unpreserved (i.e., sorting living specimens) samples (accuracy). In the few studies that compared sorting results between preserved and unpreserved samples, unpreserved samples were sorted in the field and preserved samples were sorted in the laboratory (e.g., Humphrey et al., 2000; Metzeling et al., 2003; Haase et al., 2004; Nichols & Norris, 1996). The findings of these studies are the result of field sorting and other aspects of sample processing rather than sorting living specimens. Therefore, sorting under laboratory conditions is studied in this thesis.

Whereas chapters 3, 4, and 5 deal with specific aspects of sampling and sample processing and their influence on variability and accuracy, Chapter 6 deals with the subject of variability from a broader perspective. The main objective of this chapter is to quantify the spatial and temporal variability of taxonomic richness metrics based on macroinvertebrates in a minimally impaired system of drainage ditches. This information makes it possible to determine the minimum number of monitoring sites required to detect changes due to anthropogenic disturbances and/or restoration measures (power analysis).

Conservation ecology

The assessment of biological quality has a long history in freshwater ecosystems. With the introduction of the WFD and the Clean Water Act this focus has become even stronger in Europe and the United States, respectively. In terms of macroinvertebrates, the assessment and monitoring of freshwater ecosystems is focused primarily on sampling the “complete” community. Terrestrial ecosystem monitoring is focused primarily on the conservation of species diversity in general, and more specifically on the conservation of rare or threatened species. Because monitoring all species is not feasible in terms of cost, a selection of individual species is used to represent the integrity of the complete ecosystem (Manley et al., 2004). As stated by Maxwell & Jennings (2005), composite indicators composed of several species have the disadvantage that positive trends in some species can mask negative trends in other species. Thus, the extinction of individual species could occur without being noticed, which might be judged as unacceptable by conservation managers. Water managers, on the other hand, are generally more interested in changes in the ecological status of macroinvertebrate communities than changes in the presence/absence or numeric abundance of individual species.
One reason for this is that natural variability in community metrics is generally much lower than natural variability in the presence-absence and numeric abundance of individual species (Fore et al., 1996). Another reason is that water managers often reason that the disappearance of individual species does not necessarily cause significant biological effects on the functioning of a complete community (e.g., Chapin et al., 1997; Holling, 1973).

Thomas (2005) concluded that no nationally reliable monitoring schemes exist for estimating long-term (i.e., 20+ years) changes in freshwater invertebrate species frequency and distribution. In the Netherlands, the introduction of the Red Data Books for Ephemeroptera, Plecoptera, Trichoptera, and Tricladida (Verdonschot et al., 2003) and the obligation arising from the Habitat Directive to report the first assessment of the conservation status of all habitats and species of Community interest, have led to an increased demand for information about the frequency and distribution of individual freshwater invertebrate species. To make monitoring programs cheaper in the future, it is an important question whether samples collected for the purpose of assessing ecological quality of surface waters, can also be used to provide conservation managers with reliable information on individual freshwater invertebrate species. Thomas (2005) already recommended that conservation organizations can take advantage of the existing monitoring programs for the biological assessment of surface waters to monitor changes in freshwater invertebrate biodiversity. Therefore, the study described in Chapter 6 aimed to determine whether water authorities’ current monitoring programs can provide the information on trends in the frequency and distribution of individual freshwater invertebrate species required by conservation managers.

Finally, a synthesis of the preceding chapters is provided in Chapter 7. The implications of the results from the previous chapters on the design of cost-effective monitoring programs will be discussed. Here, the question of how the results from this thesis can be applied 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 is addressed. Furthermore, Chapter 7 deals with some other important issues in biological assessment: (1) the need for biological assessment in addition to assessment based on physical and chemical water quality variables, (2) the lessons that can be learned from the development of biological assessment systems in the past and present, (3) the lack of diagnostic power of current biological assessment systems, and (4) the role of species traits in developing ‘new’ tools for biological assessment. Figure 1.2 provides a schematic overview of the structure of this thesis, including the relationships between the different chapters.
Development of a biological assessment system
(chapter 2)

variability, accuracy and costs associated with sampling and sample processing

- Sample size (chapter 3)
- Sampling period (season) (chapter 4)
- Sorting (chapter 5)

Spatial and temporal variation (chapter 6)

Combining monitoring objectives? (chapter 6)

Design of a cost-effective monitoring program (chapter 7)

Figure 1.2: Schematic overview of the structure of this thesis including the respective chapters.

References


quality of fresh waters: RIVPACS and other techniques. Freshwater Biological Association, Cumbria, United Kingdom.


