Hydrobiologia (2006) 566:523–542

M.T. Furse, D. Hering, K. Brabec, A. Buffagni, L. Sandin & P.F.M. Verdonschot (eds), The Ecological Status of European Rivers: Evaluation and Intercalibration of Assessment Methods
DOI 10.1007/s10750-006-0074-7

Influence of macroinvertebrate sample size on bioassessment of streams

Hanneke E. Vlek[1,*], Ferdinand Šporka[2] & Il'ja Krno[3]
[1]*Alterra, Green World Research, P.O.Box 47, 6700 AA Wageningen, The Netherlands*
[2]*Department of Hydrobiology, Institute of Zoology, Slovak Academy of Sciences, Dúbravská cesta 9, SK-84506 Bratislava, Slovakia*
[3]*Department of Ecology, Faculty of Natural Sciences of Comenius University, Mlynská dolina B-2, SK-84215 Bratislava, Slovakia*
*(*Author for correspondence: E-mail: hanneke.vlek@wurl.nl)*

Key words: sample size, costs, macroinvertebrates, metrics, bioassessment, streams, the Netherlands, Slovakia

Abstract

In order to standardise biological assessment of surface waters in Europe, a standardised method for sampling, sorting and identification of benthic macroinvertebrates in running waters was developed during the AQEM project. The AQEM method has proved to be relatively time-consuming. Hence, this study explored the consequences of a reduction in sample size on costs and bioassessment results. Macroinvertebrate samples were collected from six different streams: four streams located in the Netherlands and two in Slovakia. In each stream 20 sampling units were collected with a pond net (25×25 cm), over a length of approximately 25 cm per sampling unit, from one or two habitats dominantly present. With the collected data, the effect of increasing sample size on variability and accuracy was examined for six metrics and a multimetric index developed for the assessment of Dutch slow running streams. By collecting samples from separate habitats it was possible to examine whether the coefficient of variation (CV; measure of variability) and the mean relative deviation from the "reference" sample (MRD; measure of accuracy) for different metrics depended only on sample size, or also on the type of habitat sampled. Time spent on sample processing (sorting and identification) was recorded for samples from the Dutch streams to assess the implications of changes in sample size on the costs of sample processing. Accuracy of metric results increased and variability decreased with increasing sample size. Accuracy and variability varied depending on the habitat and the metric, hence sample size should be based on the specific habitats present in a stream and the metric(s) used for bioassessment. The AQEM sampling method prescribes a multihabitat sample of 5 m. Our results suggest that a sample size of less than 5 m is adequate to attain a CV and MRD of ≤10% for the metrics ASPT (Average Score per Taxon), Saprobic Index and type Aka + Lit + Psa (%) (the percentage of individuals with a preference for the akal, littoral and psammal). 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 even adequate to attain a CV and MRD of ≤20%. Accuracy of the multimetric index for Dutch slow running streams can be increased from ≤20 to ≤10% with an increase in labour time of 2 h. Considering this low increase in costs and the possible implications of incorrect assessment results it is recommended to strive for this ≤10% accuracy. To achieve an accuracy of ≤10% a multihabitat sample of the four habitats studied in the Netherlands would require a sample size of 2.5 m and a labour time of 26 h (excluding identification of Oligochaeta and Diptera) or 38 h (including identification of Oligochaeta and Diptera).

Introduction

One of the objectives of the European Water Framework Directive (WFD; European Commission, 2000) is to standardise the biological assessment of surface waters in Europe. In the AQEM project assessment systems based on macroinvertebrates, which meet the requirements of the WFD (Hering et al., 2004), were developed. For example, an assessment system for slow running streams was developed in the Netherlands (Dutch AQEM assessment system; Vlek et al., 2004). For the development of the assessment systems data were collected in eight European countries using a standardised method for sampling, sorting and identification (Hering et al., 2004). This standardised AQEM method requires a pond net (width 25 cm) or kick sample collected over a length of 5 m, divided into 20 sampling units of 25 cm. The 20 sampling units are proportionally distributed over the habitats present in a stream consistent with their relative coverage. The AQEM method has proved to be relatively time-consuming, i.e., sample processing of Dutch samples can take 155 h per sample (Vlek, 2004). Before water managers are willing to apply the AQEM method for the purpose of biological monitoring the costs associated with the method will have to be drastically reduced.

Costs of monitoring can, among others, be reduced by reducing the sample size. The interpretation of the concept of sample size is variable. Cao et al. (1997) and Bartsch et al. (1998) interpreted sample size as the number of samples (replicates), while Metzling & Miller (2001) interpreted sample size as the physical size of a sample. In most cases a decrease in the costs of biological monitoring programs has been achieved by limiting the number of samples or restricting the number of organisms picked (Metzling & Miller, 2001). The implications of these measures to reduce costs have been the subject of many studies (e.g., Needham & Usinger, 1956; Chutter, 1972; Elliot, 1977; Barbour et al., 1996; Somers et al., 1998; Lorenz et al., 2004). The implications of reducing the physical sample size, however, have hardly been studied. Also, investigations concerning the number of replicate samples are not relevant in the context of biological monitoring by water managers, since water managers usually take only one multihabitat sample for the purpose of biological monitoring. This multihabitat sample consists of several sampling units from different habitats and all sampling units together form one composite multihabitat sample. In this study we, therefore, addressed the influence of physical sample size instead of the number of replicate samples.

Two important aspects of biological monitoring results should be considered in making decisions on the applied sample size: variability and accuracy. Biological monitoring usually has two purposes: (1) to estimate variables of interest at one site and (2) to make comparisons among sites or times. Variables of interest in biological monitoring are primarily metric values (e.g., the number of taxa, ASPT values, BMWP values) and ecological quality indications resulting from assessment systems. Accuracy is a very important aspect of estimating metric values, since accuracy refers to the closeness of a measurement to its true value (Norris et al., 1992). For the purpose of this study the definition of accuracy by Norris et al. (1992) has been adopted. The aspect of variability is very important in making comparisons, because the validity of conclusions depends on data variability (Norris et al., 1992). Higher variability and lower accuracy increase the risk of incorrect assessment results. In case the 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 outweigh the savings associated with a smaller sample size (Doberstein, 2000).

Given the importance of accuracy, variability and costs in the process of decision-making, the aim of this study was to assess the implications of changes in sample size for different habitats on (1) the variability and accuracy in metric values, (2) the variability and accuracy of assessment results calculated with the Dutch AQEM assessment system and (3) the costs of sample processing.

Methods

Study site and data collection

The Netherlands

Streams dominated by a single habitat (coverage > 50%) were selected to enable sampling of that habitat over a total length of 5 m. In total, four sites at four different streams (the Oude beek, the Heelsumse beek, the Tongerensche beek and the Molenbeek) were sampled. Each stream is dominated by a different habitat. The streams represent slow flowing (current velocity < 50 cm/s) middle and downstream reaches of poor to moderate ecological quality in the Netherlands, except for the Oude Beek. The Oude Beek is an upstream reach of good ecological quality. The catchment area of all streams is smaller than 100 km^2 and is located between 0 and 200 m a.s.l. Fine to medium-sized gravel (0.2–2 cm; akal) was sampled in the Oude Beek (N 52° 9′ 47.9″ E 5° 57′ 30.1″), submerged macrophytes (*Callitriche* sp.) in the Heelsumse beek (N 51° 58′ 40.7″ E 5° 45′ 30.6″), sand in the Tongerensche beek (N 52° 20′ 22.9″ E 5° 55′ 47.3″) and FPOM (fine particulate organic matter) in the Molenbeek (N 51° 59′ 26.2″ E 5° 43′ 53.5″). The Heelsumse beek, the Tongerensche beek, and the Molenbeek were selected because they represent a stream type and ecological quality which frequently occurs in the Netherlands. The Oude Beek was selected because gravel is frequently found in streams of good ecological quality.

Sampling took place between June and September 2002. From each stream 20 sampling units of the dominant habitat were collected. A sampling unit was collected by pushing a rectangular pond net (25×25 cm, mesh size 500 μm) through the upper part of the substratum (2–5 cm) over a length of approximately 25 cm. A ruler was used to visually point out the length of approximately 25 cm. The 20 sampling units were collected in buckets, and kept separately during sample processing. In the laboratory the sampling units were stored overnight in a refrigerator, where they were oxygenated until sorting. The sampling units were washed through a 1000 and a 250 μm sieve prior to sorting. Live organisms were sorted from the sampling units by eye and preserved in 70% ethanol, except for Oligochaeta and Hydracarina. Oligochaeta were preserved in 4% formaldehyde and Hydracarina in Koenike fluid. Organisms were identified to the lowest taxonomic level possible, i.e., species level for almost all specimens. Literature used for identification purposes is listed in AQEM consortium (2002: p. 156, Appendix 8). Time spent on sorting and identification of all specimens in each sampling unit was recorded.

Slovakia

In Slovakia, four different habitats were sampled in two streams: Pokútsky potok (N 48° 34′ 14.8″ E 18° 40′ 16.5″) and Hostiansky potok (N 48° 29′ 36.3″ E 18° 28′ 40.1″). Both streams are siliceous mountain streams in the West Carpathian. Their catchment is smaller than 100 km^2 and is located between 200 and 500 m a.s.l. Pokútsky potok represents streams of high ecological quality and Hostiansky potok represents streams of good to moderate ecological quality. Two dominating habitats were sampled in both streams: macrolithal (20–40 cm) and mesolithal (6–20 cm) in Pokútsky potok, akal and microlithal (2–6 cm) in Hostiansky potok. The streams were selected because they represent a range in ecological quality that is frequently found in small siliceous mountain streams in the West Carpathian.

Sampling took place in June 2003. From each habitat 20 sampling units were collected as described for the Dutch streams. The 20 sampling units were collected in buckets, preserved in 4% formaldehyde, and kept separately during sample processing. The buckets were transported to the laboratory. The sampling units were washed through a 1000 μm and a 500 μm sieve in the laboratory prior to sorting. Preserved organisms were sorted from the sampling units by stereomicroscope and preserved in 70% ethanol. Organisms were identified to the lowest taxonomic level possible, i.e., species level for almost all specimens. Literature used for identification purposes is listed in AQEM consortium (2002: p. 143, Appendix 8).

Data analysis

In total 158 sampling units were collected from eight different habitats. The assumption was made that the 20 pooled sampling units from one habitat would accurately represent the macroinvertebrate

community composition of the respective habitat. The 20 pooled sampling units (with a total sample size of 5 m) are therefore referred to as the ''reference'' sample. The sample size is expressed as the length over which the pond net was pushed through the substratum. This length can be easily converted into the sampled area by multiplying it by 0.25 m (width of the pond net). Different numbers and combinations of sampling units were pooled per habitat to ''construct'' composite samples of different sizes. To gain insight into the effect of sample size on variability and accuracy the sampling units from each habitat were randomly reordered 50 times. In case of one sampling unit or 19 sampling units it was only possible to reorder 20 times. For each sample size the randomly selected sampling units were pooled to form a composite sample. Sampling units were selected randomly without replacement because in the field the same area is normally not sampled twice. The described procedure resulted in 50 or 20 replicate (composite) samples per sample size with sample size ranging from 0.25 to 4.75 m. For example, 50 randomly selected combinations of eight sampling units were used to study a sample size of 2 m.

For evaluation, six metrics were selected from an extensive list of metrics that can be calculated with the program ASTERICS version 1.0 (AQEM/STAR Ecological RIver Classification System; http://www.aqem.de): the Saprobic Index (Zelinka & Marvan, 1961), the Average Score per Taxon (ASPT; Armitage et al., 1983), the number of individuals, the number of taxa, the percentage of Epehemeroptera, Plecoptera and Trichoptera taxa (EPT-taxa (%); Lenat, 1988), and the percentage of individuals with a preference for the akal, littoral and psammal (type Aka + Lit + Psa (%); Schmedtje & Colling, 1996). The first reason to select these metrics was that they represent a variety of metric types (taxon richness, community composition, tolerance-intolerance, habitat preference, population attributes). Second, some of these metrics are frequently used in Europe. Third, EPT-taxa (%), type Aka + Lit + Psa (%) and ASPT have proven to be well correlated to anthropogenic stress in Dutch slow running streams and are incorporated in a revised version of the multimetric index for the assessment of Dutch slow running streams described by Vlek et al. (2004). Fourth, EPT-taxa (%) and ASPT have proven to be well correlated to anthropogenic stress in streams with habitats similar to the habitats present in Slovakian mountain streams (Hering et al., 2004).

Metric values were calculated for all composite samples and plotted against the sample size (number of pooled sampling units) (Heyer & Berven, 1973; Bartsch et al., 1998). Species abundances in a sample of a certain size were always standardised to a sample size of 5 m (abundance $\times$ 5/sample size (m)), e.g., species abundances in a composite sample consisting of 10 sampling units (2.5 m) were multiplied by 2 to make them comparable to the species abundances in a composite sample consisting of 20 pooled sampling units (5 m). To compare accuracy between metrics, habitats and sample size, the relative deviation of the metric value for each composite sample from the ''reference'' sample (true value) was calculated. The information concerning accuracy was summarised by calculating the mean relative deviation (MRD) over all composite samples of a certain size.

The coefficient of variation (CV = SD/mean), a measure of variability, was calculated for the metric values of each sample size per habitat. The minimal sample size required to attain a CV and MRD of both $\leq 10\%$ and $\leq 20\%$ was graphically depicted to facilitate the comparison of the effect of sample size on accuracy and variability for different metrics and habitats. The minimal sample size, henceforth referred to as the sample size, required to achieve a certain level of variability or accuracy is used as a measure for variability and accuracy. This is possible because sample size is correlated with variability/accuracy; a larger sample size implies lower variability or higher accuracy. The sample sizes required to reach a CV or MRD of both $\leq 10\%$ and $\leq 20\%$ for the individual habitats (FPOM, sand akal and submerged macrophytes in the Netherlands; akal, macrolithal, mesolithal and microlithal in Slovakia) were summed per country to gain insight into the sample size required for a multihabitat sample.

For all composite samples from Dutch habitats, ecological quality classes were calculated with a revised version of the multimetric index described by Vlek et al. (2004), in order to determine the effects of sample size and habitat on the variability and accuracy in assessment results. The ecological quality class for the samples from

Slovakia was not calculated because no suitable multimetric index was available for the assessment of samples from Slovakian streams.

Sample processing time (time spent on sorting and identification) was recorded for each Dutch sampling unit. The mean sample processing time, including and excluding the time needed for the identification of Oligochaeta and Diptera, was plotted against sample size per habitat to study the consequences of an increase in sample size in terms of costs. A *t*-test ($\alpha = 0.05$) was performed per sample size to look for significant differences in sample processing time between habitats. Residuals were plotted against predicted values to check for normality in sample processing time. No deviations from normality in sample processing time were found.

Results

Variability and sample size

The mean and standard deviation for sample sizes ranging from 0.25 to 4.75 m are given for each metric and habitat in the supplementary material[1]. Depending on the metric, the effect of increasing sample size on metric values showed different types of responses (supplementary material). A decrease in variation with increasing sample size and a relative stable mean (e.g., Fig. 1) was observed for the following metrics: number of individuals, Saprobic Index, type Aka + Lit + Psa (%) and EPT-taxa (%) (supplementary material). A decrease in variation and an increase in the mean value with increasing sample size (e.g., Fig. 2) was observed for the number of individuals and the number of taxa (supplementary material). The type of metric response to increasing sample size was identical for all habitats and streams in both the Netherlands and Slovakia (supplementary material). The ASPT values showed either one of the two described responses or an intermediate response (Fig. 3), depending on the habitat (supplementary material).

[1] Electronic supplementary material is available for this article at <http://dx.doi.org/ 10.1007/s10750-006-0074-7> and accessible for authorised users.

The Saprobic Index and the metric type Aka + Lit + Psa (%) showed relatively low variability (Fig. 4). A sample size of 0.5 m or less was in all cases sufficient to reach a CV of ≤ 10%, with two exceptions: (1) in case of the habitat akal (NL) and the Saprobic Index a sample size of 2.5 m was required to reach a CV of ≤ 10% and (2) in case of the habitat submerged macrophytes (NL) and the metric type Aka + Lit + Psa (%) a sample size of 1.5 m was required to reach a CV of ≤ 10% (Fig. 4).

The ASPT and the number of taxa showed intermediate variability (Fig. 4). The sample size required to achieve a CV of ≤ 20% for the ASPT was 0.25 m. However, to achieve a CV of ≤ 10% for the ASPT the sample size had to be much larger for the habitats akal (1.25 m) and sand (1.75 m) in the Netherlands. For the other habitats the sample size required to achieve a CV of ≤ 10% varied between 0.25 and 0.75 m. For the number of taxa the sample size required to achieve a CV of ≤ 20 % was low (0.25–0.75 m). As for the ASPT, however, the sample size had to be much larger to achieve a CV of ≤ 10% (0.75–2 m) and differences between habitats became obvious. Variability in the number of taxa did not increase as a function of the number of taxa or the number of individuals collected from a habitat. For example, the metric number of taxa showed higher variability for sand samples than FPOM samples (Fig. 4), while the number of individuals and the number of taxa collected from the FPOM samples were higher than the number of individuals and taxa collected from the sand samples (Table 1).

The EPT-taxa (%) and the number of individuals showed high variability in most cases (Fig. 4). The sample size required to achieve a CV of ≤ 10% for the EPT-taxa (%) varied highly from 0.5 to 4.25 m in both countries, depending on the habitat. Results for the EPT-taxa (%) from the habitat FPOM are not depicted in Figure 4, because EPT-taxa were only found in three of the 20 sampling units and in very low percentages (3.4% on average). The sample size required to achieve a CV of ≤ 10% for the EPT-taxa (%) was 2.5 m on average, whereas it was 1 m on average to achieve a CV of ≤ 20%. To achieve a CV of ≤ 10%, all habitats required a sample size of at least 1.75 m, except for the habitats akal (NL) and macrolithal (S). The differences between habitats were

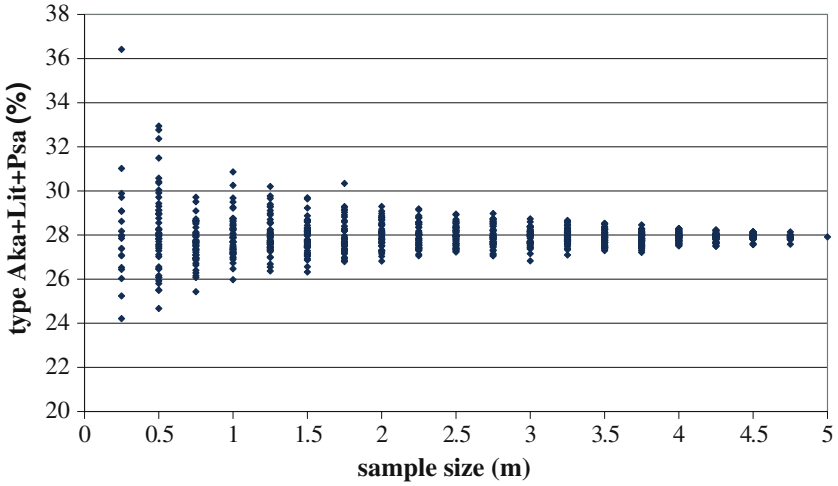

Figure 1. Response of type Aka + Lit + Psa (%) values to increasing sample size for composite FPOM samples from the Molenbeek.

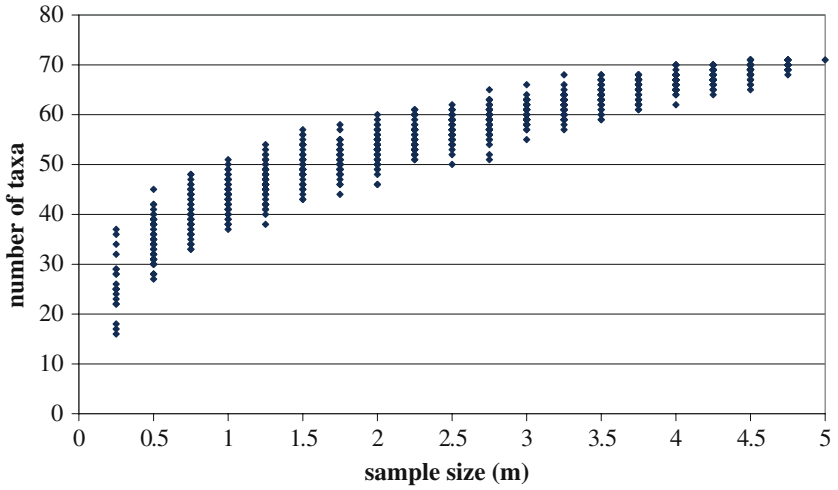

Figure 2. Response of the number of taxa to increasing sample size for composite FPOM samples from the Molenbeek.

somewhat smaller for the number of individuals than for the EPT-taxa (%) with the sample size required to achieve a CV of ≤ 10% ranging from 2.5 to 4 m. On average sampling of 3 m (CV of ≤ 20%) and 1.5 m (CV of ≤ 10%) was required for the number of individuals.

Akal was the only habitat sampled both in the Netherlands and in Slovakia. The difference in the sample size required to achieve a CV of ≤ 10% for this habitat between the Netherlands and Slovakia was less than 0.75 m for the number of individuals, the ASPT and the metric type Aka + Lit + Psa (%) (Fig. 4). The differences in the sample size required to achieve a CV of ≤ 10% were much higher for the number of taxa (1 m), the EPT-taxa (%) (2 m) and the Saprobic Index (2.25 m).

The sample size required to reach a CV of ≤ 10% and ≤ 20% for a multihabitat sample from streams in the Netherlands and Slovakia is shown in Table 2. The sample size required to attain a CV ≤ 10% for the Saprobic index, the metric type Aka + Lit + Psa (%) and the ASPT was considerable smaller than 5 m (between 1.5 and 3.75 m). The minimal sample size required to attain a CV of ≤ 10% for the metrics number of taxa, number of individuals and EPT-taxa (%)

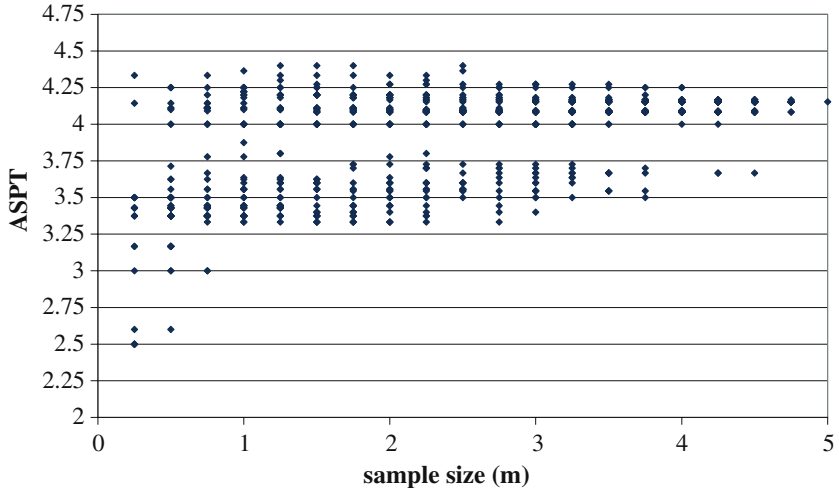

Figure 3. Response of the ASPT values to increasing sample size for composite FPOM samples from the Molenbeek.

varied between 4.5 and 13.75 m. To reach a CV of ≤20% the metrics ASPT, number of taxa, Saprobic Index and type Aka + Lit + Psa (%) required a considerable smaller minimal sample size compared to the EPT-taxa (%) and the number of individuals (between 2.75 and 6.5 m smaller). All metrics, except the number of individuals, required a minimal sample size of less than 5 m to attain a CV of ≤20%.

Accuracy and sample size

The same patterns were observed in the relative accuracy of metrics as in the relative variability of metrics: high accuracy corresponds to low variability. Like the differences in variability (Fig. 4), the differences in accuracy between metrics were high (Fig. 5). The Saprobic Index and the metric type Aka + Lit + Psa (%) showed relative high accuracy (Fig. 5). For both metrics a sample size of 0.25–0.5 m was sufficient to reach a MRD of ≤10%, with two exceptions: (1) in case of the habitat akal and the Saprobic index a sample size of 2.25 m was required and (2) in case of the habitat submerged macrophytes and the metric type Aka + Lit + Psa (%) a sample size of 1.5 m was required.

The ASPT showed intermediate accuracy (Fig. 5). The sample size required to achieve a MRD ≤20% for the ASPT was low (0.25–0.5 m). However, the sample size required to attain a MRD of ≤10% varied from 0.25 to 1.5 m depending on the habitat.

The EPT-taxa (%), number of individuals and number of taxa showed relatively low accuracy (Fig. 5). The sample size required to attain a MRD of ≤10% was 3 m on average for all three metrics. To attain a MRD of ≤20% this was 1.5 m on average. The pattern in relative accuracy for the number of taxa differed (Fig. 5) from the pattern in relative variability (Fig. 4). The metric showed intermediate variability compared to low accuracy.

The differences in accuracy and variability between habitats for the different metrics showed similar patterns (Figs. 4, 5). Differences in accuracy between habitats were larger when the deviation from the "reference" sample was higher, except for the number of taxa (Fig. 5). Differences in variability and accuracy between habitats were highest for the EPT-taxa (%) (Figs. 4, 5). Differences between habitats were minimal for the Saprobic Index values and the metric type Aka + Lit + Psa (%) for both variability and accuracy, with two exceptions: (1) the habitat akal showed low accuracy and high variability for the Saprobic Index and (2) the habitat submerged macrophytes showed low accuracy and high variability for the metric type Aka + Lit + Psa (%) compared to all other habitats (Figs. 4, 5). The difference in accuracy between habitats for the number of taxa was low compared to the differences in variability.

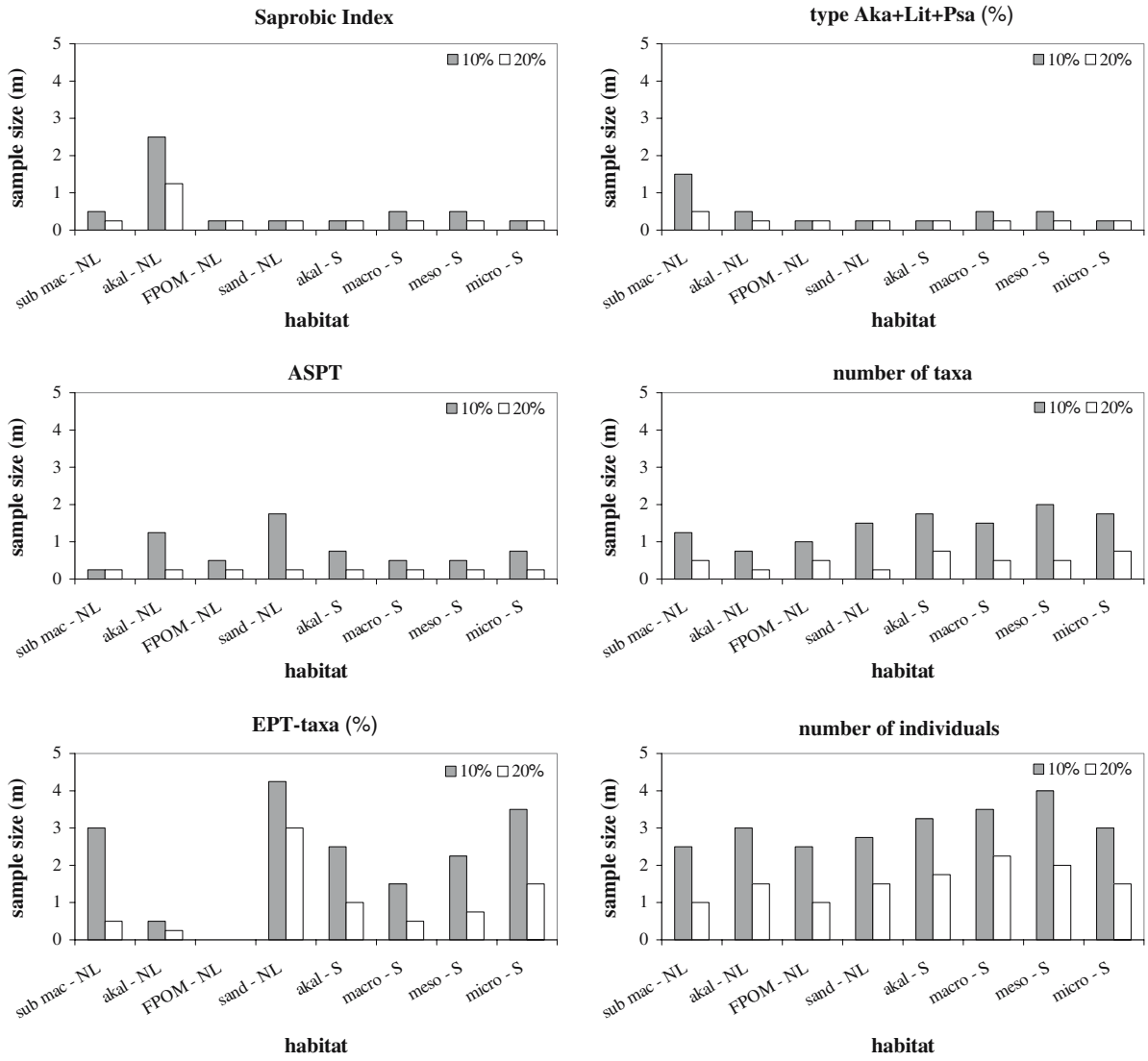

Figure 4. Overview of the minimal sample size required to attain a CV of ≤10% and ≤20 % (maximum) for each combination of habitat and metric (sub mac = submerged macrophytes; macro = macrolithal; micro = microlithal, meso = mesolithal; NL = Netherlands; S = Slovakia).

The differences in the sample size required to attain a MRD of ≤10% for the habitat akal between the Netherlands and Slovakia was less than 0.75 m for all metrics, except for the Saprobic Index (2 m; Fig. 5).

The sample size required to reach a MRD of ≤10% and ≤20% for a multihabitat sample from streams in the Netherlands and Slovakia is shown in Table 2. The sample size required to attain a MRD of ≤10% for the Saprobic index, the metric type Aka + Lit + Psa (%) and ASPT was smaller than 5 m (between 1.25 and 4 m). The sample size required to attain a MRD of ≤10% for the metrics number of taxa, number of individuals and EPT-taxa (%) varied between 6.75 and 15.5 m. To reach a MRD of ≤20% the metrics ASPT, Saprobic Index and type Aka + Lit + Psa (%) required a considerable smaller sample size compared to the EPT-taxa (%), the number of taxa and the number of individuals (between 1 and 10.25 m smaller). All metrics, except the EPT-taxa (%) from Dutch streams and the number of taxa,

Table 1. Overview of the number of individuals and number of taxa collected from the 20 sampling units per habitat and country

Habitat	Number of individuals	Number of taxa
The Netherlands		
Akal	2759	59
Submerged Macrophytes	3032	44
FPOM	7693	71
Sand	5404	63
Slovakia		
Akal	3246	54
Microlithal	2152	59
Mesolithal	1056	58
Macrolithal	1198	66

required a sample size of less than 5 m to attain a CV of ≤20%.

Assessment and sample size

The relation between sample size and the deviation from the ecological quality class associated with the ''reference'' sample differed between habitats. Assessment results for the habitat FPOM did not depend on sample size; a sample size of only 0.25 m resulted in all cases in an ecological quality class identical to that of the ''reference'' sample (Table 3). Assessment results for the habitat sand deviated from the ''reference'' samples for sample sizes varying between 1 and 1.75 m, but only in 4% of the cases (Table 3). In many cases small samples (0.25–0.75 m) from the habitats submerged macrophytes and akal showed a deviation in ecological quality class from the ''reference'' sample. To reduce the percentage of samples indicating an ecological quality class deviating from the ''reference'' sample to less than 10%, a sample size of at least 1 m is required when collecting samples from submerged macrophytes or akal (Table 3).

Sample processing costs

Mean sample processing time (or costs) increased with sample size for all habitats (Fig. 6). A twofold increase in sample size resulted in approximately a doubling of the costs. The relative increase in costs with an increase in sample size of 0.25 m (for sample sizes larger than 0.5 m) was relatively low (≤factor 1.3). The absolute increase in costs, however, was considerable, e.g., between 139 and 519 min for an increase in sample size from 0.75 to 1 m.

Costs varied considerably between habitats (Fig. 6). Irrespective of sample size, costs significantly differed between habitats ($p < 0.001$), except

Table 2. Overview of the minimal multihabitat sample size required to attain a CV of ≤10%, a CV of ≤20%, a mean relative deviation of ≤10% and mean relative deviation of ≤20% for each combination of metric and country (NL = The Netherlands; S = Slovakia)

Metric	Country	Sample size (m)			
		CV ≤10%	Mean relative deviation ≤10%	CV ≤20%	Mean relative deviation ≤20%
Type Aka + Lit + Psa (%)	NL	2.5	2.5	1.25	1.25
Type Aka + Lit + Psa (%)	S	1.5	1.25	1	1
EPT-taxa (%)	NL	7.75	9.75	3.75	5.25
EPT-taxa (%)	S	9.75	9.25	3.7	3.5
Number of individuals	NL	10.75	6.75	5	2.75
Number of individuals	S	13.75	12.25	7.5	6
ASPT	NL	3.75	3	1	1.5
ASPT	S	2.5	4.5	1	1
Number of taxa	NL	4.5	13.75	1.5	9.75
Number of taxa	S	7	15.5	2.5	11.25
Saprobic Index	NL	3.5	3.25	2	1.75
Saprobic Index	S	1.5	1.25	1	1

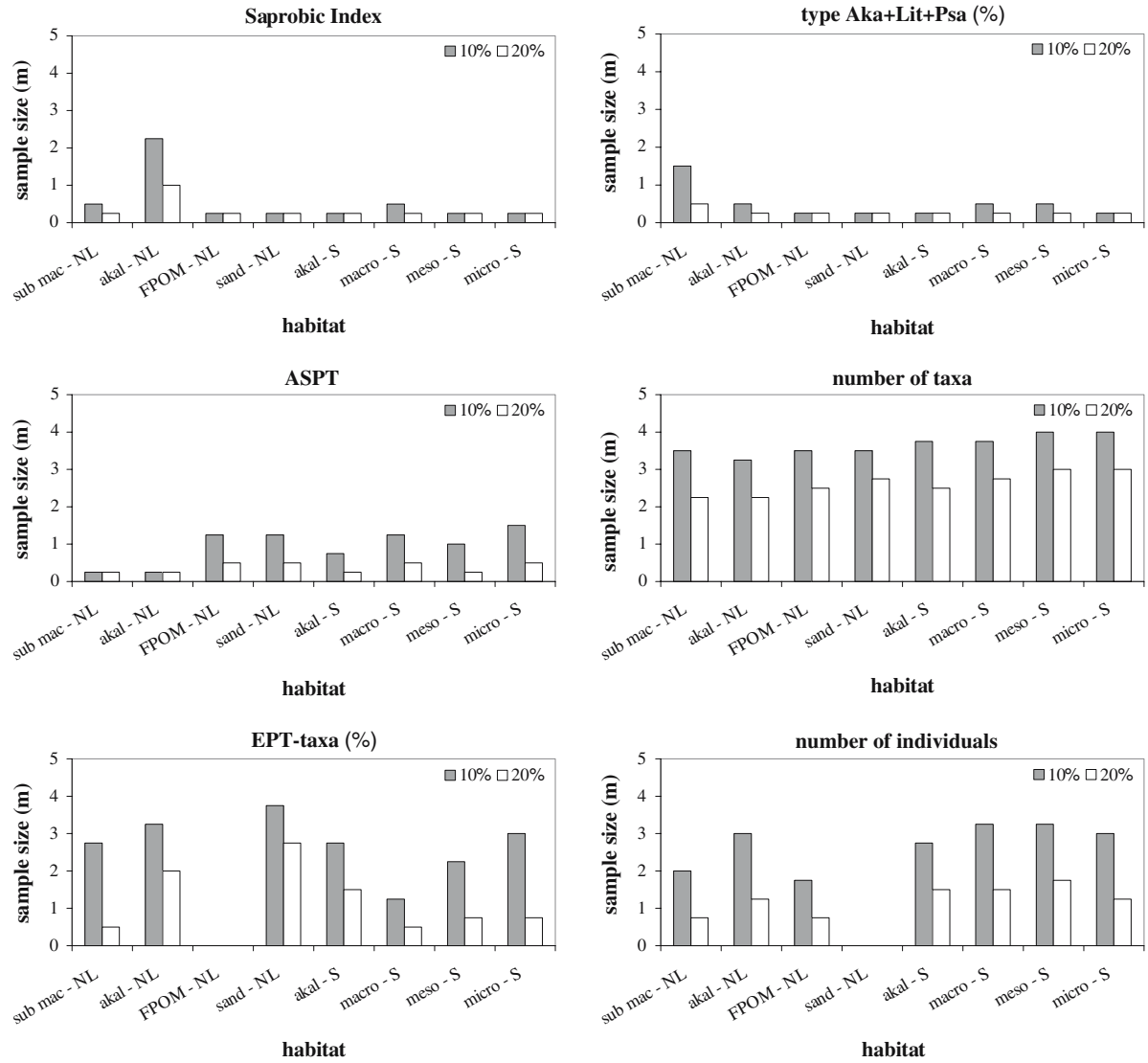

Figure 5. Overview of the minimal sample size required to attain a mean relative deviation of ≤ 10% and ≤ 20 % for each combination of habitat and metric (sub mac = submerged macrophytes; macro = macrolithal; micro = microlithal, meso = mesolithal; NL = The Netherlands; S = Slovakia).

for costs between sand and akal samples that did not differ significantly for a sample size of 0.25 m ($p = 0.053$). Processing of FPOM samples proved to be the most costly, followed by samples from the habitat sand, akal and submerged macrophytes, respectively (Fig. 6). The differences in costs between sand, akal and submerged macrophytes samples were relatively small compared to the differences in costs between FPOM samples and samples from all other habitats (Fig. 6).

Costs were related to the number of individuals collected from a sample. Costs for FPOM samples were relatively high, and so was the number of individuals collected from the FPOM samples (Fig. 6 and Table 1). The costs of FPOM samples were high compared to sand samples (factor 2.2 higher) and so was the number of individuals collected from FPOM samples (factor 1.4 higher). However, the differences in costs between FPOM and sand samples could not be completely

Table 3. Overview of the percentage of samples indicating an ecological quality class different from the ''reference sample'' per habitat (sampled in the Netherlands) and sample size

Sample size (m)	Habitat			
	Submerged macrophytes	Akal	FPOM	Sand
0.25	25	26	0	0
0.5	55	30	0	0
0.75	16	24	0	0
1	6	6	0	2
1.25	8	2	0	4
1.5	0	0	0	4
1.75	2	0	0	4

Percentages for sample sizes larger than 1.75 m are not listed, because these were zero.

explained by the differences in the number of individuals; the costs of FPOM samples were much higher than expected based on the number of individuals.

Costs were greatly reduced by not identifying Oligochaeta and Diptera (Figs. 6, 7). The costs of sand samples were reduced with a factor 2.7, of FPOM samples with a factor 1.9, of akal samples with a factor 1.3, and of submerged macrophytes with a factor 1.2. These reductions in costs were related to the number of Oligochaeta and Diptera individuals present in the samples. The FPOM and sand samples consisted for approximately 70% of Oligochaeta and Diptera individuals, while this percentage was only 40% for akal samples and 18% for submerged macrophytes samples. Even when Oligochaeta and Diptera were not identified the costs of FPOM samples were still the highest, followed by samples from the habitat akal, submerged macrophytes and sand (Fig. 7). Despite the decrease in costs associated with not identifying Oligochaeta and Diptera, a twofold increase in sample size still resulted in approximately a doubling of the costs.

The cost that had to be made to reach a CV of $\leq 10\%$ and $\leq 20\%$ for the individual habitats and the multihabitat samples are given in Table 4. The costs in Table 4 are directly related to the sample size. Only the costs related to variability are shown in Table 4 because results for accuracy and variability were similar (Figs. 4, 5). The costs of FPOM samples for the EPT-taxa (%) were not included in Table 4 because EPT-taxa were only found in 3 of the 20 sampling units, which means that the total costs for the EPT-taxa (%) were underestimated. The total costs (costs for a multihabitat sample) to achieve a CV of $\leq 20\%$ were high for the number of individuals and the EPT-taxa (%), 96 and 62 h, respectively (Table 4). The total cost to achieve a CV of $\leq 20\%$ for the other metrics varied between 20 and 34 h. To reduce CV from $\leq 20\%$ to $\leq 10\%$ an increase in total costs by a factor of 1.6 (19 h) for the Saprobic Index and by a factor of 1.5 (12 h) for the metric type Aka + Lit + Psa (%) was required (Table 4). The other metrics required an increase in total costs by a factor of 1.8 to a factor of 3.4, or an absolute increase in hours between 50 and 199.

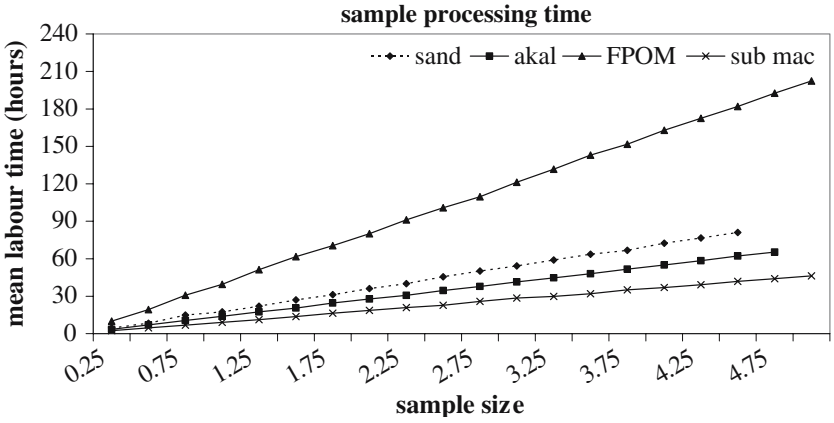

Figure 6. Mean sample processing time as a function of sample size for the habitats sand, akal, FPOM and submerged macrophytes from Dutch streams.

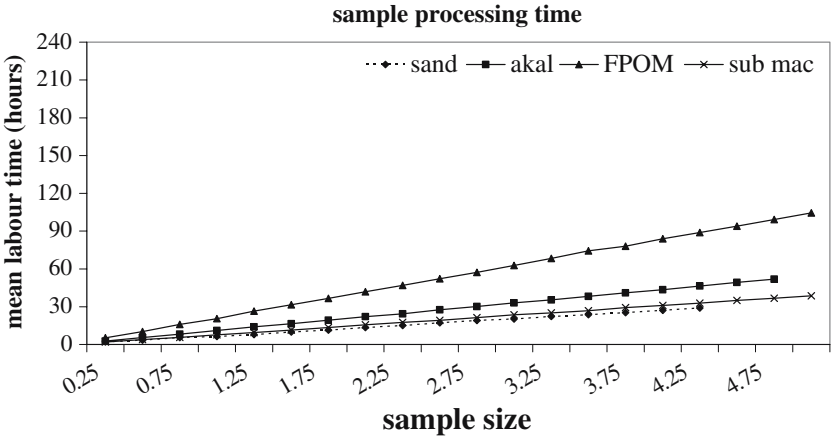

Figure 7. Mean sample processing time (excluding the identification of Oligochaeta and Diptera) as a function of sample size for the habitats sand, akal, FPOM and submerged macrophytes from Dutch streams.

The differences in total costs between metric to reach a CV of ≤10% were much larger than the differences in total costs between metrics to reach a CV of ≤20%. The total costs to reach a CV of ≤10% were low for the Saprobic Index (54 h) and the metric type Aka + Lit + Psa (%) (35 h) compared to the others metrics (between 70 and 215 h) (Table 4).

The absolute differences in total costs between metrics were lower when the costs for the identification of Oligochaeta and Diptera were not included, while the relative differences in total costs between metrics remained similar. When Oligochaeta and Diptera were not identified an increase in total costs by a factor of 1.7 for the Saprobic Index (16 h) and for the metric type Aka + Lit + Psa (%) (10 h) was required to reduce CV from ≤20% to ≤10% (Table 4). The other metrics required an increase in costs by a factor of 2.1 to a factor of 3.1, or an absolute increase in hours between 26 and 69, when Oligochaeta and Diptera were not identified (Table 4).

To gain accuracy in assessment results, by reducing deviations from the ecological quality class with the "reference" sample, from ≤20% to ≤10% sample size (and costs) did not have to be increased for the habitats FPOM, sand and akal (Table 3). The habitat-submerged macrophytes required an increase in sample size from 0.75 to 1 m to achieve this gain in accuracy (Table 3), which is equal to an increase in labour time of 2 h (Fig. 6).

Discussion

Methodological approach

The optimal sample size is the largest possible (Green, 1979). One of the restrictions of this study was that variation and accuracy were studied based on the assumption that a sample size of 5 m would cover all variation of one habitat at a site. The data showed decreasing variation in metric values and increasing accuracy with increasing sample size. The decrease in variation with sample size might have been more gradual in reality. Samples of different sizes were created by randomly combining samples from the complete pool of 20 sampling units. The question is whether variation might have been higher if the samples of different sizes had been collected in the field. It is difficult to judge whether the 5 m sampled in this study covers all variation at a site. Compared to the sample sizes applied in biological surveillance monitoring an area sampled of 1.25 m^2 (=sampling over a length of 5 m) from *one habitat* is quite large, e.g., the mean area sampled in macroinvertebrate monitoring programs by USA state agencies is 1.7 m^2 for a *mulitihabitat* sample (Carter & Resh, 2001).

The sample size of the individual sampling units was approximately 25 cm. It was not possible to sample exactly 25 cm without disturbing the substrate prior to sampling. The small variation in sample size between the sampling units is not

Table 4. Overview of the sample processing time required to attain a CV of ≤10% and ≤20% including and excluding (labour time excl.) the identification of Oligochaeta and Diptera per habitat and metric

Metric	Habitat	Labour time (hours)		Labour time excl. (hours)	
		CV ≤10%	CV ≤20%	CV ≤10%	CV ≤20%
ASPT	Akal	17	3	14	3
	FPOM	19	10	10	5
	Sand	31	5	11	2
	Sub mac	2	2	2	2
	Total	**70**	**20**	**38**	**12**
EPT-taxa (%)	Akal	7	3	5	3
	FPOM	0	0	0	0
	Sand	77	54	29	20
	Sub mac	28	5	24	4
	Total	**112**	**62**	**58**	**27**
Number of individuals	Akal	41	20	33	16
	FPOM	101	40	52	20
	Sand	50	27	19	10
	Sub mac	23	9	19	8
	Total	**215**	**96**	**123**	**55**
Number of taxa	Akal	11	3	8	3
	FPOM	40	19	20	10
	Sand	27	5	10	2
	Sub mac	11	5	9	4
	Total	**88**	**32**	**48**	**19**
Saprobic Index	Akal	35	17	28	14
	FPOM	10	10	5	5
	Sand	5	5	2	2
	Sub mac	5	2	4	2
	Total	**54**	**34**	**39**	**23**
Type Aka + Lit + Psa (%)	Akal	7	3	5	3
	FPOM	10	10	5	5
	Sand	5	5	2	2
	Sub mac	14	5	11	4
	Total	**35**	**23**	**24**	**14**

Sample processing time was only recorded for habitat samples collected from streams in the Netherlands.

expected to have consequences regarding the applicability of the results of this study, since it will always be a problem to determine the exact sample size when sampling with a pond net in slow running streams.

Samples in this study have been collected between June and September. The fact that the habitats were not sampled simultaneously might have influenced the results. Studies performed in the Netherlands and in Slovakia, however, indicated that there are no significant differences in the number of individuals, the number of taxa, the EPT-taxa (%), ASPT values or Saprobic Index values between months (Šporka et al., 2006; Vlek, 2004). These findings make it unlikely that differences in variability between habitats were the result of differences between months.

In many European countries samples are preserved prior to sorting, while the samples (from the Netherlands) collected during this study were not preserved. Findings by Vlek (2004) suggest that the choice to preserve a sample or not will not influence variability and accuracy in metric values, i.e., Vlek (2004) detected no significant differences in

the number of individuals, the number of taxa, the EPT-taxa (%), ASPT values or Saprobic Index values between preserved and unpreserved macroinvertebrate samples collected in the Netherlands.

The samples collected in this study came from different streams which makes it difficult to determine the effect of sample size on variability in metric values of a multihabitat sample. In this study the assumption was made that by reaching a CV (or MRD) of $\leq 10\%$ for the individual habitats, a CV (or MRD) of $\leq 10\%$ for the multihabitat samples would be guaranteed. Unfortunately, it was not possible to test this assumption since the habitats in this study came from different streams. Generally, macroinvertebrate community composition differs more among streams than within sites (e.g., Doberstein et al., 2000; Sandin & Johnson, 2000). Consequently, variability would be much higher in combining habitat samples from different streams than combining habitat samples from one stream. According to Beisel (1998) the variability in taxon richness and total abundance does not depend on the number of habitats sampled. This would suggest that metric values based on multihabitat samples would not be more variable than metric values based on single habitat samples, as was assumed in this study. Another difficulty was that the relation between variability/accuracy and multihabitat sample size was based on the four specific habitats sampled in the Netherlands and in Slovakia. This relation will have to be adjusted depending on the number and type of habitats present in the stream that is subjected to monitoring. Carter & Resh (2001) suggested that multihabitat samples would be more variable than single habitat samples, since sampling from multiple habitats in proportion to their cover is most likely to be operator-dependent and therefore more difficult to standardize than collecting from a single habitat samples. The variability in habitat coverage estimates is an extra source of variation that should be studied in the future.

Variability and sample size

High variability in metric values creates problems with assessment. As a result of high variability metric values will overlap between ecological quality classes. This overlap makes it impossible to distinguish between many ecological quality classes, complicating assessment (Doberstein et al., 2000).

When considering costs the metrics type Aka + Lit + Psa (%), Saprobic Index and ASPT should be preferred over the number of individuals, the number of taxa and EPT-taxa (%), for these showed relative low variability and high accuracy, which means that the required sample size to attain a certain degree of variability is smaller. For biological assessment it is important to know whether these metrics are also (highly) correlated to anthropogenic stress. Both the ASPT and the Saprobic Index are frequently applied in Europe and have proven to be highly correlated to organic pollution. The ASPT has been incorporated in multimetric indices in the Czech Rebuplic (Brabec et al., 2004), Greece (Skoulikidis et al., 2004), Italy (Buffagni et al., 2004), Sweden (Dahl et al., 2004) and the United Kingdom (Clarke et al., 2002). The Saprobic Index (or derivations from this index) has been incorporated in multimetric indices in Austria (Ofenböck et al., 2004), the Czech Republic (Brabec et al., 2004), Germany (Rolauffs et al., 2004), the Netherlands (Vlek et al., 2004) and Sweden (Dahl et al., 2004). A possible correlation between anthropogenic stress and type Aka + Lit + Psa (%) values are yet to be established.

The number of taxa and the number of individuals are notoriously poor metrics (Karr & Chu, 1999). The number of individuals showed high variation compared to the other metrics evaluated in this study. Apparently, significant variation in faunal densities occurs over small spatial scale, possibly caused by invertebrate aggregations (Downes et al., 1993).

Differences in variability between habitats depended on the metric studied, indicating that differences in variability between habitats could not be explained based on general assumptions about habitat heterogeneity. In general, metrics characterised by higher variability showed larger differences between habitats.

The large differences in variability for the number of taxa, the EPT-taxa (%) and the Saprobic Index between akal samples from the Netherlands and Slovakia might have been the result of regional differences or different sample processing protocols. The Slovakian

samples were washed through a 500 μm mesh size sieve, while the Dutch samples were washed through a 250 μm mesh size sieve. It is not clear why the differences in variability are so high for the EPT-taxa (%) and the Saprobic Index compared to the other metrics.

Accuracy and sample size

As long as metric values are highly correlated to anthropogenic stress, high accuracy is not per definition required for assessment purposes, since class boundaries applied in an assessment system should always be calibrated based on data. In cases where scientists are interested in the 'true' community composition instead of biological assessment, accuracy (apart from variability) becomes very important. It is difficult to obtain accurate measurements of richness due to the collector's curve phenomenon (Colwell & Coddington, 1994; Fig. 2). This phenomenon resulted in high costs to establish accurate values for the number of taxa and the percentage of EPT-taxa. Colwell & Coddington (1994) stated that the number of taxa encountered in a sample increases asymptotically as a function of both the area sampled and the number of individuals in a sample. Lorenz et al. (2004) suggested that the curve is also a function of taxa diversity and that in streams with lower species diversity richness measures are likely to approach an asymptote at a smaller sample size. In this study no evidence was found to suggest that the number of taxa collected increased as a function of the number of individuals or the number of taxa in a sample. Cao et al. (2002) and Clarke et al. (2002) found that sampling variability in the number of taxa increased with the mean number of taxa recorded at a site. Doberstein et al. (2000) found low variances in metric values in streams with relatively few taxa. This study did not confirm the findings of Doberstein et al. (2000), Cao et al. (2002) and Clarke et al. (2002) because no evidence was found to suggest that the number of taxa collected increases as a function of the number of taxa in a sample and only minor differences were detected between habitats (determines the number of taxa in a sample) in variability and accuracy in the number of taxa compared to Cao et al. (2002). Where Cao et al. (2002) compared differences between habitats in the same river or site we compared habitats from different streams in different countries. Cao et al. (2002) detected differences in total taxon richness of more than 30% (based on one sampling unit). We detected differences in total taxon richness between Dutch habitats of 8% and between Slovakian habitats of 18%. An explanation for the differences between our study and that of Doberstein et al. (2000), Cao et al. (2002) and Clarke et al. (2002) might be the range in the number of taxa collected from the habitats in our study (between 44 and 71 taxa). This assumption is supported by Cao et al. (2002), who showed that relative differences in total taxon richness (%) are much larger when comparing a community of 20 taxa with a community of 60 taxa, than when comparing a community of 60 with a community of 100 taxa. So, caution should be taken in basing decisions concerning sample size on the results of this study when sampling habitats with less than 44 taxa.

Differences in accuracy between habitats depended on the metric studied, indicating that differences in accuracy between habitats could not be explained based on general assumptions about habitat characteristics. In general, metrics characterised by lower accuracy showed lager differences between habitats.

The large differences in accuracy for the Saprobic Index between akal samples from the Netherlands and Slovakia might have been the result of regional differences or different sample processing protocols.

Sample processing costs

Costs were based on identifications to species level and identification of all specimens. Some metrics, however, do not necessitate identification to species level or identification of all groups. For example, the calculation of the Saprobic Index, the metric type Aka + Lit + Psa (%), the ASPT or the EPT-taxa (%) does not require the identification of Oligochaeta and Diptera. In the Netherlands Oligochaeta and Diptera can make up a large part of the total number of individuals in a sample. Instead of determining the costs for the different metrics separately, which would be lengthy, the costs excluding the identification of Oligochaeta and Diptera were determined. This

means that the costs for the ASPT and the EPT-taxa (%) are in reality lower than indicated in this study because these metrics do not necessitate the identification of other groups besides Oligochaeta and Diptera. The assumption made in this study was that often a combination of metrics (multimetric) will be used for assessment, thereby requiring the identification of the majority of the groups. For this reason, differences in costs between metrics were not taken into account. In case these differences in costs are taken into account the metrics ASPT and EPT-taxa (%) might still be calculated against reasonable costs, despite their high variability.

Apart from the groups that are identified, taxonomic resolution plays an important role in the costs associated with sample processing. All cost-related comparisons made in this study have been based on identifications to species level. The ASPT is a metric that requires identification to family level only. When the ASPT is the only metric used for bioassessment purposes and identifications can be performed at family level, the cost associated with the ASPT would probably be comparable to the costs associated with the Saprobic Index or the metric type Aka + Lit + Psa (%).

Differences in sample processing costs between habitats could not completely be related to the number of individuals collected. Other factors, e.g., the characteristics of the collected material sampled (large amounts of small dark particulate matter makes it more difficult to detect organisms) or previous experience of the analysts with the taxa collected also might have played a role.

The samples in this study were collected by pushing the net through the upper layer of the substratum, collecting the complete upper layer. The amount of material and the number of individuals collected through kick sampling or jabbing the substratum would have been much lower (Vlek, 2004). Since costs are directly related to the amount of material and the number of individuals collected (Barbour & Gerritsen, 1996), sample processing costs can expected to be much lower in case of kick sampling or jabbing the substratum instead of sampling the complete upper layer of the substratum.

Assessment and sample size

Reason for this study was the large amount of time that is needed for the processing of samples collected with the AQEM method. In the AQEM project multimetric indices were developed based on multihabitat samples collected according to the AQEM method (Hering et al., 2004). The assessment of anthropogenic stress with multimetric indices based on multihabitat samples has been frequently applied in the United States (Ohio EPA, 1987; Plafkin et al., 1989; Barbour et al., 1992; Kerans et al., 1992; Barbour et al., 1996; Major et al., 1998 and Maxted et al., 2000) and Europe (Hering et al., 2004). Arguments in favour of this approach are (1) by collecting macroinvertebrates from all the habitats present in proportion to their coverage a sample is a better representative of the habitats (and organisms) present in the sampled reach than when collecting from a single habitat (Carter & Resh, 2001); limiting sampling to a single habitat means that certain kinds of anthropogenic stress, which only influence specific habitats, may go undetected (Kerans et al., 1992); (2) multimetric indices provide detection capability over a broader range and nature of stressors and give a more complete picture about ecosystem health (Karr et al., 1986; Barbour et al., 1996).

The calculation of ecological quality classes in this study was based on samples from one habitat. However, the multimetric index used to calculate the classes was calibrated based on multihabitat samples (Vlek et al., 2004). Calculations of the ecological quality classes based on multihabitat samples would most likely have resulted in different classes compared to the calculations based on samples from one habitat. Still, the acquired information is very valuable in the sense that it gives an idea about the sensitivity of assessment results to reductions in sample size.

The differences in the percentage of misclassifications (a deviation in ecological quality class from the "reference" sample) between habitats could not be explained based on general assumptions about habitat heterogeneity; otherwise the variability in metric values would have been higher for samples from submerged macrophytes and akal than for samples from sand and FPOM for all metrics studied. Of the metrics evaluated in this

study the metrics EPT-taxa (%), ASPT and type Aka + Lit + Psa (%) are incorporated in the multimetric index. The differences in misclassification between habitats could neither be explained by the variation in EPT-taxa (%) values. Variability in EPT-taxa (%), ASPT and type Aka + Lit + Psa (%) values together were not higher for submerged macrophytes and akal samples than for sand and FPOM samples. The differences in misclassification between habitats seemed to be related to other metrics incorporated in the multimetric index. The low number of misclassifications for the sand samples did not reflect the relatively high variation in EPT-taxa (%) values, two possible explanations can be (1) EPT-taxa (%) values did not happen to fall near a breakpoint in the scoring criteria (Fore et al., 2001) and/or (2) the combination of several metrics makes the multimetric index robust.

It is difficult to predict the influence of variability/accuracy for different individual metrics on the variability and accuracy of the final assessment result (Vlek, 2004). This is, among others, due to the fact that it is very important whether metric values for a single sample happen to fall near a breakpoint in the scoring criteria (Fore et al., 2001). Water managers will be interested in the probability that assessment results indicate less than good ecological quality while in reality ecological quality is good (false positives, type I error), because false positives will lead to unnecessary restoration measures (CIS working group 2.3, 2003). Organisations dealing with nature conservation will of course be interested in the probability that assessment results indicate good quality while in reality the ecological quality is less than good (false negatives, type II error). It is unlikely that water managers will take more than one multihabitat sample for the purpose of routine biological monitoring, due to costs considerations. So, instead of calculating the number of samples necessary to achieve a low error, they would be interested in knowing the error associated with taking only one sample. With information on the variability in individual metric values, the program STARBUGS (Clarke, 2004) can be used to calculate the effect of differences in estimates of habitat coverage and the effect of variability in individual metric values on the final assessment result of individual samples. The information on variability in the supplementary material can be used to perform the mentioned calculations for different multimetric indices. However, assumptions will have to be made about the variability of multihabitat samples based on single habitat variability. Because it is not clear whether the differences in variability and accuracy between samples from the Netherlands and Slovakia were caused by regional differences or different sample processing protocols, the application of the information in the supplementary material should be limited to the studied stream types in Slovakia and in the Netherlands.

The information in this paper gives scientists and water managers the opportunity of weighing a decrease in variability and an increase in accuracy on the one hand against the increase in costs on the other hand. Hopefully, the outlined approach shows water managers that the consequences of poor decision making potentially outweigh the savings associated with smaller sample area (Doberstein et al., 2000).

Conclusions and recommendations

Accuracy and variability varied depending on the habitat and the metric examined. This leads to the conclusion that sample size applied for biological monitoring should be based on the specific habitats present in a stream and the metric(s) used for bioassessment.

Assessment based on the number of taxa, the ASPT, the EPT-taxa (%) or the number of individuals is relative expensive compared to assessment based on the Saprobic Index or the metric type Aka + Lit + Psa (%), when specimens are identified to species level and a CV of 10% is aspired. These relative expensive metrics also require a high absolute increase in costs to realise a decrease in CV from $\leq 20\%$ to $\leq 10\%$, while this decrease in costs requires (for most habitats) a relative low (or even no) increase in costs for the Saprobic Index and the metric type Aka + Lit + Psa (%). The increase in costs necessary to reduce variability for the Saprobic Index and the metric type Aka + Lit + Psa (%) is certainly justifiable given the possible implications of incorrect assessment results. When assessment of Dutch streams is based on the Saprobic Index or the metric type Aka + Lit + Psa (%) it is, therefore, recommended to strive for a CV of $\leq 10\%$. A CV of $\leq 10\%$ can be achieved by sampling 3.5 m (54 h,

including identification of Oligochaeta and Diptera) in case of the Saprobic Index or 2.5 m (35 h, Oligochaeta and Diptera) in case of the metric type Aka + Lit + Psa (%). The indicated sample sizes for multihabitat samples are based on streams in the Netherlands where the habitats FPOM, akal, submerged macrophytes and sand are present. For streams in Slovakia (small siliceous mountain streams in the West Carpathian) a CV of ≤ 10% can be achieved by sampling 1.5 m in case of both metrics. The indicated sample size is based on multihabitat samples from streams in Slovakia where the habitats akal, macrolithal, mesolithal and microlithal are present.

The recommended multihabitat sample sizes are based on a fixed sample size per habitat and do not depend on the coverage of the individual habitats in a stream. Results of this study suggested that a multihabitat sample size of less than 5 m is also adequate to attain a CV and MRD of ≤ 10% for the metric ASPT. The metrics number of taxa, number of individuals and EPT-taxa (%) require 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 even adequate to attain a CV and MRD of ≤ 20%,

Accuracy of the multimetric index for Dutch slow running streams depends on the sampled habitat(s). No extra costs are associated with an increase in accuracy from ≤ 20% to ≤ 10% for akal, FPOM and sand samples. However, the sample size of submerged macrophytes samples has to be increased from 0.75 to 1 m to achieve this increase in accuracy. This increase in sample sizes equals an increase in labour time of 2 h, which is no much considering the possible implications of incorrect assessment results. Hence, it is recommended to strive for a an accuracy of ≤ 10%, which requires a multihabitat sample size of 2.5 m (0.25 m FPOM, 0.25 m sand, 1 m akal and 1 m submerged macrophytes) and a labour time of 26 h (excluding Oligochaeta and Diptera) or 38 h (including Oligochaeta and Diptera).

Acknowledgements

This study was carried out within the STAR project funded by the European Commission, 5th Framework Program, Energy Environment and Sustainable Development, Key Action Water, Contract No. EVK1-CT-2001-00089. We are very grateful for the helpful comments made by Rebi Nijboer and two anonymous reviewers on an earlier version of the manuscript. We would like to thank Tjeerd-Harm van den Hoek, Martin van den Hoorn, Rink Wiggers, Tomáš Derka, Eva Bulánková, Daniela Illéšová, Zuzana Pastuchová and Zuzana Zaťovičová for their efforts in collecting the data on which this study was based.

References

AQEM consortium, 2002. Manual for the Application of the AQEM System. A Comprehensive Method to Assess European Streams using Benthic Macroinvertebrates, Developed for the purpose of the Water Framework Directive. Version 1.0, February, 2002.

Armitage, P. D., D. Moss, J. F. Wright & M. T. Furse, 1983. The performance of a new biological water quality score system based on macroinvertebrates over a wide range of unpolluted running-water sites. Water Research 17: 333–347.

Barbour, M. T. & J. Gerritsen, 1996. Subsampling of benthic samples: a defense of the fixed-count method. Journal of the North American Benthological Society 15: 386–391.

Barbour, M. T., J. Gerritsen, G. E. Griffith, R. Frydenborg, E. McCarron, J. S. White & M. L. Bastian, 1996. A framework for biological criteria for Florida streams using benthic macroinvertebrates. Journal of the North American Benthological Society 15: 185–211.

Barbour, M. T., J. L. Plafkin, B. P. Bradley, C. G. Graves & R. W. Wisseman, 1992. Evaluation of EPA's rapid bioassessment benthic metrics: metric redundancy and variability among reference stream sites. Environmental Toxicology and Chemistry 11: 437–449.

Bartsch, L. A., W. B. Richarson & T. J. Naimo, 1998. Sampling benthic macroinvertebrates in a large floodplain river: considerations of study design, sample size and cost. Environmental Monitoring and Assessment 52: 425–439.

Beisel, J. N., P. Usseglio-Polatera, S. Thomas & J. C. Moreteau, 1998. Effects of mesohabitat sampling strategy on the assessment of stream quality with benthic invertebrate assemblages. Archiv für Hydrobiologie 142: 493–510.

Brabec, K., S. Zahrádlová, D. Němejcová, P. Pařil, K. Kokeš & J. Jarkovský, 2004. Assessment of organic pollution effect considering differences between lotic and lentic stream habitats. Hydrobiologia 516: 331–346.

Buffagni, A., S. Erba, M. Cazzola & J. L. Kemp, 2004. The AQEM multimetric system for the southern Italian Apennines: assessing the impact of water quality and habitat degradation on pool macroinvertebrates in Mediterranean rivers. Hydrobiologia 516: 313–329.

Cao, Y., W. P. Williams & A. W. Bark, 1997. Effects of sample size (replicate number) on similarity measures in river

benthic Aufwuchs community analysis. Water Environment Research 69: 107–114.

Cao, Y., D. Williams & D. P. Larsen, 2002. Comparison of ecological communities: the problem of sample representativeness. Ecological Monographs 72: 41–56.

Carter, J. L. & V. H. Resh, 2001. After site selection and before data analysis: sampling, sorting, and laboratory procedures used in stream benthic macroinvertebrate monitoring programs by USA state agencies. Journal of the North American Benthological Society 20: 658–682.

Chutter, F. M., 1972. A reappraisal of Needham and Usinger's data on the vriability of a stream fauna when sampled with a Surber sampler. Limnology and Oceanography 17: 139–141.

Common Implementation Strategy (CIS) Working Group 2.3 – REFCOND, 2003. Guidance on Establishing Reference Conditions and Ecological Status Class Boundaries for Inland Surface Waters. European Commission, Version 7.0, 93 pp.

Clarke, R. T., 2004. Error/Uncertainty Module Software STARBUGS (STAR Bioassessment Uncertainty Guidance Software) User Manual. STAR (Standardisation of River Classifications) Deliverable 9. Produced under European Union 5th Framework Programme Contract EVK1-CT 2001-00089.

Clarke, R. T., M. T. Furse, R. J. M. Gunn, J. M. Winder & J. F. Wright, 2002. Sampling variation in macroinvertebrate data and implications for river quality indices. Freshwater Biology 47: 1735–1751.

Colwell, R. K. & J. A. Coddington, 1994. Estimating terrestrial biology through extrapolation. Philosophical Transactions of the Royal Society (Series B) 345: 101–118.

Dahl, J., R. K. Johnson & L. Sandin, 2004. Detection of organic pollution of streams in southern Sweden using benthic macroinvertebrates. Hydrobiologia 516: 161–172.

Doberstein, C. P., J. R. Karr & L. L. Conquest, 2000. The effect of fixed-count subsampling on macroinvertebrate biomonitoring in small streams. Freshwater Biology 44: 355–371.

Downes, B. J., P. S. Lake & E. S. G. Schreiber, 1993. Spatial variation in the distribution of stream macroinvertebrates. Implications of patchiness for models of community organization. Freswater Biology 30: 119–132.

Elliot, J. M., 1977. Some Methods for the Statistical Analysis of Benthic Invertebrates, 2nd edn. Sci. Publ. No. 25, Freshwater Biological Association, Ferry House, U.K., 156 pp.

European Commission, 2000. Directive 2000/60/EC of the European Parliament and of the Council – Establishing a Framework for Community Action in the Field of Water Policy. Brussels, Belgium, 23 October 2000.

Fore, L. S., K. Paulsen & K. O'Laughlin, 2001. Assessing the performance of volunteers in monitoring streams. Freshwater Biology 46(1): 109–123.

Green, R. H., 1979. Sampling Design and Statistical Methods for Environmental Biologists. John Wiley and Sons, New York, 257 pp.

Heyer, R. W. & K. A. Berven, 1973. Sepcies diversity of herpetofauna samples from similar microhabitats at two tropical stations. Ecology 54: 642–645.

Hering, D., O. Moog, L. Sandin & P. F. M. Verdonschot, 2004. Overview and application of the AQEM assessment system. Hydrobiologia 516: 1–20.

Karr, J. R., K. D. Fausch, P. L. Angermeier, P. R. Yant & I. J. Schlosser, 1986. Assessing biological integrity in running waters: a method and its rationale. Illinois National History Survey, Champaign, Illinois, Special Publication 5.

Karr, J. R. & E. W. Chu, 1999. Restoring Life in Running Waters: Better Biological Monitoring. Island Press, Washington, DC.

Kerans, B. L., J. R. Karr & S. A. Ahlstedt, 1992. Aquatic invertebrate assemblages: spatial and temporal differences among sampling protocols. Journal of the North American Benthological Society 11: 377–390.

Lenat, D. R., 1988. Water quality assessment of streams using a qualitative collection method for benthic macroinvertebrates. Journal of the North American Benthological Society 7: 222–233.

Lorenz, A., L. Kirchner & D. Hering, 2004. 'Electronic subsampling' of macrobenthic samples: how many individuals are needed for a valid assessment result? Hydrobiologia 516: 299–312.

Major, E. B., M. T. Barbour, J. S. White & L. S. Houston, 1998. Development of a Biological Assessment Approach for Alaska Streams: A Pilot Study on the Kenai Peninsula. Environment and Natural Resources Institute, University of Alaska Anchorage, Anchorage, AK. Report for Alaska Department of Environmental Conservation, Anchorage, AK, 31 pp.

Maxted, J. R., M. T. Barbour, J. Gerritsen, V. Poretti, N. Primrose, A. Silvia, D. Penrose & R. Renfrow, 2000. Assessment framework for mid-Atlantic coastal plain streams using benthic macroinvertebrates. Journal of the North American Benthological Society 19: 128–144.

Metzeling, L. & J. Miller, 2001. Evaluation of sample size used for the rapid bioassessment of rivers using macroinvertebrates. Hydrobiologia 444: 159–170.

Needham, P. R. & R. L. Usinger, 1956. Variability in the macrofauna of a single riffle in Prosser Creek, California, as indicated by the Surber sampler. Hilgardia 24: 383–409.

Norris, R. H., E. P. McElravy & V. H. Resh, 1992. The sampling problem. In Calow, P. & G. E. Petts (eds), Rivers Handbook. Blackwell Scientific Publications, Oxford: 282–306.

Ofenböck, T., O. Moog, J. Gerritsen & M. Barbour, 2004. A stressor specific multimetric approach for monitoring running waters in Austria using benthic macro-invertebrates. Hydrobiologia 516: 251–268.

Ohio EPA (Enviromental Protection Agency), 1987. Biological Criteria for the Protection of Aquatic LifeI–IIIOhio EPA, Division of Water Quality Monitoring and Assessment, Surface Water Section, Columbus, Ohio.

Plafkin, J. L., M. T. Barbour, K. D. Porter, S. K. Gross & R. M. Hughes, 1989. Rapid Bioassessment Protocols for Use in Streams and Rivers: Benthic Macroinvertebrates and Fish. EPA/440/4–89/001. U. S. EPA Office of Water, Washington, DC.

Rolauffs, P., I. Stubauer, S. Zahrádlová, K. Brabec & O. Moog, 2004. Integration of the saprobic system into the European Union Water Framework Directive. Hydrobiologia 516: 285–298.

Sandin, L. & R. K. Johnson, 2000. The statistical power of selected indicator metrics using macroinvertebrates for

assessing acidification and eutrophication of running waters. Hydrobiologia 422/423: 233–243.

Schmedtje, U. & M. Colling, 1996. Ökologische Typisierung der aquatischen Makrofauna. Informationsberichte des Bayerischen Landesamtes für Wasserwirtschaft 4/96.

Skoulikidis Th., N., K. C. Gritzalis, T. Kouvarda & A. Buffagni, 2004. The development of an ecological quality assessment and classification system for Greek running waters based on benthic macroinvertebrates. Hydrobiologia 516: 149–160.

Somers, K. M., R. A. Reid & S. M. David, 1998. Rapid biological assessments: how many animals are enough? Journal of the North American Benthological Society 17: 348–358.

Šporka, F., H. E. Vlek, E. Bulánková & I. Krno, 2006. Influence of seasonal variation on bioassessment of streams using macroinvertebrates. Hydrobiologia 566: 543–555.

Vlek, H. E. (ed.), 2004. Comparison of (Cost) Effectiveness between Various Macroinvertebrate Field and Laboratory Protocols. European Commssion, STAR (Standardisation of River Classifications), Deliverable N1, 78 pp.

Vlek, H. E., P. F. M. Verdonschot & R. C. Nijboer, 2004. Towards a multimetric index for the assessment of Dutch streams using benthic macroinvertebrates. Hydrobiologia 516: 173–189.

Zelinka, M. & P. Marvan, 1961. Zur Präzisierung der biologischen Klassifikation der Reinheit fließender Gewässer. Archiv für Hydrobiology 57: 389–407.