Hydrobiologia (2006) 566:433–439
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/s10750-006-0079-2

Errors and uncertainty in bioassessment methods – major results and conclusions from the STAR project and their application using STARBUGS

Ralph T. Clarke[1,]* & Daniel Hering[2]
[1]*Centre for Ecology and Hydrology, Winfrith Technology Centre, Winfrith Newburgh, DT2 8ZD Dorchester, Dorset, UK*
[2]*Department of Hydrobiology, University of Duisburg-Essen, D-45117 Essen, Germany*
*(*Author for correspondence: E-mail: rtc@ceh.ac.uk)*

Key words: replicate, sampling variation, sub-sampling, macroinvertebrates, multi-metric, uncertainty, STARBUGS, simulations, software, audit, quality assurance, sample coherence

Abstract

The STAR project's extensive replicated sampling programmes have provided the first ever quantitative comparative studies of the susceptibility of a wide range of national macroinvertebrate sampling methods and taxonomic metrics to uncertainty resulting from the effects of field sampling variability and subsequent sub-sampling and laboratory (or bank-side) procedures and protocols. We summarise six STAR project papers examining various aspects of the potential sources of uncertainty in the observed fauna and observed metric values. The use of new simulation software STARBUGS (STAR Bioassessment Uncertainty Software System) to incorporate the effects of these potential errors into quantitative assessments of the uncertainty in assigning water bodies to WFD ecological status classes is discussed.

Introduction

Any indices of freshwater biological quality are of little value without some knowledge and quantitative estimates of their precision and of the confidence in assigning individual water bodies (river sites or lakes) to ecological status classes. This is a requirement of the Water Framework Directive (WFD), which states that 'Estimates of the confidence and precision attained by the monitoring system used shall be stated in the river basin monitoring plan' (European Commission, 2000). In particular, given the importance being placed in the WFD on determining whether a water body is in 'good' or better status class, we would like to be able to estimate the probability that a water body could actually be of 'moderate' or worse status.

The WFD requires each Member State to express bioassessment results as Ecological Quality Ratios (EQRs), where the ratios represent the relationship between the values of the biological parameters observed for a water body and the values for these parameters in the reference conditions applicable to that water body. The directive also requires each country to use these EQRs to classify water bodies into five ecological status classes and to monitor any changes in the status of water bodies (European Commission, 2000). When the ecological condition of a river site is assessed in two different years, the observed estimates of site quality will usually differ and the ecological status class may also have changed. We need to be able to place some confidence on the likelihood that a real change in quality or change in status class has occurred or whether the observed changes are just due to the inherent errors and sampling variation in the whole site assessment process.

As a general guide to the likely levels of uncertainty in assignment of sites to WFD ecological status classes, Figure 1 and Table 1 show the probability of misclassifying a site of any particular true quality (i.e., EQR value) according to the size of the errors or uncertainty in the

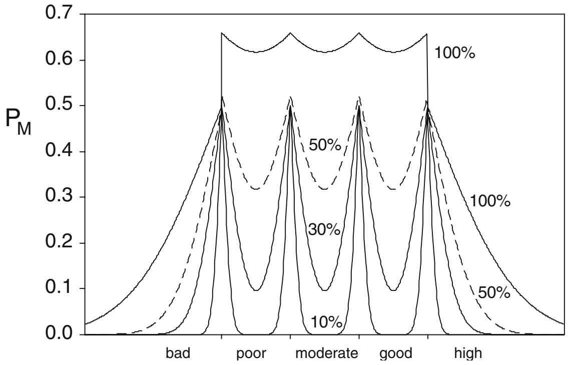

Figure 1. Plot of the probability (P_M) of classifying a site into a different status class versus its true Environmental Quality Ratio (EQR) value for a range of error/uncertainty standard deviations (σ) in the observed sample EQR value. The EQR range has been divided into the five WFD classes (high, good, moderate, poor and bad) with the middle three classes each of width W. Plots are shown for $\sigma = 10$, 30, 50 and 100% of W, where the broken line indicates the 50% plot.

Table 1. Mean and range of misclassification rates (P_M) for sites with true qualities in a middle (i.e., not top or bottom) ecological status class for each of a range of error/uncertainty standard deviations (σ) in their observed sample EQR values, where σ is expressed as a percentage of the EQR range of each middle class

σ (%)	Mean % misclassification (%)	Range (%)
10	8	0–50
30	24	10–50
50	39	32–52
100	63	62–66

metric's EQR values expressed as a percentage of the width of the status classes for the EQR (see Clarke et al., 1996 for the mathematical derivation). When the uncertainty standard deviation of the EQR values is only 10% of the width of status class, then sites whose true quality lies in the centre of a status class would never be misclassified. Sites whose true quality lies on the border of two classes will always have at least a 50% chance of being placed in the wrong class. With uncertainty standard deviations of 10% of class width, the overall misclassification rate for sites in a middle class (i.e., 'good', 'moderate' or 'poor') (assuming an even spread of true qualities across the class) is only 8% (Table 1). If however, the error standard deviation is 50% of the class width, then even sites in the centre of a middle class have a roughly one in three chance of being placed in the wrong class and roughly 40% of all sites in the class will be misplaced into either a higher or lower class. If the error standard deviation is equal to the class width (i.e., 100%), as if possible for metrics with high sampling variability, then all sites whose true quality lies within a middle class will more likely than not be placed in the wrong class (Fig. 1 and Table 1). Sites with EQR values either well above the high/good boundary or well below the poor/bad boundary will obviously have the lowest probabilities of being misclassified.

Sources of uncertainty in the observed biota

The sources of variation in the fauna observed at a site are due to:

(i) *Sampling variation and sampling method.* Within each site there will still be spatial heterogeneity in the microhabitats and distributions of macroinvertebrates and other organisms. Thus taxonomic richness and composition will vary between samples taken during the same period. The precision of a sampling method will therefore be influenced by the number of sampling units, the range of habitats and/or the total area sampled at a site.

(ii) *Sample processing and taxonomic identification errors.* Sub-sampling in the laboratory (or in the field) will also lead to increased uncertainty. In sorting the material in a test site's sample and identifying the taxa, some taxa may be missed or misidentified by less experienced staff. This may lead to biases and under-estimation of any index involving some from of taxonomic richness.

(iii) *Natural temporal variation.* There will also be what might be called 'natural' temporal variation whereby the taxa present at a site, not just in the sample, will vary 'naturally' over time for reasons other than stress or pollution.

(iv) *The effects of pollution or environmental stress on the biota.* This is what we are trying primarily to detect and quantify.

It may be difficult to distinguish between (iii) and (iv). For example, biological effects of a reduction in river discharge due to the weather may be considered natural, but reductions in river flow when abstraction is present may be considered a man-induced stress.

The potential sources of error in estimates of the expected fauna and Reference Condition (RC) for a site include having an inadequate set of reference sites, the choice of statistical prediction method or modelling technique, not involving all relevant environmental predictor variables and errors in measuring these variables for new sites (for further details see Clarke et al., 1996). For example, the WFD permits the determination of RC for a site from the average biota of the reference sites in the same stream type, where WFD System A types are based on only 3–4 classes of altitude, catchment area and geology. System B types and site-specific predictive models such as RIVPACS (Clarke et al., 2003) which use more site variables, might be expected to give more precise target RC, as recently shown for RIVPACS-type models in the UK, Sweden and the Czech Republic (Davy-Bowker et al., 2006).

The STAR project's extensive replicated sampling programme and the subsequent analysis of results has provided the first ever quantitative comparative study of the susceptibility of each of a wide range of established and 'national' macroinvertebrate sampling methods and a wide range of metrics to uncertainty resulting from the effects of field sampling method variability and subsequent subsampling and laboratory (or bank-side) procedures and protocols (Furse et al., 2006). We provide an integrated summary of six STAR project papers examining various aspects of the potential sources of uncertainty in the observed fauna and observed metric values.

Sampling method and sample size

Most commonly used macroinvertebrate sampling methods for rivers involving sampling each of the major habitats at a site and combining these basic sampling units into one overall composite sample for the site. Usually only one composite sample is obtained and thus there is no replication. Vlek et al. (2006) examined the effect of varying the number of sampling units involved in the composite sample on the precision of six commonly used macroinvertebrate metrics. They took repeated random subsets of 20 sampling units (each 25 cm sampling length using a 25 cm wide pond-net) from the dominant habitat type at each of four sites in the Netherlands and from each of two different habitats in each two streams in Slovakia. Although, as expected, the precision of all metrics increased with sample size (i.e., number units), the typical number of sampling of units required to achieve a 10% coefficient of variation (CV) for the composite sample varied from 1–2 (e.g., Saprobic index), to 3–8 (ASPT and 'Number of taxa'), while '% EPT-taxa' and 'total number of individuals' often required 10–17 sampling units. Accuracy was measured by treating the metric values based on all 20 sampling units combined as the 'truth', this was not ideal as it forces the any systematic bias to decrease as sample size approaches the maximum 20 units. The two most precise metrics also showed no systematic bias or trends with increasing sample size. However, ASPT values tended to under-estimate the 'true' ASPT when based on very few sampling units (<4–10) (Supplementary material in Vlek et al., 2006). This example reminds us that the sampling methods and protocols used to estimate the observed values of metrics should be exactly the same as those used at the reference sites involved in setting the target RC values; otherwise there may be systematic biases in the EQR values.

Vlek et al. (2006) found that, for a fixed sample size, precision was fairly similar across most habitat types for most metrics. However, there were exceptions, especially for '% EPT-taxa', which suggests caution in extrapolating estimates of sampling precision from one habitat or stream type to another.

Sample processing time was also found to increase linearly with sample size. Although the number of sampling units needed to achieve a target precision for a particular metric was similar for many stream types and metrics, the costs in terms of sample processing time for a given sample size varied significantly between habitats (Vlek et al., 2006). Samples from habitats, which had the most individuals per sample (and often the most taxa) (e.g., Fine particulate organic matter

(FPOM) riverine habitats in the Netherlands) tended to take longer to process, as might be expected with a method, which identifies all of the individuals in a sample.

Sampling variation

The STAR project involved the first ever-extensive replicated sampling programme to estimate and compare the overall effects of sampling variation on a wide range of 27 commonly used metrics for nine macroinvertebrate sampling methods across Europe. Replicate samples were taken in each of two seasons at a subset of 2–6 sites of varying pre-classified ecological status within each of 18 stream types spread over 12 countries, using both the STAR-AQEM method and a national sampling method or, where unavailable, the RIVPACS sampling protocol.

Clarke et al. (2006a) analysed these data to provide the first comparative estimates of the susceptibility to sampling variability of a range of macroinvertebrate sampling methods and metrics, including the six metrics involved in the proposed Inter-calibration Common Metric multi-metric index (ICMi, Buffagni et al., 2006). Clarke and colleagues determined the transformation scale for each metric, which made the replicate sampling standard deviation (SD) the most homogeneous, enabling a single best estimate of sampling SD of a metric to be determined for any particular method and stream type. These estimates can be then used to simulate the likely uncertainty in metric values associated with any other single sample taken from the same stream type using the same method (Table 2); as incorporated in the STAR project's STARBUGS software (see below).

Clarke et al. (2006a) estimated the precision of the combination of method and metric by expressing the replicate sampling variance as a percentage P_{samp} of the total variance in metric values with a stream type. High percentages indicate low sampling precision and low repeatability and hence that such a combination of sampling method and metric is unlikely to have much power to detect differences in ecological status class. The national methods used in the Czech Republic, Denmark, France, Poland and the RIVPACS method used in the UK and Austria all had percentage sampling variances <10% for most metrics. Because two methods were used on the same set of sites within a stream type, the P_{samp} values provided a valid comparison of their relative sampling precision. Most national methods, including RIVPACS, had sampling precisions at least as good as those for the STAR-AQEM method. In contrast, none of the metrics had percentage sampling variances <10% when based on either the Italian (IBE) method, which used bank-side sorting, or the Latvian national method, which identified only a limited set of taxa. When averaged over all stream types and methods, the three Saprobic metrics had the lowest average percentage sampling variances (3–6%). Obviously, metrics with high sampling precision and repeatability may still not be good ecological metrics or accurate indicators of ecological status class.

Lorenz & Clarke (2006) assessed the taxonomic community similarity of all pairs of samples taken within a stream type. They introduced the new concept of sample 'coherence' as a measure of the relative strength of within-site, within-season and within-method similarity. Site-coherence (i.e., the percentage of samples which are most similar to another sample from the same site) amongst sites

Table 2. Mathematical procedure used to simulate random sampling values of metrics with sampling SD (σ), which are constant on a particular transformation scale. X denotes the user-supplied untransformed observed value for a site. Z denotes a random standard normal deviate with a mean of zero and SD of σ

Transformation	Mathematical notation	Simulated value of metric in untransformed units
None	x	$X+Z$
Square root	$\surd x$	$(\surd X+Z)^2$
Double square root	$\surd\surd x$	$(\surd\surd X+Z)^4$
Arcsine square root for proportions	arcsine($\surd x$)	sine(arcsine($\surd X$) + Z))^2
Arcsine square root for percentages	arcsine($\surd(x/100)$)	sine(arcsine($\surd(X/100)$) + Z))^2

with replicate samples varied between 83% and 100%. Season-coherence of samples was nearly 100% even if different sampling methods were compared; indicating that time of year has a major influence on in-stream fauna. The STAR-AQEM method is most comparable in relative community similarity to the Nordic, Portuguese and Czech (PERLA) national methods and less comparable to the Italian (IBE) and Latvian methods. Samples collected by these latter methods had higher similarities to other sites sampled with the same methods than to samples from the same site obtained using the STAR-AQEM method, thus there was low site-coherence. Lorenz & Clarke (2006) found that replicate samples are less coherent within site, within season or within sampling method if the taxonomic resolution is family rather than species.

Sample processing and taxonomic identification errors

Having obtained a sample in the field, the procedures used to process the sample can all influence the overall reliability of the recorded taxonomic information. For example, the STAR-AQEM method requires the sub-sampling and taxonomic identification of at least one-sixth of the sample and at least 700 individuals. To assess the effect of this on the precision of results, replicate STAR-AQEM sub-samples were taken at most of the STAR sites where replicate samples were taken. Clarke et al. (2006b) found that STAR-AQEM sub-sampling effects caused more than 50% of the overall variance between replicate samples values for 12 of the 27 macroinvertebrate metrics analysed and was generally greatest for metrics that depend on the number of taxa present.

Sorting and identifying a larger fraction of the sample would reduce this source of variation (in the extreme, sorting the whole sample would eliminate it); but at increased costs. Vlek (2004) found that, on average across the sampled sites, STAR-AQEM samples took 18 h to process (including sorting and identification, whilst RIVPACS samples took only 9 h – half the amount of time. As the RIVPACS method led to no more than marginally higher average percentage sampling variances within the four countries where both methods were used, the RIVPACS method may be more cost-effective than the STAR-AQEM method.

Since the early 1990s, the UK government's environment agencies have used internal quality assurance and external auditing schemes to monitor the quality of their processing and taxonomic identification of RIVPACS macroinvertebrate samples (Dines & Murray-Bligh, 2000). Using this experience, a sample-auditing scheme involving 10 countries was implemented within the STAR project to assess the joint impact of sorting and identification errors of macroinvertebrate samples collected and analysed using different methods (notably STAR-AQEM and RIVPACS) (Haase et al., 2006). Haase and colleagues analysed differences in terms of 'gains' and 'losses' of taxa between the original and audited recorded lists of taxa for a sample. They found a surprising degree of sorting and identification errors, the total impact of which was reflected in many functional metrics and in metrics indicative of taxonomic richness. The results stress the importance of implementing quality control mechanisms in macroinvertebrate assessment schemes to monitor, improve and maintain sample-processing performance.

Natural temporal variation

Within STAR, Šporka et al. (2006) made an assessment of the effect of natural temporal seasonal variability on macroinvertebrate community composition and metric values. They took replicate multi-habitat samples at two-monthly intervals for a year from two stretches of a calcareous stream in the Carpathian Mountains and found major seasonal distinct differences in community composition. Moreover, seasonal differences were detected for many metrics, often related to the amount of organic material present. This study re-enforces the problem of deciding when to sample a stream for biomonitoring. Ignoring natural seasonal variability can confound the detection of anthropogenic environmental change. In the context of the WFD, it is important that the RC value of one or more metrics for a water body are not only appropriate for that type of site, but are determined from samples taken at roughly the same type of year as the sampling season(s) used in the monitoring programme. Sampling in more than one season (e.g., spring and autumn) and perhaps combining the samples (to determine both observed and RC metric values) can lead to more reliable estimates of ecological status, as shown by Clarke et al. (2002).

Implications for uncertainty in ecological status assessments and STARBUGS

As part of the STAR project, a new simulation software package called STARBUGS (STAR Bioassessment Uncertainty Guidance Software, Clarke, 2004) has been produced to help assess the effect of the various sources of variation and errors in the observed and RC values of one or more metrics on the overall uncertainty in assignment of water bodies to ecological status classes. See www.eu-star.at for further details about downloading the software and user manual.

Within STARBUGS, the ecological status class assessment for individual metrics can be based on just the observed (O) values of metrics or on Ecological Quality Ratios (EQRs) involving the ratio of the observed metric values to the RC (or RIVPACS-type Expected) values (E_1) of the metric. More generally, EQRs are determined by:

$$\mathrm{EQR} = \frac{O - E_0}{E_1 - E_0} \qquad (1)$$

where O = observed value, E_1 = Reference Condition value (= value of metric for which EQR = 1) and E_0 = value of metric for which EQR = 0.

Statistical distributions of the uncertainty in the estimated EQR values are obtained in STARBUGS using stochastic simulations, as follows

$$\text{Simulated EQR} = \frac{O + S + B - E_0}{E_1 + R - E_0} \qquad (2)$$

where S = random sampling (+ sub-sampling) variation term, B = random sorting and identification bias and variation term, R = random error in estimating RC value E_1.

The estimates of overall replicate sampling (including perhaps sub-sampling) SD for the term S can be obtained from Clarke et al. (2006a), the Deliverable 8 report on the STAR web-site www.eu-star.at, or elsewhere as appropriate for the metric(s), stream type and sampling methods. The sample sorting and identification term B is more complex as such errors can lead to both additional variances and systematic biases. For example, inexperienced staff tends to miss some taxa present and under-estimate metrics involving taxonomic richness. An estimate of error SD for the RC values could, for example, be obtained from the standard error of the simple or weighted mean of the metric values for the reference sites' samples on which the RC value was based. STARBUGS uses these various estimates of the components of uncertainty to generate many random simulations of the potential metric values for a site, from which the pre-defined metric-based classification rules and class boundaries are used repeatedly on each simulation to build up estimates of the probabilities that a particular water body belongs to each of the WFD ecological status classes.

It should always be remembered that there is no absolute truth. The uncertainty in any approach can only be assessed using the limited information available.

References

Buffagni, A., S. Erba, M. Cazzola, J. Murray-Bligh, H. Soszka & P. Genoni, 2006. The STAR common metrics approach to the WFD intercalibration process: Full application for small, lowland rivers in three European countries. Hydrobiologia 566: 379–399.

Clarke, R. T., 2004. 9th STAR Deliverable. Error/Uncertainty Module Software STARBUGS (STAR Bio Assessment Uncertainty Guidance Software) User Manual. www.eu-star.at.

Clarke, R. T., M. T. Furse, J. F. Wright & D. Moss, 1996. Derivation of a biological quality index for river sites: comparison of the observed with the expected fauna. Journal of Applied Statistics 23: 311–332.

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.

Clarke, R. T., J. F. Wright & M. T. Furse, 2003. RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. Ecological Modelling 160: 219–233.

Clarke, R. T., J. Davy-Bowker, L. Sandin, N. Friberg, R. K. Johnson & B. Bis, 2006a. Estimates and comparisons of the effects of sampling variation using 'national' macroinvertebrate sampling protocols on the precision of metrics used to assess ecological status. Hydrobiologia 566: 477–503.

Clarke, R. T., A. Lorenz, L. Sandin, A. Schmidt-Kloiber, J. Strackbein, N. T. Kneebone & P. Haase, 2006b. Effects of sampling and sub-sampling variation using the STAR-AQEM sampling protocol on the precision of macroinvertebrate metrics. Hydrobiologia 566: 441–459.

Dines, R. A. & J. A. D. Murray-Bligh, 2000. In J. F. Wright, D. W. Sutcliffe & M. T. Furse (eds), Assessing the Biological Quality of Freshwaters: RIVPACS and Similar Techniques, Freshwater Biological Association, Ambleside, pp. 71–78.

European Union, 2000. Directive 2000/60/EC. Establishing a Framework for Community Action in the Field of Water Policy. European Commission PE-CONS 3639/1/100 Rev 1, Luxemburg.

Furse, M., D. Hering, O. Moog, P. Verdonschot, R. K. Johnson, K. Brabec, K. Gritzalis, A. Buffagni, P. Pinto, N. Friberg, J. Murray-Bligh, J. Kokes, R. Alber, P. Usseglio-Polatera, P. Haase, R. Sweeting, B. Bis, K. Szoszkiewicz, H. Soszka, G. Springe, F. Sporka & I. Krno, 2006. The STAR project: context, objectives and approaches. Hydrobiologia 566: 3–29.

Haase, P., J. Murray-Bligh, S. Lohse, S. Pauls, A. Sundermann, R. Gunn & R. Clarke, 2006. Assessing the impact of errors in sorting and identifying macroinvertebrate samples. Hydrobiologia 566: 505–521.

Lorenz, A. & R. T. Clarke, 2006. Sample coherence – a field study approach to assess similarity of macroinvertebrate samples. Hydrobiologia 566: 461–476.

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., 2004. Comparison of cost effectiveness between various macroinvertebrate field and laboratory protocols. European Commission, STAR (Standardisation of river classifications), Deliverable, N1, 78 pp, www.eu-star.at.

Vlek, H. E., F. Šporka & I. Krno, 2006. Influence of macroinvertebrate sample size on bioassessment of streams. Hydrobiologia 566: 523–542.