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

Citation for published version (APA):
6 Quantifying spatial and temporal variability of macroinvertebrate metrics

6 Quantifying spatial and temporal variability of macroinvertebrate metrics

Hanneke E. Keizer-Vlek, Piet F.M. Verdonschot, Ralf C.M. Verdonschot & Paul W. Goedhart


Abstract

Since the introductions of the Habitat Directive and the European Water Framework Directive, water authorities are now obliged to monitor changes in conservation value/ecological quality on larger spatial scales (opposed to site scale), as well as 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, analyses of the statistical power of the monitoring programs should be implemented. Currently, the statistical properties associated with aquatic monitoring programs are often unknown. We collected macroinvertebrate samples from 25 meso-eutrophic drainage ditches in the Netherlands and selected 7 taxonomic richness metrics for the evaluation of spatial and temporal variability. Simulations were performed to investigate the effects of changes in (1) the total number of species included in a taxonomic richness metric and (2) the relative number of rare species included in a taxonomic richness metric. Of the 7 metrics evaluated, the number of common species required the smallest number of monitoring sites, followed by the number of Gastropoda species, and the number of species. Also, results showed that metric variability will decrease when the proportion of rare species included in a taxonomic richness metric is reduced or the total number of species included is increased. Irrespective of the metric applied a large effort will be required to detect change within drainage ditches in the Wieden, due to high spatial variability. Therefore, we need to explore the possibilities of applying alternative more cost-effective methods for sampling and sample processing in biomonitoring programs.

Keywords: spatial variability, temporal variability, rare species, macroinvertebrates, biomonitoring, detection of change
Introduction

The ecological quality of surface waters in the Netherlands has been monitored for several decades by regional water authorities. Until recently, they were focused on determining the ecological quality at a specific site and they collected a single sample at the site for this purpose, as is also common practice in the United States (Carter & Resh, 2001). Since the introductions of the Habitat Directive and the European Water Framework Directive, water authorities are now obliged to monitor changes in conservation value/ecological quality on larger spatial (regional) scales, as well as 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 new requirements, the process of designing monitoring programs and interpreting the data resulting from these programs should implement analyses of the statistical power of the programs.

Power analysis (assessing the ability of a program to accurately detect change) could help avoid unnecessary expenditures on monitoring programs that cannot provide meaningful results or that will lead to overspending. In the final step of testing a statistical hypothesis, a decision will be made about the validity of the null hypothesis. Two types of errors can be made in making this decision, a type I error or a type II error. A type I error can be described as “drawing the conclusion that change has occurred when in fact it has not”. Conversely, concluding that change has not occurred when in fact it has is called a type II error. Both errors can have large consequences. Type I errors could lead to serious negative financial effects if costly and unnecessary restoration measures are taken. Type II errors could have serious negative effects on reaching ecological goals if failure to detect a negative trend leads to the dismissal of required restoration measures (Taylor & Gerrodette, 1993).

The probability of making a type I error is usually denoted as \( \alpha \) (statistical significance), and the probability of making a type II error is usually denoted as \( \beta \). Power \((1 - \beta)\) is defined as the probability that change will be detected (Gerrodette, 1987). Statistical power depends on a number of factors: (1) statistical significance, (2) the magnitude of effect to be detected (i.e., effect size), (3) sample size and variability, and (4) statistical assumptions (e.g., use of one-tailed tests versus two-tailed tests). Currently, the statistical properties associated with aquatic monitoring programs are often unknown.

Changes in ecological quality can be the result of restoration measures or anthropogenic disturbance. However, such changes can be masked by several sources of variation, sampling effects, spatial variation, and temporal variation. To determine whether change is the result of anthropogenic
Quantifying spatial and temporal variability of macroinvertebrate metrics

disturbance requires the determination of natural variability (Johnson, 1998; Leunda et al., 2009; Resh & Rosenberg, 1989). Insight into both spatial and temporal variability is required. Most studies that have quantified temporal and/or spatial variability were focused on lotic ecosystems (e.g., Dolph et al., 2010; Downes et al., 1993; Gehler, 2004; Springe et al., 2006). The majority of surface waters in the Netherlands, however, are lentic ecosystems. Of these lentic ecosystems, drainage ditches are particularly interesting because they are important drivers of biodiversity in agricultural areas (Armitage et al., 2003; Herzon and Helenius, 2008; Painter, 1999). They are also a prominent feature in the landscape of the lowlands of northwestern Europe; in the Netherlands alone, total ditch length is approximately 300,000 km (Verdonschot et al., 2012).

There are several studies that have dealt with spatial or temporal variation of macroinvertebrate communities in lentic systems in relation to biological assessment (Hämäläinen et al. 2003; Kashian & Burton, 2000; Tangen et al., 2003). However, only a few have quantified both spatial and temporal variations with the purpose of defining statistical properties of future monitoring programs (Johnson, 1998; Trigal et al., 2006). The first objective of this study was therefore to quantify 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 disturbance and/or restoration measures.

The decision whether to include rare species in analysis for bioassessment purposes may affect statistical power (Cao et al., 2001). Many studies have addressed the use of rare species in biological assessment. Some advocate the use of rare species, because they may be good indicators of ecological quality (e.g., Lenat & Resh, 2001; Lyons et al., 1995; Nijboer & Schmidt-Kloiber, 2006; Poos & Jackson, 2012). Others, favor the exclusion of rare species because they add noise to the analysis (e.g., Gauch, 1982; Marchant, 2002), thus diminishing power. None of these studies, however, have looked at the effects of including/excluding rare species from metrics and the effect that this has on metric variability. The second objective of this study was therefore to determine the influence of rare species on variability of taxonomic richness metrics.
Methods

Study area

Macroinvertebrate samples were collected from 25 drainage ditches in the Netherlands. The Netherlands can be characterized as a mostly flat agricultural landscape. The ditches were located in the natural preserve the Wieden. The Wieden is a peatland covering about 100 km², of which a large part is open water. The Wieden can be characterized as a cultural landscape, which has been formed as a result of peat excavations in the past in combination with wind erosion and reed cutting. The area consists of fen-meadows, reed beds and quaking fens. Despite the artificial origin of the drainage ditches the influence of point and non-point sources on these ditches is minimal. Therefore, we considered the spatial variation in the Wieden as natural spatial variation. The drainage ditches in the Wieden are naturally meso-eutrophic (Table 6.1). The 25 sampled drainage ditches all belonged to the same watertype: buffered ditches in peatland areas with a maximum width of 8 m (Elbersen et al., 2003).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Median</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>conductivity (μS/cm)</td>
<td>382</td>
<td>156</td>
<td>519</td>
</tr>
<tr>
<td>pH</td>
<td>7.24</td>
<td>6.07</td>
<td>7.94</td>
</tr>
<tr>
<td>total nitrogen (mg/l)</td>
<td>1.35</td>
<td>0.68</td>
<td>3.35</td>
</tr>
<tr>
<td>total phosphorus (mg/l)</td>
<td>0.05</td>
<td>0.04</td>
<td>0.9</td>
</tr>
<tr>
<td>depth (cm)</td>
<td>60</td>
<td>33</td>
<td>116</td>
</tr>
<tr>
<td>width (m)</td>
<td>4</td>
<td>3.5</td>
<td>7</td>
</tr>
</tbody>
</table>

Sampling and laboratory processing

Macroinvertebrate samples were collected from the drainage ditches in the months May to June in 2006, 2007, and 2008. Using a D-frame dip net (25 cm × 25 cm, 500-μm mesh size), we collected a composite sample at each site from three habitats over set distances: the emergent vegetation (1.5 m), the submerged and floating vegetation (2 m), and the (otaherganic) bottom substrate (1.5 m). The composite samples were transferred to buckets and transported to the laboratory, where they were stored in a refrigerator and oxygenated. The samples were washed through 1000-μm and 250-μm sieves. Next, live organisms belonging to the groups Odonata, Gastropoda, Trichoptera, and Ephemeroptera were sorted from the samples by eye and...
preserved in 70% ethanol. Organisms were identified to the lowest taxonomic level possible, which was the species level for almost all specimens.

Data analysis

Rare species

Several criteria have been used to define rare species in ecological studies. In most benthic studies, species are considered rare when they occur at low abundance and/or have a small distribution range (Cao et al., 1998). Resh et al. (2005) used temporal occurrence to define rarity. In this study, we have defined rare species based on the frequency of collection (using a combination of spatial and temporal occurrence). A species was considered rare when it occurred in 5% or less of the 75 samples collected (Resh et al., 2005). A species was considered common when it occurred in 70% or more of the 75 samples collected.

Individual species

In total, 75 samples were collected from 25 sites during three consecutive years. The frequency of collection was calculated for each species by determining the proportion of samples from which the species was collected. The frequency of collection was divided into 10 distribution classes. For each distribution class, we calculated the proportion of species compared to the total number of species. This was done for each of the 4 different groups: Ephemeroptera, Gastropoda, Odonata, and Trichoptera. All taxa that could not be identified to species level were excluded from the analyses.

To examine abundance patterns of rare and common species, we determined average species density by dividing the summed density of all samples by the number of samples from which a species was collected. Log$_{10}$ transformed density (number of individuals/1.25 m$^2$) was plotted against the frequency of collection and this relationship was fitted with a linear regression.

The number of monitoring sites required to detect a change in the frequency of collection of an individual species between two points in time depends on (1) the probabilities of occurrence at the two time points, (2) the significance level $\alpha$ of the test, and (3) the required power ($1 - \beta$). The required number of monitoring sites was calculated using the improved approximate method for testing the equality of two binomial proportions (Casagrande et al., 1978). This method gives larger samples sizes than those based on the “arc-sin formula,” for example, as used by Cochran & Cox (1975). We used a two-sided
test and assumed that the number of monitoring sites at the two time-points were equal. Different levels of statistical significance were used ($\alpha = 0.05$, $\alpha = 0.1$, and $\alpha = 0.2$) and statistical power was set at 80% ($\beta = 0.2$). Calculations were performed for 2 effect sizes: 20% and 40% change.

**Taxonomic richness metrics**

Many multimetric indices that are currently used for biological assessment apply taxonomic richness metrics (e.g., Bloxsom et al., 2002; Dahl & Johnson, 2004; Menetrey et al., 2011; Purcell et al., 2009; Vlek et al., 2004). Several studies have showed that taxonomic richness metrics are far less variable than those based on density or biomass, and are thus more effective at detecting change (e.g., Johnson, 1998; Resh & McElravy, 1993; Smith et al., 2005; Springe et al., 2006; Vlek, 2004). Therefore, we selected 7 taxonomic richness metrics for the evaluation of spatial and temporal variability, including number of species, number of indicator species, number of ET (Ephemeroptera and Trichoptera) species, number of Trichoptera species, number of Gastropoda species, number of rare species, and number of common species. The number of indicator species was based on a list of indicator species that was developed especially for drainage ditches (Nijboer, 2000). The list contains a combination of species that should be present in drainage ditches of good ecological quality. The coefficient of variation (CV; standard deviation divided by the mean, reported as a percentage) was used as a measure of variability and was calculated based on the 25 samples collected in 2006, 2007, and 2008.

Three sources of variation can be distinguished: spatial variation ($\sigma_s^2$), temporal variation ($\sigma_t^2$), and remaining variation ($\sigma_r^2$); the last component is a combination of different sources of variability, for example, analytical variation, variation at lower temporal scales (e.g., within season), and variation at lower spatial scales (e.g., within site). We estimated $\sigma_s^2$, $\sigma_t^2$, and $\sigma_r^2$ using restricted maximum likelihood where each variance was held positive. The power of a statistical test to detect a change in a metric between two points in time depends on the variance of the difference between the averages at the two time points. When $n$ sites are monitored in one year and another $n$ sites are monitored in another year, this variance equals $2(\sigma_s^2 + \sigma_r^2)/n + 2\sigma_t^2$. The term $\sigma_s^2$ cancels when the same sites are used. Note that an increase in the number of sampled sites only reduces spatial and remaining variations, not temporal variation. The baseline variance $2\sigma_t^2$ implies that a change in the order of magnitude of 2 times the baseline standard deviation, i.e., $2\sqrt{2}(\sigma_t^2)$, will never be significant. Based on the estimates of the variance components, a power of 95%, and a significance level of 0.05, the number of monitoring sites
(n) required to detect a change of 25% was calculated by means of the non-central t distribution. The calculation is similar to the one used by Cochran & Cox (1975), except that we took into account the baseline variance $\sigma^2$. Note that the number of samples equals 2n, since n samples are taken at two points in time.

Simulations were performed to investigate the effects of changes in (1) the total number of species included in a taxonomic richness metric and (2) the relative number of rare species included in a taxonomic richness metric. To simulate taxonomic richness metrics with different numbers of species, random species lists were generated 50 times for several combinations of a certain number of species. Combinations ranged from 6 to 26 species in total with a given percentage of 50% rare and 50% common species. To simulate taxonomic richness metrics with different proportions of rare species, random species lists were generated based on the complete species list from the 75 samples. Lists of 13 species each with combinations ranging from 0 rare species and 13 common species to 13 rare species and 0 common species were randomly generated, 50 times for each combination. For each list, in both experiments, we summed the number of species collected from each sample, and then calculated the coefficient of variation for the number of species, based on the total of 75 samples.

Results

Individual species

Frequency of collection and abundance

During the three years of sampling, 3 Ephemeroptera species, 25 Gastropoda species, 18 Odonata species, and 28 Trichoptera species were collected. For all 4 macroinvertebrate groups, differences in the number of species collected were small between years, with a maximum difference of 2 species (Table 6.2). The number of different Gastropoda, Odonata, and Trichoptera species collected in total (during the 3 years) differed considerably from the numbers of species collected during each of the individual years (Table 6.2).

Frequency of collection was high (>0.55) for all Ephemeroptera species (Fig. 6.1). Comparatively, the frequency of collection was low for many Trichoptera and Odonata species; 36% and 44%, respectively, had frequencies of collection of 0.05 or less (found in ≤3 samples) (Fig. 6.1). The group Gastropoda was represented by species with both high and low frequencies of
collection (Fig. 6.1). In total, 28% of the species was collected at a frequency of 0.05 or less.

Increased density was correlated with higher frequency of collection (Fig. 6.2; p < 0.001). All 10 species with an average of 30 or more individuals per 1.25 m² (1.5 log₁₀-transformed) had a frequency of collection >0.69, with the exception of *Segmentina nitida*. *S. nitida* was the only species collected in high numbers (74 individuals/1.25 m²) with a relative low frequency of collection (0.25). *Brachytroon pratense* also stands out due to its low average density (1.8 individuals/1.25 m²) and relatively high frequency of collection (0.29) (Fig. 6.2).

**Table 6.2:** Overview of the number of species collected per macroinvertebrate group, from drainage ditches in the Wieden in 2006, 2007, 2008 (25 samples each), and all three years together (75 samples).

<table>
<thead>
<tr>
<th>Macroinvertebrate group</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>Total number of species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ephemeroptera</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Gastropoda</td>
<td>18</td>
<td>17</td>
<td>16</td>
<td>25</td>
</tr>
<tr>
<td>Odonata</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>18</td>
</tr>
<tr>
<td>Trichoptera</td>
<td>12</td>
<td>10</td>
<td>11</td>
<td>28</td>
</tr>
</tbody>
</table>

**Figure 6.1:** Frequency of collection distribution (the percentage of total species collected against the proportion of total samples in which these species occurred) for Ephemeroptera, Gastropoda, Odonata, and Trichoptera species collected from 25 drainage ditches in the Wieden in 2006, 2007, and 2008.
Detection of change

Monitoring is required to detect changes in the frequency of collection for individual species within a conservation area. In general, the number of monitoring sites required to detect change decreases with an increase in the frequency of collection. An increase in the level of significance (α) has relatively little effect on the required number of monitoring sites. However, the degree of change (effect size) has a large influence on the number of required monitoring sites (Appendix A, Fig. A.1). For example, to detect a 20% change for species with a collection frequency of 0.45, 506 monitored sites are required, whereas detection of a 40% change, only requires 131 sites (α = 0.05 and β = 0.2) (Appendix A, Fig. A.1 and A.2). To detect a 40% change, species with a frequency of collection ≤0.7 will require more than 50 monitoring sites (α = 0.05 and β = 0.2) (Appendix A, Fig. A.2).

Taxonomic richness metrics

Variability

Differences in average metric values between years were small, with a maximum difference of 2 species. For the number of Gastropoda species and...
number of common species, there were no differences in metric values between years (Table 6.3). Estimates of variance components showed no temporal variation for the following metrics: number of indicator species, number of Gastropoda species, and number of rare species (Table 6.4). Compared to spatial and remaining variation, temporal variation was negligible for all other metrics (Table 6.4).

Variation was the highest for the number of rare species (CV = 155%, 3-year average), followed by the number of Trichoptera species and the number of ET species. Variation was the lowest for the number of common species (CV = 17%, 3-year average) (Table 6.3). CVs differed between years for all metrics to varying degrees, with a maximum between-year difference of 50% for the number of rare species. On the other hand, spatial variation for the number of Trichoptera species showed only minimal differences between years (Table 6.3).

<table>
<thead>
<tr>
<th>Metric</th>
<th>Year</th>
<th>Average</th>
<th>CV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of species</td>
<td>2006</td>
<td>23</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>2007</td>
<td>23</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>2008</td>
<td>25</td>
<td>20</td>
</tr>
<tr>
<td>number of indicator species</td>
<td>2006</td>
<td>9</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>2007</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>2008</td>
<td>9</td>
<td>24</td>
</tr>
<tr>
<td>number of Ephemeroptera and Trichoptera</td>
<td>2006</td>
<td>9</td>
<td>40</td>
</tr>
<tr>
<td>species</td>
<td>2007</td>
<td>8</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>2008</td>
<td>9</td>
<td>35</td>
</tr>
<tr>
<td>number of Trichoptera species</td>
<td>2006</td>
<td>7</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>2007</td>
<td>6</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>2008</td>
<td>7</td>
<td>42</td>
</tr>
<tr>
<td>number of Gastropoda species</td>
<td>2006</td>
<td>12</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>2007</td>
<td>12</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>2008</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td>number of rare species</td>
<td>2006</td>
<td>2</td>
<td>163</td>
</tr>
<tr>
<td></td>
<td>2007</td>
<td>1</td>
<td>177</td>
</tr>
<tr>
<td></td>
<td>2008</td>
<td>1</td>
<td>127</td>
</tr>
</tbody>
</table>
Table 6.4: Estimates of spatial, temporal, and remaining variance components for the 7 selected metrics. The last column indicates the number of monitoring sites required (n) to detect a 25% change (effect size) in average metric values between two points in time, calculated according to Cochran and Cox (1957) (α = 0.05, β = 0.05).

<table>
<thead>
<tr>
<th>Metric</th>
<th>Spatial</th>
<th>Temporal</th>
<th>Remaining</th>
<th>Number of sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of species</td>
<td>15.8</td>
<td>0.9</td>
<td>12.7</td>
<td>23</td>
</tr>
<tr>
<td>number of indicator species</td>
<td>2.8</td>
<td>--</td>
<td>2.8</td>
<td>33</td>
</tr>
<tr>
<td>number of Ephemeroptera and Trichoptera</td>
<td>7.3</td>
<td>0.4</td>
<td>3.6</td>
<td>62</td>
</tr>
<tr>
<td>species</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>number of Trichoptera species</td>
<td>4.5</td>
<td>0.4</td>
<td>3.0</td>
<td>76</td>
</tr>
<tr>
<td>number of Gastropoda species</td>
<td>3.0</td>
<td>--</td>
<td>3.6</td>
<td>21</td>
</tr>
<tr>
<td>number of rare species</td>
<td>0.01</td>
<td>--</td>
<td>0.8</td>
<td>1017</td>
</tr>
<tr>
<td>number of common species</td>
<td>2.1</td>
<td>0.1</td>
<td>1.3</td>
<td>13</td>
</tr>
</tbody>
</table>

Detection of change

Based on the variation in the dataset we estimated spatial, temporal, and remaining sources of variation and determined that a 25% change in the average number of indicator species between 2 points in time could be detected with 33 monitoring sites (Table 6.4; α = 0.05 and β = 0.05). To detect the same change in the total number of species, only 23 monitoring sites were required. The smallest number of sites (13) was required to detect change in the number of common species (Table 6.4).

Variability and rare species

Variation decreased with a decrease in the number of species incorporated into the “simulated” taxonomic richness metric (Fig. 6.3). However, depending on the species that were randomly selected to construct the metric, CV varied considerably within each number of species incorporated in the metric (illustrated by wide error bars in Fig. 6.3).

Variation also decreased with a decrease in the proportion of rare species that were incorporated in the ‘simulated’ taxonomic richness metric (Fig. 6.4). The average CV was 17% for a metric that consisted of only common species. This 17% gradually increased to 23% as the metric was gradually adjusted to consist of 8 rare and 5 common species. Inclusion of 9 or more rare species led to a considerable increase in average CV and a metric that consisted of 13 rare species gave an average CV of 188% (Fig. 6.4).
Figure 6.3: Relationship between the number of macroinvertebrate species included in a taxonomic richness metric and the coefficient of variation. For each number of species, different combinations of species were randomly reordered 50 times. Each combination consisted of 50% common and 50% rare species. Variation in the total number of species (per sample) was calculated based on the 75 samples collected at 25 sites in the Wieden. Squares represent average CV and error bars represent ± standard deviation in CV’s.

Figure 6.4: Relationship between the proportion of rare species included in a richness metric and the coefficient of variation. For each proportion of rare species, different combinations of species were randomly reordered 50 times. Given a random list of 13 species, spatial variation in the total number of species (per sample) was calculated based on the 75 samples collected at 25 sites in the Wieden. Squares represent average CV and error bars represent ± standard deviation in CV’s.
Discussion

Individual species

Our study showed that 28% of the species collected from drainage ditches in the Wieden were rare. A 20-y study by Resh et al. (2005) reported similar percentages of 20–30% rare taxa from a Californian stream. Unlike Resh et al. (2005), we did not differentiate between species that were spatially rare and those that were temporarily rare, because temporal variability will affect spatial variability and vice versa.

We observed differences in the frequency of collection distribution between the three different macroinvertebrate groups (Ephemeroptera are not considered here, because only three species were collected). The Trichoptera and Odonata were characterized by a relative high number of species with a frequency of collection \( \leq 0.15 \), while Gastropoda exhibited a relatively even distribution over the different frequency classes. These differences in the frequency of collection distributions might result from differences in the relationship between rarity and density. In accordance with Resh et al. (2005) we found a significant correlation between increased density and increased frequency of collection \( (R^2 = 0.68) \). A positive relationship between the density of a species and extent of its spatial distribution is also commonly observed in terrestrial ecosystems (Gaston, 1996). Studying the relationship between rarity and density for the individual groups, \( R^2 \) was 0.48 for the Gastropoda, and 0.75 for the Trichoptera, indicating that the relationship between rarity and density varies between macroinvertebrate groups. The Odonata were not considered, because only 2 species were collected with frequencies higher than 0.3.

Large numbers of sites must be monitored to detect changes in the frequency of collection of individual macroinvertebrate species due to restoration measures or anthropogenic disturbance, especially in the case of rare species. To detect a 40% change \( (\alpha = 0.05 \text{ and } \beta = 0.2) \) in the frequency of collection, more than 1000 sites must be sampled to monitor rare species (frequency of collection < 0.05), while common species with a frequency of collection \( \geq 0.7 \) will require less than 50 monitoring sites. Unfortunately, conservation managers are mostly interested in the rare species. It is unlikely that more than 50 sites will be monitored at a regional scale just to detect changes in drainage ditches (or any other water type for this matter); however, for monitoring at the national scale, much higher numbers of monitoring sites might be acceptable. One should also keep in mind that spatial variance in this study was based on a relatively small region. At the national scale, spatial...
variation will be higher and the frequency of collection of the individual species will probably be far lower (than at the regional scale), resulting in more sites for monitoring to achieve equal power.

The results from this study clearly indicate that in general it will be easier to detect change based on metrics than on individual species (Fig. A.1 and Table 6.4). As already stated by Maxwell and Jennings (2005), composite indicators (composed of several species) have the disadvantage that positive trends in some species can mask negative trends in other species. This means the extinction of individual species could take place without being noticed, which might be judged to be unacceptable by conservation managers. Water managers, on the other hand, are generally more interested in changes in the ecological status of macroinvertebrate communities than they are interested in the changes in 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 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).

**Taxonomic richness metrics**

**Spatial and temporal variation**

Both spatial and temporal variations can vary at different scales. A wide variety of studies have examined variation at different spatial and temporal scales, i.e., among-season, among-year, within-site, within-reach, and among-streams/lakes (e.g., Gebler, 2004; Sandin & Johnson, 2000; Springe et al., 2006; Trigal et al., 2006). Apart from spatial and temporal differences, variation can also result from analytical error. We did not explicitly examine each of these different scales, or the variation due to analytical error. This study was merely meant to gain insight into the sampling effort required to detect changes in ecological quality within a system of drainage ditches between years. For this purpose, it was not necessary to tease out variation at different scales. We combined analytical error and variation at lower temporal scales (i.e., within-season) and spatial scales (i.e., within-site variability) into one overall term (sample variation). Information about the different sources of variation can be very valuable when you want to increase statistical power by making changes to your sampling protocol and/or sampling design.
To reduce variability, sampling was stratified in time, i.e., all samples were taken in May/June, and one operator collected all samples. When using the results of this study to design a monitoring program, it should be kept in mind that variability will increase if (1) more than one operator collects samples and/or (2) sample collection is not limited to one season. Studies by Trigal et al. (2006) and Clarke et al. (2002) indicate the extents to which these sources can contribute to overall variation. Trigal et al. (2006) showed seasonal variation of 32% for sweep-net samples collected in June, July, February, and May in a Mediterranean shallow lake. Clarke et al. (2002) estimated that less than 12% of sampling variation was due to inter-operator sampling effects using trained staff.

The magnitude of spatial variation in this study varied considerably between metrics. Although this is in line with the findings of many others (e.g., Gebler, 2004; Johnson, 1998; Trigal et al., 2006), we encountered some difficulties when we tried to compare coefficients of variation for the number of species (taxon richness) with those reported by others. Different studies covered different temporal and spatial scales, different water types, different habitats, and different sampling and sample processing protocols. Caution should be taken in making comparisons between studies, because variability in metric values can differ depending on the water type (Clarke et al., 2006) and even among different water bodies of the same water type (Porst & Irvine, 2009); the magnitude of metric variability also varies between sampling protocols (Vlek, 2004). This also makes it crucial to define variability at the scale appropriate for the aim of your study when developing a monitoring scheme.

In this study, (among-year) temporal variation appeared to be negligible for all 7 metrics, especially compared to spatial variation and remaining sources of variation. This is not in line with findings of Johnson (1998), who showed that among-year variability in total taxon richness was higher than among-sample and among-lake variability for littoral habitats in 16 Swedish lakes. In our study, temporal variation was calculated based on only three collections (2006, 2007, and 2008) and thus might have been underestimated; this needs to be studied further in the near future. These first results, however, suggest that temporal variation will hardly influence the monitoring effort required to detect change.

Our results showed that very large differences in metric values can be observed within a relatively small region, e.g., CV 38% for the number of ET species. This implies that making inferences at a higher spatial scale based on one site might lead to completely erroneous conclusions. This is in line with
the findings of Downes et al. (2000) and Gebler (2004), who each concluded that individual sites cannot be representative of larger stream sections.

(Reduction of) sampling effort

Of the 7 metrics evaluated, the number of common species required the smallest number of monitoring sites to detect change in the Wieden, followed by the number of Gastropoda species and the number of species. The largest number of monitoring sites was required for the number of rare species, with 1017 sites needed to detect a 25% change.

Results showed that an increase in the number of species included in a taxonomic richness metric will reduce variation in metric values, and thus decrease the number of monitoring sites required to detect change. Including more species in a metric reduces the chance of high variability, due to the fact that the absence of one species may be compensated by the presence of another. It would be advantageous to increase the number of species included in a metric to reduce spatial variability and increase statistical power of a given number of monitoring sites. However, Fore et al. (1996) suggested that in some cases, signal may be lost in the noise, i.e., a strong response by a few taxa can be missed because macroinvertebrate communities are usually dominated by taxa that are neither sensitive nor insensitive to human impact. Both statistics and ecological relevance should be balanced in developing or selecting metrics for the monitoring of changes in ecological quality. Increasing the number species included in an index may have statistical advantages, but can also make it more difficult to detect a relevant ecological signal. In this study we did not consider the sensitivity of metrics to anthropogenic disturbance. To develop a reliable assessment system the sensitivity of the metrics applied in this study needs to be determined. For this purpose, information on metric values at sites of different ecological qualities is required. Studies by Vlek et al. (2004) and Verdonschot et al. (2012) are examples of methods that can be applied to select metrics that are sensitive to anthropogenic disturbance.

A negative trend was observed between the relative number of rare species included in a taxonomic richness metric and the variability of metric values. This means that, from the point of statistical power, the inclusion of rare species in richness metrics should be restricted. Före et al. (1996) stressed that, "excluding rare taxa for statistical purposes only is contradictory to biological common sense". We agree with Cao et al. (2001) that statistics should be used to look for important ecological signals and that these might not be the strongest statistical signals. On the other hand, what is the point of monitoring if spatial variability is so high it becomes impossible to detect any
Quantifying spatial and temporal variability of macroinvertebrate metrics

Again, we want to stress the importance of weighing statistics and ecological relevance and the necessity to determine the sensitivity to anthropogenic disturbance of the metrics applied in this study.

Despite taking into account the number of species and the relative number of rare species in constructing a richness metric, a large effort will be required to detect change within a region. For example, we calculated that 13 sites must be monitored (at 2 points in time) to detect a 25% change in the number of common species in the Wieden ($\alpha = 0.05$, $\beta = 0.05$). Such a sampling effort would be considered too costly by water authorities in the Netherlands. Johnson (1998) also concluded that sample sizes required to document changes between sites or years were so high that they are seldom used in field assessments of environmental impact. He suggested increasing statistical power by stratifying sampling in space and time (Johnson, 1998). However, in our study, sampling was already stratified in time and stratifying in space within ditches is almost impossible because they are much smaller than lakes and it is difficult to discern between habitats. Another option is to increase statistical power by increasing $\alpha$ and $\beta$. The question at what level statistical significance and power should be set has been dealt with by numerous authors (e.g., Field et al., 2007; Mapstone 1995). The traditionally applied 5%-level of statistical significance resulting from adherence to the “five-eighty” convention (Di Stefano, 2003) places the “burden of proof” with those trying to prove environmental change due to human impact. A commonly voiced opinion is that this task should be shifted towards those who are trying to prove that no environmental change has taken place (e.g., Dayton, 2001; Field et al., 2004; Gray 1990). To balance the burden of proof, Field et al. (2004) derived a cost function approach that minimizes the total costs of both type I and type II errors. However, such a cost function requires information on the costs of type I and type II errors. In the case of macroinvertebrates, no information is available on the costs of type II errors, which are difficult to determine because macroinvertebrates do not have a direct economic function/value, e.g., like fish that serve as a food source or coral reefs that attract tourist. Maxwell & Jennings (2005) considered the intrinsic value of macroinvertebrate species to be higher than the costs associated with unnecessary management actions, and therefore relaxed $\alpha$ from the traditional 0.05–0.2. The levels of statistical significance and power applied in this study are intended only as examples. Prior to developing monitoring schemes, appropriate levels of significance and power should be discussed with all stakeholders. Important in this discussion is the realization that by increasing $\alpha$ and $\beta$, error rates will increase up to a point where one might
question the purpose of monitoring, i.e., the costs of wrong decisions can become far higher than the costs of monitoring.

In cases where statistical power can not be increased through stratifying of sampling and/or increasing $\alpha$ and/or $\beta$, the only option is to develop more cost-effective methods for sampling and sample processing. For example, Verdonschot (2010) has applied activity traps in drainage ditches, which saved 65% time compared to sweep net sampling (R. C. M. Verdonschot, Alterra, Wageningen University and Research Centre, personal communication.). Another more cost-effective method would be to target specific organism groups (i.e., Trichoptera). We need to explore the possibilities of applying alternative more cost-effective methods for sampling and sample processing in biomonitoring programs.

Since the introductions of the Habitat Directive and the European Water Framework Directive, water authorities are now obliged to monitor changes in conservation value/ecological quality on larger spatial (regional) scales. Therefore, it is remarkable that the issue of probability sampling in aquatic monitoring programs has not received many attentions in Europe. Probability sampling is well suited to eliminate selection bias since, by construction, every site has a known nonzero probability of being selected (Cochran, 1977). In Europe the selection of sample sites by water authorities is often based on their assumed representativeness, or practical matters like accessibility. This manner of site selection is called non-probability sampling. The problem with non-probability sampling is that statistically based inferences about trends at higher/larger spatial scales cannot be made (Edwards, 1998; Stoddard et al., 1998; Parr et al., 2002). To our knowledge EMAP (Environmental Monitoring and Assessment Program), developed in the United States, is the first and only attempt to use probability sampling for the purpose of site selection in the design of aquatic monitoring programs. The use of probability sampling in aquatic monitoring programs should also be considered in Europe.

Conclusions

This study shows that, large numbers of sites must be monitored to detect changes in the frequency of collection of individual macroinvertebrate species, due to restoration measures or anthropogenic disturbance, especially in the case of rare species and rare species based metrics. Unfortunately, conservation managers are most interested in these rare species. The required monitoring effort automatically implies, that data collected by water authorities in biomonitoring programs developed to meet the requirements of the European
Water Framework Directive, will not meet the requirements of conservation managers. When interested in an individual species, sampling methods will have to be adjusted to this specific species to increase the frequency of collection.

The results from this study clearly indicate that in general it will be easier to detect change in a drainage ditch network based on metrics than on individual species. Of the 7 metrics evaluated in this study, the number of common species required the smallest number of monitoring sites, followed by the number of Gastropoda species, and the number of species. Also, results showed that metric variability will decrease when the proportion of rare species included in a taxonomic richness metric is reduced or the total number of species included is increased. Irrespective of the metric applied a large effort will still be required to detect change within the drainage ditch network of the Wieden, due to high spatial variability. Therefore, we need to explore the possibilities of applying alternative more cost-effective methods for sampling and sample processing in biomonitoring programs.

Acknowledgements

This study was funded by the Ministry of Economic Affairs, Agriculture and Innovation (Ministry of Agriculture, Nature, and Food Quality), Contract no. BO-02-007-002. We would like to thank Martin van den Hoorn for his efforts in collecting the data on which this study was based.

References


Appendix A

Figure A.1: Theoretical relationship between the frequency of collection and the number of sites required to detect a 20% change in the proportion of sites with observations of a species given three different levels of \( \alpha \) and \( \beta = 0.2 \).

Figure A.2: Theoretical relationship between the frequency of collection and the number of sites required to detect a 40% change in the proportion of sites with observations of a species given three different levels of \( \alpha \) and \( \beta = 0.2 \).