Environment Agency

using science to create a better place

Uncertainty in WFD assessments for rivers based on macroinvertebrates and RIVPACS

Integrated catchment science programme

Science report: SC060044/SR4

SNIFFER

SCOTLAND & NORTHERN IRELAND FORUM FOR ENVIRONMENTAL RESEARCH

The Environment Agency is the leading public body protecting and improving the environment in England and Wales.

It's our job to make sure that air, land and water are looked after by everyone in today's society, so that tomorrow's generations inherit a cleaner, healthier world.

Our work includes tackling flooding and pollution incidents, reducing industry's impacts on the environment, cleaning up rivers, coastal waters and contaminated land, and improving wildlife habitats.

This report is the result of research commissioned and funded by the Environment Agency's Science Programme.

Published by:
Environment Agency, Rio House, Waterside Drive, Aztec West, Almondsbury, Bristol, BS32 4UD
Tel: 01454 624400 Fax: 01454 624409
www.environment-agency.gov.uk

ISBN: 978-1-84911-038-9

© Environment Agency May 2009

All rights reserved. This document may be reproduced with prior permission of the Environment Agency.

The views and statements expressed in this report are those of the author alone. The views or statements expressed in this publication do not necessarily represent the views of the Environment Agency and the Environment Agency cannot accept any responsibility for such views or statements.

This report is printed on Cyclus Print, a 100% recycled stock, which is 100% post consumer waste and is totally chlorine free. Water used is treated and in most cases returned to source in better condition than removed.

Further copies of this report are available from:
The Environment Agency's National Customer Contact Centre by emailing:
enquiries@environment-agency.gov.uk
or by telephoning 08708 506506.

Author(s):
Clarke, R.

Dissemination Status:
Publicly available / Released to all regions

Keywords:
Macroinvertebrates|RIVPACS|uncertainty|replicatesampling|WFD|termporalvariance|spatialvariation

Research Contractor:
Ralph Clarke, Centre for Ecology & Environmental Change (CCEEC), School of Conservation Sciences, Bournemouth University, Dorset House, Talbot Campus, Poole, Dorset, BH12 5BB

Environment Agency's Project Manager:
Veronique Adriaenssens
Environment Agency Science Department

Collaborator(s):
Scotland and Northern Ireland Forum for Environmental research (SNIFFER)

Science Project Number:
SC060044/SR4

Product Code:
SCHO0409BPUE-E-P

Science at the Environment Agency

Science underpins the work of the Environment Agency. It provides an up-to-date understanding of the world about us and helps us to develop monitoring tools and techniques to manage our environment as efficiently and effectively as possible.

The work of the Environment Agency's Science Department is a key ingredient in the partnership between research, policy and operations that enables the Environment Agency to protect and restore our environment.

The science programme focuses on five main areas of activity:

- **Setting the agenda**, by identifying where strategic science can inform our evidence-based policies, advisory and regulatory roles;
- **Funding science**, by supporting programmes, projects and people in response to long-term strategic needs, medium-term policy priorities and shorter-term operational requirements;
- **Managing science**, by ensuring that our programmes and projects are fit for purpose and executed according to international scientific standards;
- **Carrying out science**, by undertaking research – either by contracting it out to research organisations and consultancies or by doing it ourselves;
- **Delivering information, advice, tools and techniques**, by making appropriate products available to our policy and operations staff.

Steve Killeen

Head of Science

Executive summary

The Water Framework Directive (WFD) requires EU member states to assess, monitor and, where necessary, develop programmes of measures to improve the ecological quality status of all water bodies (lakes, rivers, coastal and transitional waters). These assessments should be based on standardised sampling or surveying methods for one or more biological quality elements (BQEs), including fish, macroinvertebrates, phytoplankton, diatoms and macrophytes, and habitats.

Whatever methods, procedures and rules are used, the WFD requires estimates of the uncertainty associated with the estimate of status classification for any water body.

The RIVPACS (River Invertebrate Prediction And Classification System) approach compares the observed (O) macroinvertebrate fauna and metric values with model-derived site-specific predictions of expected (E) values (based on environmentally similar high quality reference sites). This approach has been used to assess the ecological condition of UK water bodies since the 1990s. RIVPACS pre-dates (and helped inform) the WFD.

Implementing the WFD in the UK has required updating the RIVPACS methods with new models. This has included: adjustments for varying reference site quality; the introduction of new metrics; revision of the methods for monitoring overall quality in space and time and uncertainty implications; and incorporation of updates into the new software package RICT (River Invertebrate Classification Tool).

The objective of this report is to describe, quantify and asses the effect of various sources of variation on the uncertainty in estimates of river quality based on macroinvertebrate sampling and RIVPACS. However, many of the topics discussed also apply to other types of water body and BQE. The report includes the best-available estimates of the various variance components, based on a mixture of past and new datasets and statistical analyses.

The following topics are covered.

- The effect of uncertainty on the confidence of ecological status class and the probability of ‘moderate or worse’ status.
- The RIVPACS approach and the simulation of uncertainty associated with each step.
- Errors in determining the expected or reference condition values of metrics.
- How the spatial and temporal scale being assessed and monitored (one site or whole water body; one point in time or one season or one year or three-year period) influences the sampling requirements and the uncertainty of estimates.
- Deriving statistical estimates of components of variance.
- Assessment and estimates of replicate sampling variability.
- The effect of sample processing errors.
- Temporal variability (short-term within season, seasonal, inter-year variation within three-year reporting periods) and its estimation.
- Spatial variability between sampling sites within a WFD water body and its estimation.
- Links to and comparison with the VISCOUS approach.
- Comparing combined season sample observed/expected (O/E) values with average single season sample O/E values.
- Uncertainty for multi-metric and minimum or worst case rules.

- Uncertainty results from the Europe-wide STAR (Standardisation of River classifications) project, including the relative precision of different metrics and sampling methods.

The following recommendations come out of this study.

- The new RICT software should include specific estimates and information on the confidence of failing to achieve good or better status in addition to the confidence of belonging to individual WFD status classes.
- There is a need to collate and analyse a much larger dataset of spatial variability between sampling sites within the new WFD water bodies, ideally with temporal and replicate sampling information on at least a subset of the same sites. This should allow improved estimates of the scale of spatial heterogeneity within rivers.
- Methods giving fixed RIVPACS predictions for each site should be developed. These should be based on either temporally-invariant Geographic Information System/map-based site and catchment environmental variables or long-term (five-year) average environmental variables. This would provide O/E values for every sample and allow direct assessment of O/E variance components.
- Some environmental parameters can be affected by flow and so current predictions can miss the impact of abstraction. There is a need to develop predictions that are not influenced by flow or new rules for using such data for WFD predictions (because flow pressures are to be considered).
- There should be further analyses of RIVPACS sample audit data to derive and incorporate direct estimates of sample processing errors and biases in other indices (in addition to NTAXA) into the RICT software for assessing confidence of class.
- The merit of using the average of single season sample O/E values as a measure of water body quality over a period should be reconsidered, and contrasted with the current combined season sample approach.
- Statistical methods to cope with any actual temporal and spatial mix of samples should be developed and these methods incorporated into either the RICT software or an extended version of the VISCOUS-type software tool.
- A standardised sampling approach for assessing non-wadeable rivers (based on Environment Agency/North South Shared Aquatic Resource/Centre for Ecology & Hydrology ‘deep rivers’ research) should be developed and a Biological Assessment Methods-like study to quantify uncertainty conducted.

Acknowledgements

I would like to thank Veronique Adriaenssens of the Environment Agency for support and encouragement for this work, both before and throughout the research, and also for help in preparing the Dove catchment data.

Contents

1 Introduction

1.1 WFD and uncertainty

The EU Water Framework Directive (WFD 2000) requires each member state to assess, monitor and, where necessary, improve the ecological quality of its water bodies (rivers, lakes, transitional and coastal waters). Assessing the quality of these water bodies involves the use of one or more Ecological Quality Ratios (EQRs). Each EQR represents the relationship between the value of a biological parameter (index or metric) observed for a water body and the expected value for that parameter if the water body were in reference condition The WFD requires each member state to use these EQRs to classify water bodies into one of five ecological status classes (see Figure 1.1).

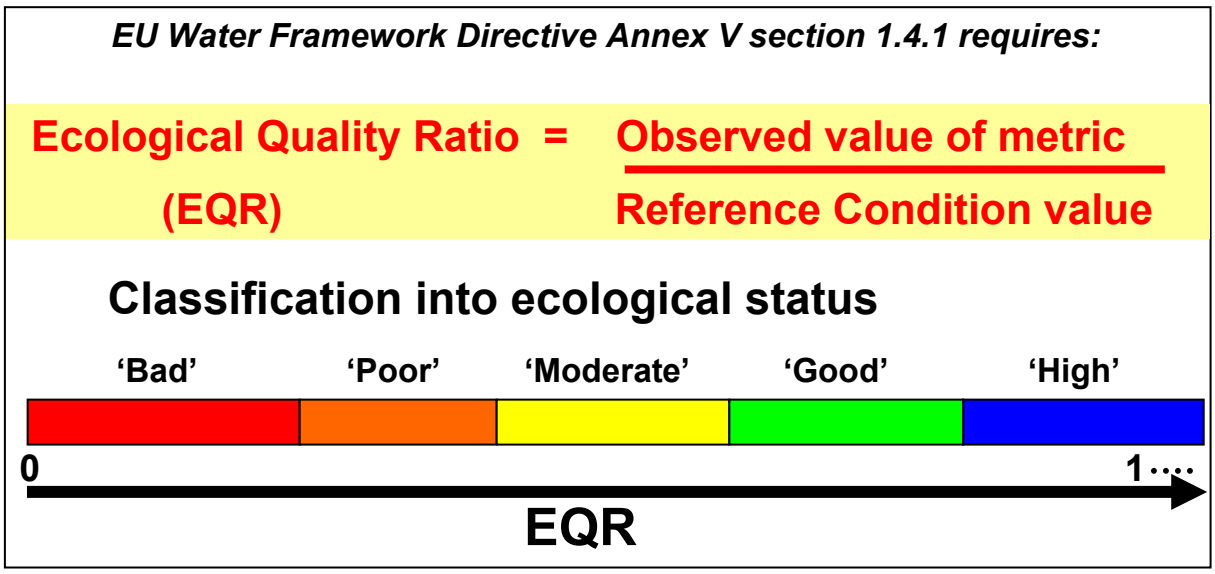


Figure 1.1 WFD requirement for the use of EQRs divided into five ecological status classes

The WFD also requires member states to establish a monitoring network and to monitor any changes in the ecological status of water bodies. Ideally, all water bodies should be in 'good' or higher ecological status by 2015 (WFD, Article 4, 1(a)(ii)). Where water bodies are currently judged to be of insufficient quality, member states are required to develop 'programmes of measures' for that river basin or sub-catchment to help improve its ecological status.

Any measures of ecological quality are of little value without some knowledge and quantitative estimates of the precision and confidence with which they assign sites and water bodies to ecological status classes. This is a requirement of the 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' (WFD, Annex V, Section 1.3).

The WFD recommends (Annex V, Section 1.3.4) that monitoring based on macro-invertebrate sampling should be based on intervals of no more than three years and that sampling 'frequencies shall be chosen so as to achieve an acceptable level of

confidence and precision'. Thus, the concept that all estimates of ecological quality for water bodies are subject to a range of sources and levels of uncertainty is an integral part of the WFD.

1.2 River assessment using macroinvertebrates and RIVPACS

The WFD requires member states to assess the ecological status of its rivers based on appropriately informative aspects of the biota at the site. These biota (referred to as Biological Quality Elements (BQE) in the WFD) can be phytoplankton, macrophytes and phytobenthos, benthic invertebrate fauna and fish fauna (WFD, Annex V, Section 1.2). River water body assessments can be based on either a single BQE or a combination of BQEs. The choice of BQEs and the metrics to be used within each BQE should depend upon their ability (statistical power and precision) and cost-effectiveness at quantifying the ecological quality of river sites, and detecting and quantifying changes in quality within monitoring programmes.

1.2.1 RIVPACS

In the UK, national biological monitoring of rivers has concentrated on the gradual collaborative development and use of RIVPACS (River Invertebrate Prediction And Classification System; Wright *et al*. 1984, Wright 2000, Wright, Sutcliffe and Furse 2000). This has mainly involved research staff at the Freshwater Biological Association (FBA), the Institute of Freshwater Ecology (IFE) and the Centre for Ecology & Hydrology (CEH), and staff within the UK government environment agencies – the Environment Agency, the Scottish Environment Protection Agency (SEPA) and the Northern Ireland Environment Agency (NIEA). RIVPACS provides standardised macroinvertebrate sampling and bioassessment methods, which are applicable to all types of wadeable streams and rivers throughout the UK.

RIVPACS works by comparing the observed fauna and observed values of derived biotic indices with the site- and season-specific expected fauna and expected values of those indices. The expected fauna are based on a previously-developed predictive statistical model that relates the observed fauna of high quality reference sites to their environmental characteristics (Figure 1.2).

RIVPACS can be used to estimate the expected value of any macroinvertebrate-based biotic index for a monitoring site, as follows:

$$\text{Expected index value} = E_I = \sum_{k=1}^{g} p_k I_k$$

where p_k = RIVPACS probability of site belonging to RIVPACS site end-group k
I_k = Average value of index for the RIVPACS reference sites in end-group k
g = number of end-groups.

Separate predictions are made for samples from each of the three RIVPACS sampling seasons: spring (February–May); summer (June–August); and autumn (September–November). It also makes predictions for combined season samples based on any combination of two or three seasons within one year.

Since the mid-1990s, national assessments of river sites have been based on the use of RIVPACS O/E (observed/expected) values for two macroinvertebrate indices identified to BMWP (Biological Monitoring Working Party; see Armitage 1983) family taxonomic level. Under the BMWP system, each family is given a score (1–10) based

on its perceived tolerance to pollution, especially organic pollution; a score of 10 indicates least tolerance and a high susceptibility to organic stress. The two indices are: (i) number of BMWP-scoring families present in a sample (denoted by the term NTAXA) and (ii) the average score per taxon present, calculated as the sum of scores of all taxa present (denoted the BMWP score) divided by NTAXA, and referred to by the term ASPT.

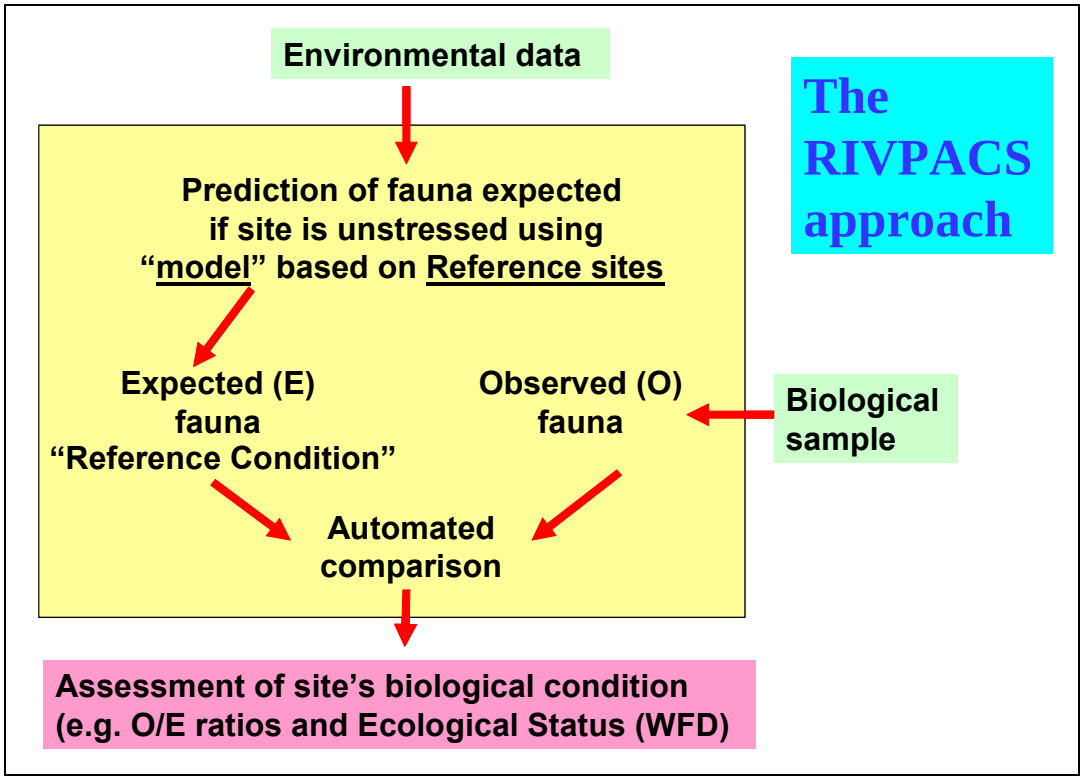


Figure 1.2 Schematic diagram of the RIVPACS bioassessment approach

A site assessment includes the following steps:

i. Calculate observed (O) sample values of each index.

ii. Calculate site- and season- specific expected (E) value of each index.

iii. Calculate O/E ratios for each index, denoted O/E_{NTAXA} and O/E_{ASPT} for these indices.

iv. Classify the site into a quality class based on pre-determined class limits for each O/E (this is done independently for each index).

v. A site's overall class is determined by the worst of its classes based on O/E_{NTAXA} and O/E_{ASPT}.

vi. Assess probability/confidence of class both for each index and overall using RIVPACS uncertainty simulation software, based on prior estimates of the various sampling variance and error terms.

This is a form of the 'worst-case' or 'one-out all-out' rule. Since 1995, sites have been classified into one of six classes (a–f) based on the limits shown in Figure 1.3. However, this system of classification is currently being revised to comply with the WFD and its prescribed five class classification. Hence, new class limits and rules will be incorporated into the River Invertebrate Classification Tool (RICT) software currently

being developed within SNIFFER (Scotland and Northern Ireland Forum for Environmental Research)/SEPA projects.

Very importantly from the point of view of standardisation and consistency, a detailed procedures manual (Environment Agency 1997) provides guidance on how to collect and analyse RIVPACS samples. This manual also provides detailed instructions on how to measure and obtain values for each of the environmental variables used in the RIVPACS software to derive predictions for the site-specific expected fauna and the expected values of biotic indices.

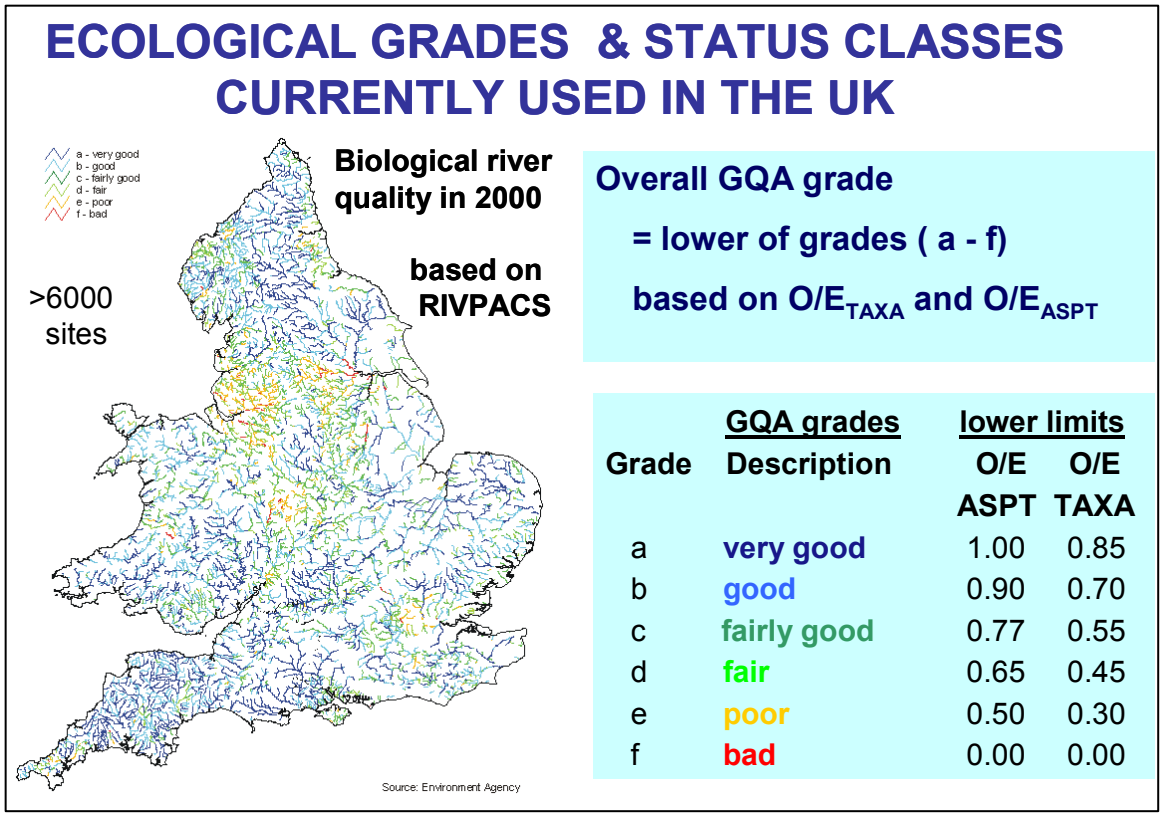


Figure 1.3 General Quality Assessment ecological grades and limits used for UK river assessments based on RIVPACS O/E values for NTAXA and ASPT

In addition, since RIVPACS was first used in national assessment surveys in 1990, the RIVPACS approach has pioneered the establishment and continued use of internal quality assurance and external quality auditing schemes. These assess and help to improve and maintain the quality of laboratory sample sorting and taxonomic identification skills.

By the early 1990s, the RIVPACS development team had already grasped the importance of understanding and quantifying the sampling errors and uncertainty associated with the RIVPACS (or any other) method of deriving estimates for the biological quality of freshwaters. This led to the carefully-designed BAMS (Biological Assessment Methods) study, which explored RIVPACS field sampling variation, the consequences of sample processing and identification errors, and the effects of errors in measuring the RIVPACS environmental predictor variables (Furse *et al*. 1995, Clarke *et al*. 2002).

The apparent success of RIVPACS and its professional approach to standardisation, consistency of methodology and attempts to measure uncertainty probably helped

contribute towards the overriding importance placed on biological/ecological assessments within the WFD.

However, the various steps involved in RIVPACS bioassessments are, as with all methods, prone to considerable sources of natural spatial and temporal variation in biota, measurement and prediction error; all of which contribute to uncertainty in site and water body assessments of EQR values and ecological status. Each of these sources of uncertainty, together with how we have attempted to assess, quantify and incorporate them, are discussed in detail in subsequent sections of this report.

1.2.2 Need for sampling and assessment methods for non-wadeable rivers

At present, RIVPACS only covers river sites that can be sampled by kick-sampling in a stream with a pond net. Non-wadeable (deep and/or fast flowing) rivers are not covered by the RIVPACS reference sites and kick-net sampling. Comparing an observed sample index value obtained using a deep-water sampling method such as air-lift with a RIVPACS expected value based on kick-net samples is not comparing like with like. Such a comparison is likely to give biased estimates of O/E according to whether a pond-net and kick-sampling provides a different number and range of taxa to an air-lift. There is still a need to develop a standardised sampling approach for assessing non-wadeable rivers (based on Environment Agency/NS Share (North South Shared Aquatic Resource)/CEH 'deep rivers' research) and to undertake a BAMS-like study to quantify uncertainty.

1.3 Need for estimates of confidence of class

1.3.1 Reasons for assessing uncertainty

We need to be able to generate statistical information on the uncertainty of any ecological quality classification scheme and the likely risk of misclassifying the status of sites and water bodies. EU member states are expected to maintain and, where necessary, help improve the overall condition of aquatic resources. This is so that, ideally by 2015, all water bodies are granted 'good' or higher ecological status, unless there are over-riding economic considerations. Given the importance, and implications, being placed on determining whether a water body is in 'moderate' or worse status class, we need to be able to estimate the probability that a water body could actually be of 'good' or better status. European environment agencies need to be confident that current river quality is inadequate in order to justify the costly measures that will be needed to get a water body to achieve 'good' or higher status. (We also need to provide evidence of how confident we are that any Programme of Measures (POMs) will actually improve the river ecology).

In addition, when we assess the ecological quality of a site in two different years or monitoring periods, the observed estimates of site quality will usually differ and the estimated 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 status class has occurred or whether the observed changes in biota and derived metrics and EQR values is just due to the errors and sampling variation inherent to the assessment process. How confident are we that a programme of measures within a catchment or sub-catchment has been effective in improving ecological status?

1.3.2 Illustrative example of misclassification rates

It is very useful, at this stage, to have some quantitative understanding of the general effects that sampling variation and other errors can have on the confidence with which we assign a site or water body to an ecological status class.

General formulae derived by Clarke *et al*. (1996) are summarised in Figure 1.4. These show the probability (P_M) of misclassifying a site/water body of any particular true quality (true EQR – plotted along the X-axis) in relation to the size of the errors or uncertainty in the estimated EQR values. In this sense, the true class can be thought of as the class of the average of all possible sample EQR values that we could have obtained for this site/water body.

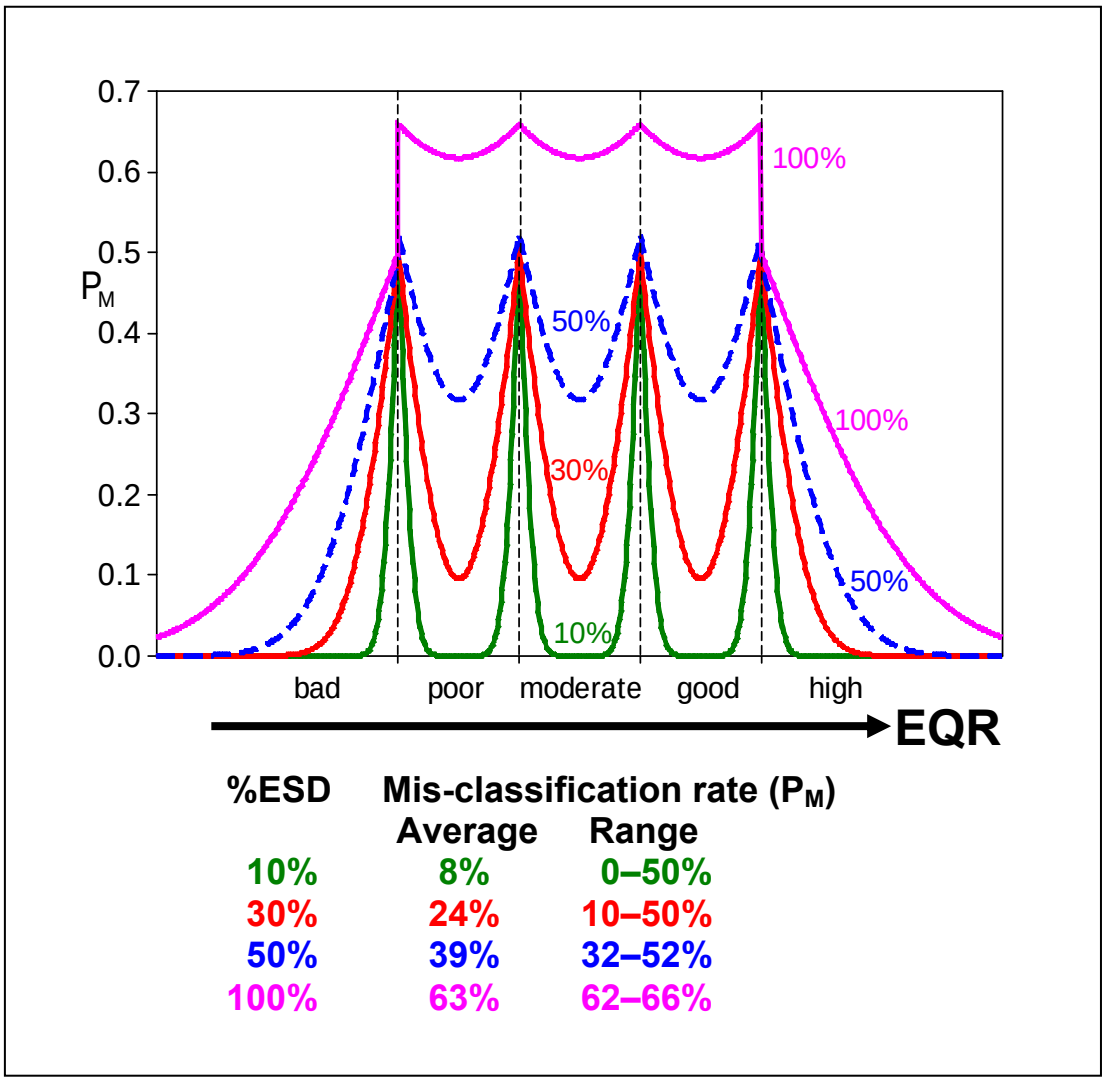
The error/uncertainty standard deviation (SD) of the EQR values represents the SD of the set of all possible EQR values that we could have obtained for that water body for the monitoring period with that sampling scheme. In other words, the SD that arises from sampling at different places within the site or water body and at different times within the period being assessed.

It is useful to express this uncertainty SD in EQR values (denoted ESD) as a percentage (denoted %ESD) of the width of an ecological status class (range of EQR values within a class). For example, if the lower limits in the EQR of a particular metric for the 'poor', 'moderate' and 'good' classes are 0.4, 0.6 and 0.8, then the width of each of these intermediate classes is 0.2. This means that if the uncertainty SD in EQR values for that metric is 0.05, then the %ESD is 25 per cent (0.05/0.2).

For illustrative simplicity, the width of the middle status classes ('poor', 'moderate' and good') are equal in Figure 1.4 and the errors/uncertainty are assumed to vary according to a normal distribution. However, in real situations, the status class widths may be unequal. In this case, for a given value of ESD, %ESD and the probability of misclassification will be higher for sites whose true class has a relatively narrower range of EQR values.

From Figure 1.4, when the uncertainty SD in the EQR values is only 10 per cent of the width of a status class (shown in green), sites whose true quality lies in the middle of the class would never be misclassified ($P_M = 0$). Sites whose true quality lies on (or almost on) the border of any two classes will always have at least a 50 per cent chance of being placed in the wrong status class. When error standard deviations are only 10 per cent of class width, the overall average misclassification rate for sites in a middle class, such as 'good', 'moderate' or 'poor', is only 8 per cent (assuming an even spread of true qualities across the class) (Figure 1.4).

If however, the uncertainty SD is 50 per cent of the class width (shown dotted in blue), even sites in the centre of a middle class have a 1 in 3 chance of being placed in the wrong class and 39 per cent of all sites in the class will be misplaced in either a higher or lower class.



%ESD	Mis-classification rate (P_M)	
	Average	Range
10%	8%	0–50%
30%	24%	10–50%
50%	39%	32–52%
100%	63%	62–66%

Figure 1.4 Probability (P_M) of misclassifying a site of any particular true quality (EQR) in relation to the uncertainty SD of the EQR expressed as a percentage (%ESD) of the width of a status class for that EQR

Note: Mean and range of P_M apply to the middle classes – good, poor and moderate – which are assumed to be of equal width.

If the error SD is equal to the class width (100 per cent; shown in purple), as is likely if the metric is highly susceptible to sampling variation and other effects, then all sites whose true quality lies within a middle class are more likely than not to be placed in the wrong status class (either a class above or below the true class). For example, a site whose true average EQR lies in the 'good' class, but very near the high/good boundary, will have a 50 per cent chance of being misclassified as high, but it will also have an additional 16 per cent chance of being misclassified as moderate (or even poor/bad), leaving only a 34 per cent (one in three) chance of being classified as good.

Obviously, sites of very 'high' or very 'bad' quality, which are well away from the boundaries of the best or worst status classes, are unlikely to be misclassified. The overall probability of misclassifying sites from the top or bottom classes ('high' or 'bad') is only half that for sites from middle classes. The probability is only one-quarter if the top or bottom class width in EQR values is twice that of the middle classes (assuming an even spread of true qualities across the class).

1.3.3 Confidence of 'good or better' or 'moderate or worse' status

Although the WFD requires that water bodies are classified into one of five ecological status classes, the primary concern is whether a water body is of good or better status or whether it is of moderate or worse status and requires a programme of measures for improvement.

If we are only interested in this dichotomy – 'good or better' versus 'moderate or worse' – then the confidence of class (probability of true class) merely depends on how far the water body's sample EQR value (denoted EQR_O) is from the critical good/moderate boundary EQR value (denoted $EQR_{G/M}$). This is relative to the size of the uncertainty SD (ESD) associated with the sample EQR value.

In assessing confidence of class, we assume that the random uncertainty variation associated with any sample EQR value follows a normal statistical distribution (a symmetrical bell shape with 95 per cent of the sample values within two SD of the true EQR mean). The same normality assumption was used by Julian Ellis and others in the development of the CAVE (Combines Appropriate Variance Estimates) and VISCOUS (Variability In Spatial Component Objectivity Unified Statistically) software and methods for assessing the effects of EQR uncertainty on confidence of class (see Ellis 2007).

If $EQR_{Diff} = EQR_O - EQR_{G/M}$ (amount sample EQR differs from the good/moderate boundary; sign (+/-) is crucial)

and

$EQR_Z = EQR_{Diff} / ESD$ (difference as multiple of the EQR uncertainty SD)

then P_{G+} = probability of a water body being of 'good or better' status
$= CDFNorm(EQR_Z)$

where CDFNorm(X) is the cumulative probability of a standard normal deviate (with zero mean and variance of unity) being less than or equal to X.

If a water body is not of good or better status, it must be of moderate or worse status, and therefore:

$P_{M-} = 1 - P_{G+}$ (probability of the water body being of 'moderate or worse' status)

This is summarised in Table 1.1.

Table 1.1 Probability of being 'good or better' (P_{G+}) and its complement, the probability of being 'moderate or worse' (P_{M-}), for a range of values of EQR_Z

EQR_Z	Probability 'good or better' (P_{G+})	Probability 'moderate or worse' (P_{M-})
-2.5	0.001	0.999
-2.0	0.023	0.977
-1.5	0.067	0.933
-1.0	0.159	0.841
-0.5	0.309	0.691
0.0	0.500	0.500
0.5	0.691	0.309
1.0	0.841	0.159
1.5	0.933	0.067
2.0	0.977	0.023
2.5	0.999	0.001

Note: These figures represent the extent to which the observed EQR exceeds the good/moderate boundary EQR value, standardised by the EQR uncertainty SD.

To have at least 95 per cent confidence that a water body is of 'good or better' status based on a single EQR, the sample EQR value needs to be at least 1.645 times the uncertainty SD above the critical good/moderate boundary value for that EQR.

Conversely, to have at least 95 per cent confidence that a water body is of 'moderate or worse' status based on a single EQR, the sample EQR value needs to be at least 1.645 times the uncertainty SD below the critical good/moderate boundary value.

These probabilities are further illustrated in Figure 1.5. This shows the probability of being of 'good or better' status for the complete range of sample EQR values, when the good/moderate boundary is set at 0.7 and the uncertainty SD varies from small (0.05) to large (0.25). When the uncertainty SD is only 0.05, the sample EQR needs to be only 0.617 or less to be at least 95 per cent confident that the water body is truly of 'moderate or worse' status (based solely on this EQR and its underlying metric(s)). However, when the uncertainty SD is larger, at 0.15 or 0.25, then the sample EQR needs to be no more than 0.453 and 0.288, respectively, to have 95 per cent confidence that the water body is not of good or better status.

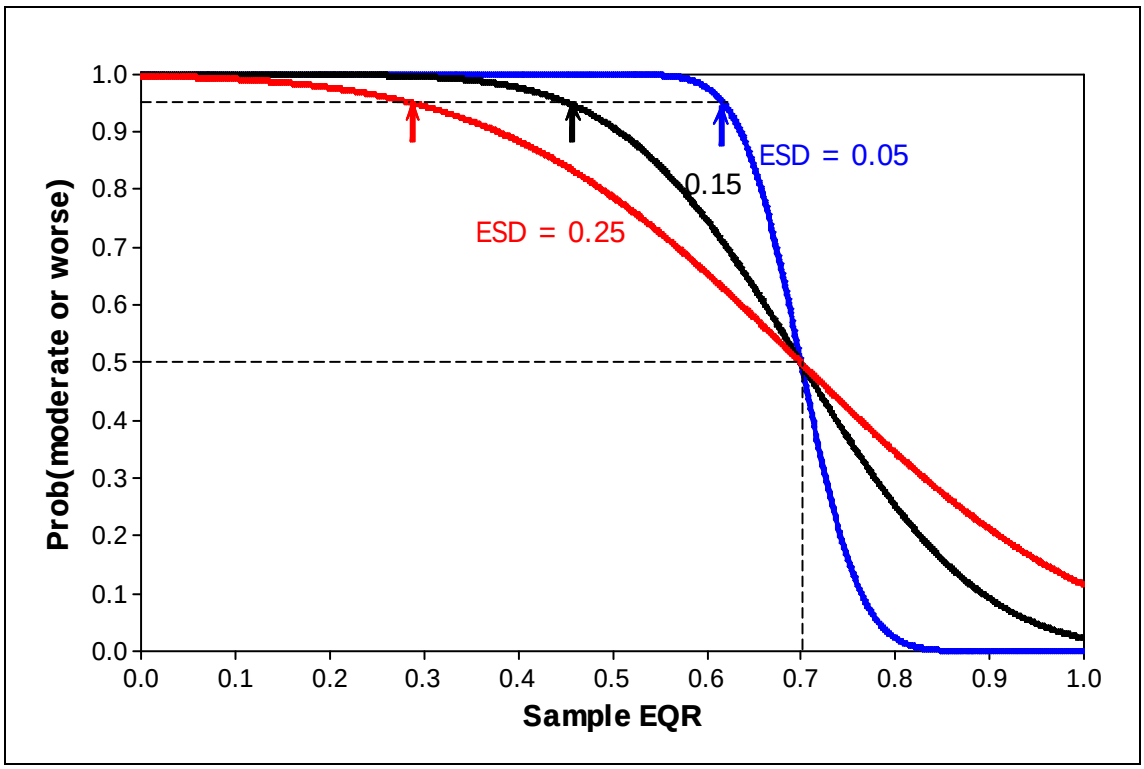


Figure 1.5 Probability (P_{M-}) that a given sample EQR value is from a water body whose true status class is 'moderate or worse' for the example situation where the good/moderate boundary value is set to 0.7 and the sample EQR uncertainty standard deviation (ESD) is 0.05 (blue), 0.15 (black) or 0.25 (red)
Note: Arrows indicate EQR values required to give 95 per cent confidence of 'moderate or worse' status class.

1.3.4 Uncertainty for sites with O/E values greater than one

For river bioassessments based on macroinvertebrates and RIVPACS, the EQR values are based on RIVPACS O/E ratios that can exceed one. Although WFD EQR ratios are supposed to be confined to the range 0–1, it is logical when assessing status class uncertainty to use the actual O/E values and derive or simulate estimates of confidence around these values in order to estimate confidence of class for the whole site. This is because if the O/E value is considerably above one then the site is definitely of high status. Whereas, if we re-set the O/E value to a maximum EQR value of one before simulating the uncertainty related to this value then we might under-estimate the confidence that the site was of high quality. If required for WFD reporting purposes, O/E values greater than one can be reset to one, but only after the confidence of class has been based on assessing uncertainty about the actual RIVPACS O/E value.

More generally, outside of RIVPACS, the WFD requirement to constrain EQR values, based on comparing observed and expected/predicted index values derived from available reference sites, to not exceed one can cause practical problems. This means great care should be taken when assessing uncertainty and confidence of class for sites with WFD EQR values set, or reset, to one.

In the CAVE approach for assessing uncertainty in EQR values, Ellis developed a method for fitting a curve to data of the relationship between the SD of sample EQR values and the mean of the sample EQR values. This forced the quadratic-like regression relationship to pass through SD=0 when the mean EQR was 0 and 1. This is done on the basis that for the (assumed true) mean EQR value to be a minimum of

zero and a maximum of one there cannot be any variation between the sample values at the site with such a mean value. However, in real situations, the mean EQR value for a site is often based on just a single, or maybe two or three, samples.

In such cases, the observed sample EQR value(s) could by chance easily all be one, especially if O/E (EQI Environmental Quality Index) values (much) greater than one are all reset to one. But other samples from the site might give EQR values less than one and there is real uncertainty in the site assessments. Therefore, it is incorrect to assume there is no sampling uncertainty when the mean sample EQR is one. This highlights an advantage in the RIVPACS-type simulation of uncertainty in the 'raw' O/E values: the simulated values can then be converted to EQR values on the 0–1 scale using the same rules as applied to the actual sample O/E values.

2 Sources of uncertainty in estimates of biotic indices and ecological status class

An estimate of any index, metric or measure of freshwater ecological quality (irrespective of whether a RIVPACS-type approach is involved) is of little value unless we have some idea of the sources and sizes of the sampling and other potential errors and uncertainty involved.

Using a RIVPACS or related WFD approach to assess ecological quality, the observed (O) values of one or more biotic indices are compared with the expected (E) or reference condition (RC) values of those indices through the use of EQRs. These are most often calculated in the form of O/E ratios, as initially developed in RIVPACS. Thus the uncertainty in any such EQR values is potentially due to sampling and other errors in the observed metric value and estimation and/or modelling errors in setting the expected or RC value of the index. Each source of uncertainty is discussed below.

2.1 Sources of variation in the observed fauna and observed index values

Catchments and rivers must be sub-divided into mutually-exclusive WFD water bodies (sections) for monitoring purposes. Within each section, one or more sampling site(s) are taken to be representative of the prevailing ecological condition in the water body as a whole.

For UK national monitoring of the ecological quality of rivers based on macroinvertebrates and RIVPACS, each river has been divided into sections, now referred to as water bodies. WFD water bodies were chosen (at least by the Environment Agency) according to simple typological criteria related to WFD System A criteria for stream types (catchment area, altitude, geology), without regard to the existing system of General Quality Assessment (GQA) monitoring reaches which were selected as relatively homogenous sections of river.

Usually, a single sampling site is taken to be representative of the whole water body. However, in some situations, there are two or more existing monitoring sites within the same new water body, and this provides valuable information on spatial variability between sites within a water body (this is discussed in Chapter 6). Unfortunately, partly because of the way water bodies are formed, there are some newly-formed water bodies with no past or current monitoring sites. The condition of these water bodies must be inferred from ‘environmentally-similar’ water bodies.

In the 1990s, when uncertainty issues were first considered within RIVPACS, the practical aim was to assess uncertainty in RIVPACS O/E estimates of ecological quality for a single site at a set point in time. Thus, only two sources of variation in observed index values were assessed and allowed for in the RIVPACS III+ software uncertainty simulation module. These were:

- replicate sampling variation (differences between replicate samples taken at the same site on the same day);
- sample processing and taxonomic identification errors.

However, with the advent of the WFD, the future strategy for national monitoring will now be based (at least by SEPA) on estimates of the ecological quality of each water body over a three-year period (see section 7.1). In particular, it is very likely to be based on an estimate of the 'average' quality across the water body as a whole over the three-year period. This estimate should be based on sampling at one or more sites within the water body at one or more times within the period. A consequence of this is that the uncertainty in such estimates of 'average' quality should now also involve a range of other sources of variation including:

- Spatial variability between sampling sites within a water body ('inter-site').
- Temporal variability between days within a RIVPACS season ('within-season').
- Temporal variability between years with the three-year period ('between-years')
- A spatio-temporal interaction if site differences within the water body vary over time.

In the extreme, if some water bodies are assessed solely according to the quality of other water bodies considered to have similar risks, then any similarity or variation between these water bodies will also affect the uncertainty and potential errors in the assessments.

In reality, spatial variability occurs at a continuity of scales within a water body, from centimetres to kilometres. But it is convenient and makes estimating spatial variability more practical if the spatial variation is sub-divided into two hierarchal levels: within-site replicate RIVPACS sampling variability and inter-site (within water body) variability.

Each of these sources of variation in the observed fauna and observed values of biotic indices are discussed in the following chapters:

Chapter 3 discusses estimating replicate sampling variability, notably through the Biological Assessment Methods (BAMS) study.

Chapter 4 discusses the potential impact of, and the means for estimating and allowing for, bias due to sample processing and taxon identification errors (based on the BAMS study and RIVPACS quality audit results from CEH).

Chapter 5 discusses estimating both within-season and between-year within period temporal variance in index values, based on a combination of Environment Agency, SEPA and NIEA datasets.

Chapter 6 discusses estimating the spatial 'inter-site' variability within water bodies, based on an analysis of a new dataset comprising 2–3 sites within each of three water bodies within the Dove catchment.

Real temporal variance implies real differences over time in the average sample biota and average index values at a site. It is measured by recording the variation in observed index values between samples taken at different times (days, weeks or years), over and above the normal expected variation between any replicate samples.

Real temporal and spatio-temporal variation and changes could be due to 'natural' effects, such as the weather (storms, droughts), or to environmental or anthropogenic stress or pollution. It may be difficult in practice to differentiate between natural temporal variation and anthropogenic causes of change in the biota at individual sites. For example, what about the biological effects of a reduction in river discharge? Reductions in flow due to weather patterns or changes in climate may be considered natural, but reductions in river flow as a result of abstraction could be considered to be a man-induced stress. For, without abstraction, the river flows in dry years would not have led to any ecological stress on the biological communities.

In general statistical modelling terminology, the natural temporal variability can be viewed as just adding to the uncertainty and being part of the background 'noise'. While stress related to real changes can be considered the 'signal' we are trying to detect, quantify and distinguish from the noise.

2.2 Sources of uncertainty in the expected fauna and expected index values

RIVPACS estimates of the season-specific fauna and biotic index values expected for a monitoring site are based on a statistical model. This is developed from the relationships between the macroinvertebrate sample community composition at a large number of high quality RIVPACS reference sites and a suite of physical and environmental variables measuring key aspects of the environmental characteristics of these sites.

The UK RIVPACS models are based on first combining the reference sites into site groups based on the similarity of their macroinvertebrate fauna using the statistical ordination and clustering technique known as TWINSPAN (Hill 1979). Then the multivariate statistical technique known as multiple discriminant analysis (MDA) is used to derive discriminant functions for estimating the probability of any site belonging to each of the RIVPACS site groups based on its values for key environmental predictor variables. These probabilities of group membership can be combined with the average values for taxa and indices within each site group to derive site-specific expected values for the taxa occurrences, abundances and biotic indices.

The errors in estimating the expected fauna and WFD reference conditions for a site are potentially due to one or more of the following factors.

2.2.1 Inadequate set of reference sites

Having too few high quality reference sites for all or certain stream types will lead to imprecise or inadequate setting of the 'target' or reference condition fauna and index values. If the reference sites are of inadequate quality, then the expected values of BMWP indices will be too low and monitoring site quality will be over-estimated. A subtler but even more likely problem is that it may be difficult to find enough, or even any, high quality sites for particular types of river (such as rivers in intensively farmed, densely populated lowland Britain). This will mean that weaker targets (or reference conditions) will be set for such types of river and EQR values and status classes may be systematically over-estimated in these regions.

Whatever method is used, if a prediction is based on a very small sample of reference sites, the estimated mean (or other percentile) index value used to set the expected or Reference Condition value is likely to be imprecise and subject to high standard errors – this was discussed in the EU REFCOND project (http://www-nrciws.slu.se/REFCOND/index.html). Although RIVPACS nominally utilises a weighted

average of all reference sites in each prediction, in reality the prediction for some extreme sites (such as the Shetlands) could be based on the observed fauna in just a few environmentally-similar reference sites.

2.2.2 Not involving all relevant environmental variables

The WFD permits the determination of reference conditions for a site from some form of the average (such as mean, median) biota of the reference sites in the same stream type. For example, System A stream types are based on only three or four classes of altitude, catchment area and geology (WFD Annex II, Section 1.2). WFD System type B and site-specific predictive models like RIVPACS, which use more site factors, might be expected to give more precise targets. The suite of environmental variables used in RIVPACS predictions is given in Table 2.1.

As the aim of the predictive models and methods is to define the index values expected in the absence of environmental stress, we try not to incorporate any variables whose values at the time of measurement may have already been altered by the stress that we are trying to assess.

This encourages the use of time-invariant variables (such as site altitude, distance from source and underlying upstream catchment geology). There is some attraction to having a fixed prediction for each site based on just time-invariant 'map-derived' variables. As well as long-term (five years) historical average values of site characteristics measured in the field prior to the occurrence of current or unacceptable levels of stress at the site (as recommended previously by CEH).

However, the macroinvertebrate community varies with the precise small-scale flow and plant habitat conditions within a site. For example, RIVPACS deliberately does not utilise any measure of macrophyte cover and composition, as this is ephemeral, seasonally-volatile and affected by recent flow regimes, but it does influence the macro-invertebrates present at a site.

Table 2.1 Environmental variables used in RIVPACS predictions

<u>Time invariant (map-based variables)</u> map location (National Grid Reference) → latitude, longitude → mean air temperature, air temperature range altitude at site (m) distance from source (km) slope (m km^{-1}) discharge category (1–10) (1 = ≤0.31, 2 = 0.31–0.62, 3 = 0.62–1.25, …, (long-term historical average) 9 = 40–80, 10 = 80–160m^3s^{-1} mean daily flow)
<u>Estimated at site at time of sampling</u> (averaged across the three RIVPACS seasons) stream width (m) stream depth (cm) Sub-stratum composition: %cover of clay/silt, sand, gravel/pebbles, cobbles/boulders →mean particle size (in phi units) water geo-chemistry: alkalinity (mg l^{-1} $CaCO_3$) supplied as annual average by Environment Agency chemists

Note: → denotes derived variable created internally within the RIVPACS software.

2.2.3 Choice of statistical prediction method or modelling technique

RIVPACS-type predictive models are based on the biological classification of sites followed by the probabilistic environmental MDA of site groups. There are a range of alternative measures for estimating the similarity of biological composition between pairs of sites (Sorensen presence-absence or Bray-Curtis (relative) abundance-based measures) and a variety of potential site/sample clustering methods (nearest-neighbour, average-linkage, hierarchical divisive (splitting) or agglomerative (combining)) for subsets of reference sites.

Alternatively, completely different statistical approaches, such as neural networks, Bayesian belief networks or direct predictive (perhaps non-linear) multiple regression models of index values for reference sites in relation to the values for the same set of RIVPACS environmental variables, could be used to derive estimates of expected values of biotic indices. For example, the EU FAME (Development, Evaluation and Implementation of a Standardised Fish-based Assessment Method for the Ecological Status of European Rivers) project developed by the European Fish Index (EFI; http://fame.boku.ac.at/downloads/manual_Version_Februar2005.pdf) is based on an average of the probabilities associated with the residuals (observed minus predicted values) from multiple regression prediction equations for each of 10 fish community metrics derived from features of the best available reference sites.

It should be possible to compare the relative accuracy of predictions of biotic indices amongst a set of reference sites produced by two or more prediction methods. For example, Walley and Fontama (1998, 2000) used neural network techniques on the RIVPACS III mainland Britain reference sites database to derive alternative predictors of the expected NTAXA and ASPT from the same set of environmental predictor variables as used on RIVPACS. The correlations (r) between observed and predicted index values produced by their best models were very similar to those based on RIVPACS III (r = 0.84–0.85 for ASPT and r = 0.67 for NTAXA for both methods).

2.2.4 Errors in measuring the environmental variables

RIVPACS predictions for a site in a particular year should be based on the average values of the environmental variables measured during each of the three seasons' sampling visits. This means that the prediction of the expected fauna changes a little each year. For future versions of RIVPACS, the intention of the Environment Agency and CEH is to derive fixed predictions of the average fauna to be expected at a site that would apply for all years. These predictions would be based on estimates of the average values of the environmental variables for the site measured over a period of at least five years (excluding any known extreme climatic years). However, little progress has yet been made in making such predictions in a consistent prescribed manner.

Errors made in measuring the values of the RIVPACS environmental variables will affect the RIVPACS MDA-based predicted probabilities of belonging to each site group and thus the expected biotic index values. This can apply both to the time-invariant variables from published maps and the variables measured in the field at the time of sampling.

Clarke *et al.* (1996) assessed the implications of errors in any one predictor variable by simulating the addition of independent random normal errors with either a fixed absolute standard error (SE of 1,2,3), or a percentage SE (%SE) for log-transformed variables. They did this for the recorded value of the variables for each RIVPACS site and then assessed the change in expected and O/E index values. They adopted a range of potential values for SE (i.e. 1, 2, 3) or %SE (i.e. 5 per cent, 10 per cent, 15 per cent, up to

100 per cent), in each case recording the frequency distribution of changes in the O/E values. The maximum acceptable errors in O/E due to variable measurement error were set at 0.01 for O/E_{ASPT} and 0.02 for O/E_{TAXA}. By requiring these error limits to be achievable by 95 per cent of the sites, it was possible to specify the minimum precision required for estimates of each environmental variable (Table 2.2).

Alkalinity and mean substratum particle size are the two variables requiring the greatest precision, which is not surprising as they are two of the most influential variables in the MDA. Estimating the substratum size classes and their percentage cover is known to be difficult, and thus, may be a significant source of error in estimating expected fauna. A site's discharge category, usually obtained from Environment Agency maps is based on long term historical discharge data for representative sites throughout each catchment. These maps must be read correctly, especially for small streams (categories 1–2). Discharge categories for sites are now usually obtained from Environment Agency hydrometric teams as a long-term average, based on modelling and spatial interpolation.

RIVPACS predictions require estimates of the annual averages of the temporally-varying environmental variables measured in the field, yet most vary seasonally and, to a lesser extent, between years. Therefore, taking replicate alkalinity samples or site measurements at one point, or over a short period, will not be sufficient to achieve acceptable standard errors. Mean values should be based on data from each season, while data from years known to be highly abnormal should be excluded. In practice, the extent to which all field-based variables are re-measured in each season of each year varies within the different environment agencies.

Table 2.2 Tolerable SE (or percentage SE) of estimates of the environmental variables used in RIVPACS, based on maximum acceptable errors of 0.02 for O/E_{TAXA} and 0.01 for O/E_{ASPT} for at least 95 per cent of all sites (condensed from Clarke *et al.* 1996)

Variable	Range of site values	Taxa O/E ≤0.02	ASPT O/E ≤0.01
Stream width (m)	0.3–2.0	20% SE	20% SE
	2–4	25% SE	25% SE
	4–120	30%SE	30% SE
Stream depth (cm)	4–10	20% SE	25% SE
	10–120	30% SE	35% SE
Stream slope (m km^{-1})	0.2–5.0	30% SE	35% SE
	5–75	25% SE	30% SE
Distance from source (km)	0.2–40	20% SE	30% SE
	40–203	30% SE	35% SE
Alkalinity (mg l^{-1} $CaCO_3$)	2–30	10% SE	20% SE
	30–150	15% SE	15% SE
	150–314	5% SE	5% SE
Mean substratum	-7.75:-6	SE=1.5	SE=1.5
(MSUBST in phi units)	-6:8	SE=1.5	SE=1
Discharge category (1–10)	1–2	no error allowed	
	3–10	none	±1

These computer simulations of variable sensitivity determine what should be considered tolerable standard errors of predictor variables for acceptable contributions to uncertainty in O/E. To assess the actual typical errors that occur in measuring and estimating the RIVPACS predictor variables, several research groups need to make independent estimates of each variable for a range of sites.

As part of the BAMS study (Furse *et al.* 1995), four researchers – two IFE (A and C) and two local National Rivers Authority (NRA) staff (B and D) – made completely independent estimates of each of the RIVPACS input predictor variables given in Table 2.1 for each of the 16 BAMS study sites. Prior to any site visits, each researcher was asked to read the relevant part of the RIVPACS procedures manual detailing how each of these variables should be measured and then any problems encountered on-site had to be resolved by each individual (as would be most realistic). Using appropriate 1:50000 Ordnance Survey maps of the National Grid reference, each researcher made independent measurements of altitude, slope and distance from source for each site. They also used NRA/Environment Agency discharge maps to estimate the long-term discharge category for each site. Each researcher also estimated the stream width, depth and substratum composition at each of the three RIVPACS seasons site sampling visits. All this was done as prescribed in the RIVPACS procedures manual, and interpreted and implemented within the researchers' respective laboratories.

A detailed summary and analysis of the differences between the four researchers in estimating the variables for each BAMS site is given in Chapter 4 of Furse *et al.* (1995), to which any interested reader is referred. Overall, they found that the standard deviation (SD) or coefficient of variation (%SD) was nearly always within acceptable limits, as determined by the previous simulation study summarised in Table 2.2.

The median %SD of variation between the researchers was calculated for the following variables: altitude (9 per cent); distance from source (12 per cent); slope (37 per cent); width (8 per cent); and depth (8 per cent). The high %SD for slope estimates nearly all occurred at lowland sites with little or no slope (less than 1m km^{-1}). Slope is used in its logarithmic form in RIVPACS predictions and such discrepancies at the extreme lowest possible slopes would not be expected to have any major impact on the RIVPACS predicted probabilities of site group membership. At one site where the recorded discharge category varied from one to five, the discrepancy was due to one person misplacing the true location of the site on the discharge map. Matching the field site to the correct map location and vice versa should be checked and verified by a second person as so much depends on it.

For most of the 16 study sites, the standard deviation between the four recorders for three-season average mean substratum composition was less than one phi unit, and it was always less than 1.5 phi units. The largest discrepancies occurred at four sites with fine substrates, where the researchers varied considerably in their assessment of the relative cover of sand versus silt/clay sediments.

To assess the typical overall combined effect of these real inter-personnel differences in estimating the RIVPACS environmental predictor variables for a site, the values of variables estimated by each researcher were input into RIVPACS. This generated four independent predictions of the expected values of the BMWP indices for each site. Furse *et al.* (1995) then calculated the SD of the four expected values for an index at each site. The SD in expected values for a site showed no consistent tendency to vary with the mean value. It was therefore assumed that the average of the SD across the 16 sites could validly be used as the estimate of uncertainty SD in RIVPACS expected values for the BMWP indices caused by errors in measuring the environmental variables.

Moreover, when these average standard deviations in expected values were determined for each of the seven possible season combinations for RIVPACS predictions, there was no relationship with the number of seasons. Therefore, Furse *et al*. (1995) concluded that it was appropriate to use the overall mean of the SD as the single estimate of the uncertainty SD in RIVPACS expected value for a site, regardless of the site type or number of seasons involved (Table 2.3).

Table 2.3 Average within-site standard deviation (SD_E) in expected values of NTAXA, ASPT and BMWP score, based on four researchers who independently derived estimates of the environmental variables used in RIVPACS predictions of expected values for 16 BAMS sites

Seasons combined	TAXA SD_E	ASPT SD_E	SCORE SD_E
Spring	0.60	0.083	4.7
Summer	0.32	0.077	2.8
Autumn	0.43	0.080	3.4
Spring+Summer	0.56	0.079	4.7
Spring+Autumn	0.65	0.086	5.0
Summer+Autumn	0.50	0.075	4.1
Three seasons	0.60	0.084	5.0
Overall mean	0.53	0.081	4.3

2.3 Simulating uncertainty in RIVPACS expected values of biotic indices

Having defined a quality index using a particular approach (such as RIVPACS), then an inadequate set of reference sites, not involving all relevant environmental variables and the choice of statistical prediction method or modelling technique are part of the definition of what is expected at each site. Hence, they are also part of the definition of the quality index. If the resulting quality index does not give sensible results, then it can be deemed as being a poor method for defining site quality.

It is not feasible to disentangle and determine the true quantitative errors in any estimates of expected values or reference conditions. Any such estimates of uncertainty are conditional on the availability of appropriate reference sites, their biological sample data, the choice and available estimates for environmental predictor variables, and our assumed model representation of the relationship between biota and environmental variables for reference sites.

It should always be remembered that there is no absolute truth. The uncertainty in any statistical or other approach can only be assessed using the limited information available. As such, Furse *et al.* (1995) and Clarke (2000) concluded that, at least at present, it was only feasible to include the estimated effects of errors in measuring the RIVPACS environmental predictor variables when estimating uncertainty in expected values for any monitoring site.

Thus the findings of the BAMS study (Table 2.3) were incorporated into the RIVPACS III+ software uncertainty simulations. Specifically, the uncertainty in expected values for the

BMWP indices is represented by a statistical normal distribution with the following standard deviations (SD_E):

$SD_E = 0.53$ for expected NTAXA
$SD_E = 0.081$ for expected ASPT
$SD_E = 4.3$ for expected BMWP score.

2.4 Deriving statistical estimates of sources of variance in observed index values

The sources of variation involved in determining the precision of an EQR estimate depend on the spatial and temporal scale over which the estimate of EQR is to apply. For example, if the EQR is to be an estimate of the average quality of a single sampling site for a single season, then the sampling uncertainty in the average observed index value only depends on replicate sampling variation (on the same day), variations between sampling personnel and temporal (between-day) within-season variation. (At this stage we ignore the effect of any sample processing errors.)

In contrast, if the estimate of EQR is intended to represent the average quality for the whole water body over a three-year monitoring period, then additional sources of variation need to be considered. These comprise of temporal variation between years (within the three-year period) and spatial variation between (potential sampling) sites within the water body.

Here, we describe the general statistical approach and methods for estimating these various sources of variance. Chapters 3–7 describe how estimates of variance components have been derived using a combination of existing Environment Agency datasets.

Analysis of variance (ANOVA) and hierarchical nested ANOVA techniques can be used to test for, and estimate, the various sources of variation and variance components contributing towards the total variance in the observed values of an index for each site or water body and time period. Unless the data is obtained from a well-designed balanced sampling programme, such as used in the BAMS replicate sampling study, the available datasets will usually only have replicate samples for a subset of sites on a subset of occasions. These datasets may also only have samples from more than one day per season for a subset of years and sites, and will not have samples for all years in a period for all sites. Such statistically unbalanced data make it more difficult to 'tease apart' and derive accurate estimates of each source of variance.

In our analyses, we used a combination of general unbalanced ANOVA procedure General Linear Model in the MINITAB statistics package (http://www.minitab.com) and General Mixed Model ANOVA and REML (Residual Maximum Likelihood) within the GenStat statistics package (www.vsn-intl.com/genstat/)

Variation in the observed values of a biotic index for macroinvertebrate samples from the same water body is potentially dependent on the following sources of variation (discussed in section 2.1):

- replicate samples;
- sampling personnel;
- short-term within-season temporal;

- inter-year temporal;
- spatial variation between sampling sites within a water body;
- spatio-temporal interactions.

Specifically, if X_{kijqr} is the observed value of the index for replicate sample *r* taken by operator *q* on day *j* in year *i* at site *k* in the water body of interest, then Y_{kijqr} can be expressed in terms of the sum of the components contributing towards the overall variation in its values.

$$X_{kijqr} = \mu + y_i + w_{ij} + s_k + (sy)_{ki} + (sw)_{kij} + o_{kijq} + e_{kijqr}$$

where μ = overall mean value of *Y* within the water body and (three-year) time period

y_i = deviation of mean value for year *i* from the overall mean value μ

w_{ij} = deviation of mean value for day *j* within year *i* from the mean for year *i*

s_k = deviation of mean value for site *k* from the overall mean value μ

$(sy)_{ki}$ = interaction deviation of site *k* in year *i* from expected based on site mean effect s_k and year mean effect y_i

$(sw)_{ki}$ = interaction deviation of site *k* on day *j* in year *i* from expected based on site mean effect s_k and day-year mean effect w_{ij}

o_{kijq} = deviation of operator *q* at site *k* on day *j* in year *i* from the mean for site *k* on day *j* in year *i*

e_{kijqr} = deviation of replicate *r* by operator *q* at site *k* on day *j* in year *i* from the mean for operator *q* at site *k* on day *j* in year *i*

and where

σ_Y^2 = variance of the y_i = variance due to differences between years in mean value

σ_W^2 = variance of the w_{ij} = variance due to differences between days within a year

σ_S^2 = variance of the s_k = variance due to differences between sites within a water body

σ_{SY}^2 = variance of the $(sy)_{ki}$ = variance due to spatio-temporal interaction between sites and years

σ_{SW}^2 = variance of the $(sw)_{kj}$ = variance due to interaction between sites and days

σ_O^2 = variance of the o_{kijq} = variance due to differences between operators within a site on the same day

σ_R^2 = variance of the e_{ijkqr} = variance due to differences between replicate samples taken by the same operator on the same day at the same site = replicate sampling variance.

Assumed minor interactions between operators and sites/days/years are ignored and treated as part of the replicate sampling variance.

This approach correctly estimates that part of the overall variance of index values at a certain site on a specific day that is due to systematic differences in the ways in which researchers take the sample (namely σ_O^2) from that part due to pure replicate sampling variability arising from small-scale spatial heterogeneity in fauna at the sampling station (namely σ_E^2).

The overall variance (σ_R^2) in observed index values at a site on any one day is the sum of the two components, namely:

$$\sigma_R^2 = \sigma_O^2 + \sigma_E^2 .$$

However, apart from the BAMS study, few RIVPACS sampling schemes yield any data on sampling differences between operators and so only their combined effect and variance (σ_E^2) can be estimated from replicate sampling data (assuming long-term sample data at sites are based on more than researcher).

The average total variance (σ_P^2) in index values at one site over the period is:

$$\sigma_P^2 = \sigma_R^2 + \sigma_W^2 + \sigma_Y^2$$

The average total variance (σ_{SP}^2) in index values across the whole water body over the period is:

$$\sigma_{SP}^2 = \sigma_R^2 + \sigma_W^2 + \sigma_Y^2 + \sigma_S^2 + \sigma_{SY}^2 + \sigma_{SW}^2$$

By consciously and explicitly structuring the sample data in this way within spreadsheets in statistics software, it becomes possible to use ANOVA software routines to derive estimates of the component of variance involved in the dataset variation. Additional sources of variation in the dataset, such as differences between sites in different water bodies and long-term temporal differences between three-year monitoring periods, also have to be allowed for using additional ANOVA factors/terms. This is to derive correct estimates of the above components of variance, which we then need to combine appropriately to derive estimates of uncertainty in EQRs and ecological status class (discussed in detail in Chapter 7).

The variance components are often reported and used in uncertainty simulation software in their SD form (for example, $SD_R = \sqrt{\sigma_R^2}$ denotes the overall replicate sampling SD within a site; Table 2.4).

Table 2,4 Definition of standard deviation (SD) form of each variance component

SD term		Description
SD_R	$= \sqrt{\sigma_R^2}$	Replicate sampling
SD_W	$= \sqrt{\sigma_W^2}$	Within-season temporal
SD_Y	$= \sqrt{\sigma_Y^2}$	Between-year temporal (within three-year period)
SD_S	$= \sqrt{\sigma_S^2}$	Spatial between site within water body
SD_{SY}	$= \sqrt{\sigma_{SY}^2}$	Site by year interaction
SD_{SW}	$= \sqrt{\sigma_{SW}^2}$	Site by within-season interaction

Chapters 3–6 summarise the process for estimating each of these variance components and standard deviations using a combination of UK government environment agency datasets and statistical analyses of variance.

3 Assessment of replicate sampling variability

Within each sampling site (of metres or tens or metres in length), there will still be natural spatial heterogeneity in river conditions and habitats for invertebrates. In particular, there will be local-scale variability in flows and the density and composition of macrophytes and other aquatic habitats.

3.1 RIVPACS macroinvertebrate sampling procedure and uncertainty

The RIVPACS prescribed sampling procedures require the field ecologist to make a visual assessment of the proportional cover of the different habitats available within the site, which must be a source of uncertainty. Having done this, the ecologist is required to sample the habitats in proportion to their relative abundance, spending a total of three minutes on active sampling. Sampling is usually done by kick-sampling the various sediments and disturbing the habitats and plants to catch the dislodged macroinvertebrates with a FBA pond-net of fixed mesh size – see RIVPACS procedure manual (Murray-Bligh 1997) for further details.

As the RIVPACS sampling process is standardised to a fixed length of active sampling time, it is accepted that not all species or families present at the site will be captured. A single three-minute sample typically contains 50 per cent of the species and 60 per cent of the families found overall in six replicate samples (Furse *et al.* 1981). Thus, RIVPACS samples are not intended or expected to catch all of the taxa at a site. This applies both to samples from monitoring sites and to samples from the RIVPACS reference sites, upon which the RIVPACS predictive models, and thus predictions of site-specific expected values of biotic indices, are based.

When comparing the observed fauna and observed biotic index values for monitoring sites with the site-specific expected or RC fauna and expected index values, it is crucial that 'like is compared with like'. As such, exactly the same sampling procedures and sample processing methods should be used for the both monitoring and reference site samples.

Most obviously, if one person samples for longer than prescribed (but at the normal efficiency), then they would tend to obtain and record more taxa and therefore over-estimate the true quality of the site and water body. In contrast, if staff are not adequately trained they may be inefficient at estimating and/or sampling all of the habitats present at the site and/or catching the dislodged macroinvertebrates and end up with a sample containing fewer taxa (and maybe fewer individuals) than would typically be obtained by more experienced personnel. This would lead to under-estimating the ecological quality and maybe the status of the site. If the same person repeatedly 'under-sampled' this monitoring site, then it might be incorrectly concluded that this site was of inadequate quality and that management steps were needed to improve the river quality. This highlights the potential problem of systematic sampling differences between personnel.

All of these issues concerning replicate sampling variability and inter-personnel effects were assessed within the Biological Assessment Methods (BAMS) study (Furse *et al*. 1995).

3.2 BAMS study sites and replicate sampling design

The BAMS study (Furse *et al*. 1995) was the first detailed attempt to quantify the effects of variation between replicate RIVPACS macroinvertebrate samples (variation between samples taken on the same day at the same river site). The size of the sampling variation in biotic index values for a river site can depend on the type of site and its ecological quality. As is common in many ecological studies, the more species (or individuals) present at a site, the greater the variation in number between replicate samples.

Therefore, it is important when trying to quantify sampling variation for an index to obtain data on sampling variation at a wide range of qualities and types of site. It is not sufficient to only estimate sampling variability at high quality or reference sites. For poorer quality sites, these estimates may over-estimate sampling variability for indices related to taxon richness. They may also under-estimate typical sampling variability in an index like ASPT, which will be based on fewer taxa at poorer quality sites and hence potentially be more prone to sampling variability.

The BAMS study therefore involved a carefully designed statistically-balanced sampling scheme (see Table 3.1).

A total of 16 study sites were selected in a stratified random manner from the full list of over 5000 NRA 1990 River Quality Survey (RQS) sites. Four sites were chosen from each of four contrasting RIVPACS II TWINSPAN end-group types: for each type, one site was randomly chosen from each of the then four RQS quality classes (A–D).

Each site was sampled once in the spring (March–May), summer (June–August) and autumn (September–November) of 1994, using standard RIVPACS three-minute sampling procedures (Environment Agency 1997). On each sampling occasion and at each site, four macroinvertebrate samples were collected. The first sample was taken by an IFE biologist (A), the second by a local NRA regional biologist (B), the third by biologist A again and the fourth sample by a second IFE researcher (C). Care was taken to minimise the possibility of re-sampling the same locations within the site in order to avoid progressive depletion of the fauna. Only the three samples from biologists A and B were sorted and identified; those from biologist C were kept in reserve. At any given site, the same biologists took the samples in each of the three seasons. For continuity of experience and efficiency, the same two IFE biologists sampled at each site but varied their roles as biologist A and C at successive sites. This scheme allowed the effects of between and within person sampling variation in both single and multiple season site assessments to be evaluated.

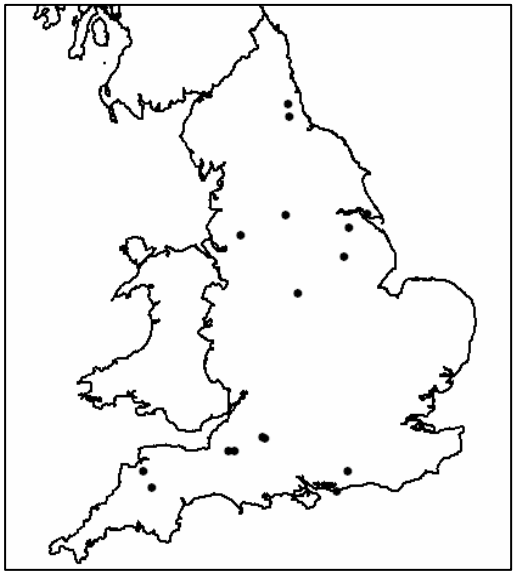
The macroinvertebrate samples were sorted and identified by experienced IFE biologists to minimise the sample processing and identification errors, which were quantified by a separate research study (Furse *et al*. 1995; see Chapter 4).

Table 3.1 Characteristics of the stratified random selection of study sites in terms of (a) ecological quality grades (A–D) as defined by range of O/E values for BMWP indices, (b) RIVPACS site group and (c) location of the full list of the 16 sites selected for replicate sampling

(a) quality grade: Range of O/E values based on:	grade A 'best' quality	B	C	D 'worst' quality
BMWP score	0.91–1.09	0.52–0.62	0.29 –0.39	< 0.18
NTAXA	0.94–1.06	0.64–0.72	0.41–0.53	< 0.30
ASPT	0.97–1.03	0.80–0.85	0.68–0.74	< 0.60

(b) RIVPACS	site group - mean value of environmental variable			
	group 3a	5b	8a	9b
distance from source (km)	15.3	8.2	11.3	33.0
width (m)	7.5	4.8	4.8	13.1
depth (cm)	19.8	21.7	32.5	77.5
altitude (m)	74	40	40	5
alkalinity (mg l^{-1} $CaCO_3$)	81	153	229	170
predominant substratum	cobbles/pebbles	gravel	gravel/sand	silt
regions of England and Wales	SW, NE, Wales	central south + midlands	east Wales to East Anglia + southern chalk streams	SE + East Anglia

(c) Site group	Grade	River name	Site name
3a	A	River Okement	South Dornaford
3a	B	River Darracott	Tantons Plain
3a	C	River Croxdale	Croxdale House
3a	D	Twyzell Burn	B6313 Bridge
5b	A	Petworth Brook	Haslingbourne Bridge
5b	B	Sheppey River	Woodford
5b	C	Sheppey River	Bowlish
5b	D	Moss Brook	PTC Bedford Brook
8a	A	Summerham Brook	Seend Bridge
8a	B	Cuttle Brook	Swarkestone
8a	C	Poulshot Stream	Jenny Mill
8a	D	Spen Beck	Dewsbury
9b	A	Old River Ancholme	Brigg
9b	B	Broad Rife	Ferry Sluice
9b	C	Skellingthorpe Drain	U/S Skellingthorpe
9b	D	Keyingham Drain	Cherry Cob



3.3 Estimation of replicate sampling variance and SD

The analysis, results and conclusions from the replicate sampling study of the BAMS sites are detailed in Clarke *et al.* (2002). The replicate sampling mean and SD of each BMWP index for each site in each season was estimated from the mean and SD of the three replicates. The sampling standard deviation of combined season samples at a site was determined from the variability in all *n* possible combinations of the appropriate single

season samples (n = 9 for each pair of seasons – spring-summer, spring-autumn and summer-autumn – and n = 27 for all three seasons combined).

3.3.1 Replicate sampling variability of NTAXA

The patterns of sampling variation within sites were very similar for each single season and for each pair of seasons, so results were combined to give overall estimates for any single season (S1) and for any two seasons combined (S2). The range and variance of replicate taxonomic richness (index NTAXA) tends to increase with the mean richness recorded for a site (Figure 3.1 a–c).

Taylor's Power Law regressions of log replicate variance against log replicate mean were used to estimate the best data transformation for equalising the replicate standard deviation for all sites (Taylor 1961, Elliott 1997). The regression slopes (standard errors in brackets) were 0.92 (± 0.26; r^2 = 22 per cent) for single season samples, 1.21 (± 0.19; r^2 = 47 per cent) for two-season combined samples and 0.94 (± 0.24; r^2= 52 per cent) for three-season combined samples (Figure 3.1 a–c). None of the slopes was significantly different from unity, which indicates that transforming to the square roots of the number of taxa (denoted by √NTAXA) should make the sampling variances independent of the number of taxa (Elliott 1977).

After transforming to the square root of the number of taxa, the replicate residual variation about the site by season mean does appear to be roughly constant for all sites and independent of the mean number of taxa present (Figure 3.1 d–f). For single season samples, there was no detectable overall tendency for sampling variation to be greater in one season than another. This was also the case for the two-season combined samples, so only the number of seasons involved in the sample is relevant (Figure 3.1).

Subsidiary influences of site quality and type

Higher quality sites have already been shown to have higher sampling variability in the (untransformed) number of taxa, as a simple consequence of greater taxon richness. However, Kruskal-Wallis non-parametric ANOVA (Siegel 1956) of the SD of replicate values of √NTAXA showed no systematic subsidiary influences ($p > 0.05$) of site quality (grades A, B, C, D) on the variability of √NTAXA (Figure 3.2). Thus all the influence of site quality on sampling variability can be determined by the observed number of taxa at the site. Kruskal-Wallis ANOVA was used to test for differences in the sampling variance of √NTAXA between RIVPACS site groups. These analyses were done separately for single, two- and three-season combined samples, and there were no cases showing any significant differences (all p>0.05).

Effects of sample order

For narrow streams especially, one might expect each sample to remove a significant fraction of the fauna and thus for taxonomic richness to decrease for subsequent samples. However, a Friedman two-way ANOVA of ranks (Siegel 1956) on sampling order (1–3) and site by season found no statistically significant (p>0.05) overall trends or differences in the number of taxa caught according to the order in which the samples were taken. This is important because it increases the validity of comparing the variation between replicate BAMS samples taken by the same person (first and third samples at each site) with those taken by different biologists (first and second samples).

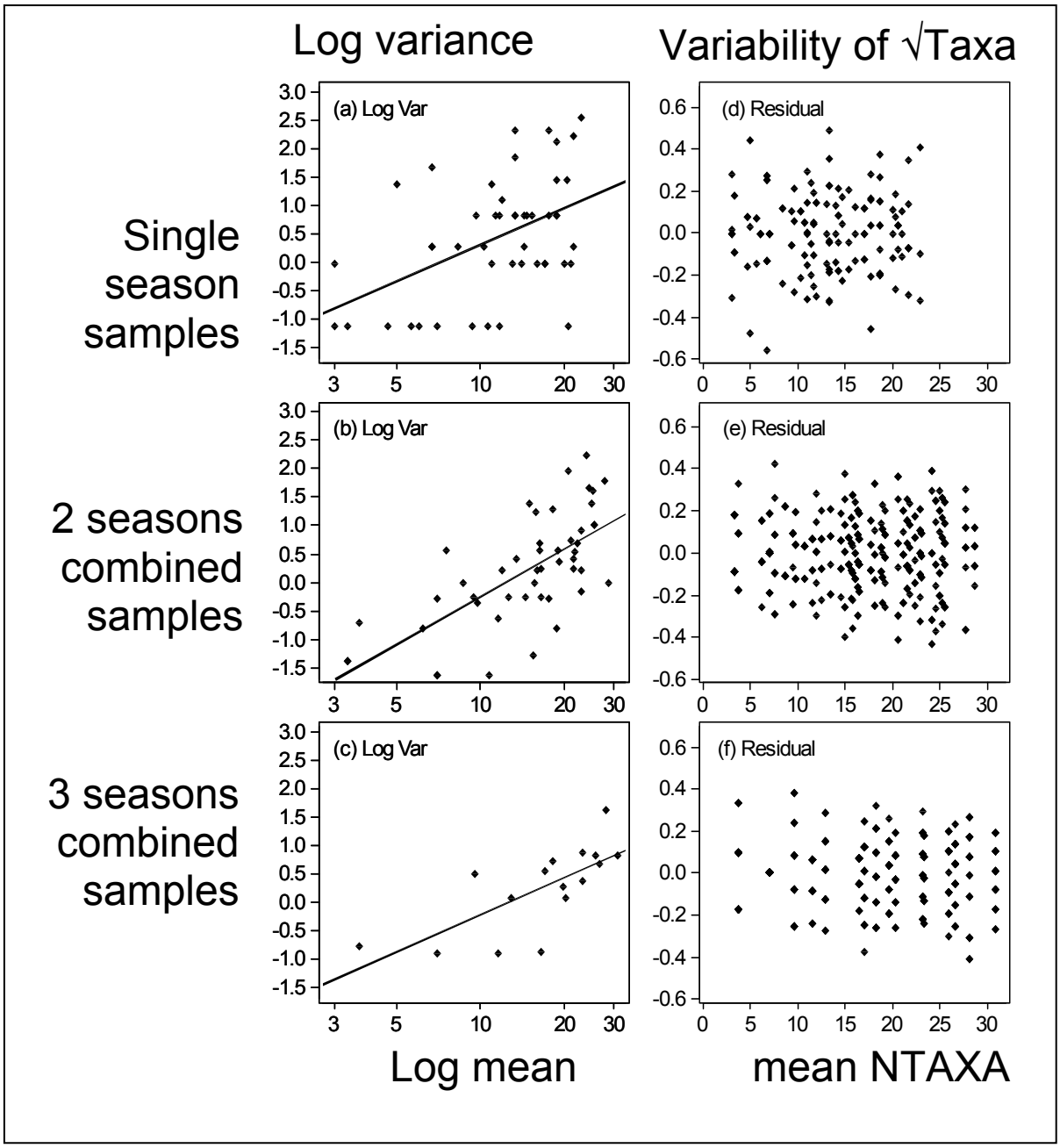


Figure 3.1 Plots of $\log_e$ variance against $\log_e$ mean of observed NTAXA in replicate samples from 16 BAMS sites for (a) single season samples (spring, summer or autumn), (b) two-season combined samples and (c) three-season combined samples, each with fitted log-log regression lines; corresponding plots (d)-(f) show residual variation of the square root of NTAXA for individual replicate samples in relation to the mean NTAXA for that site and seasonal combination

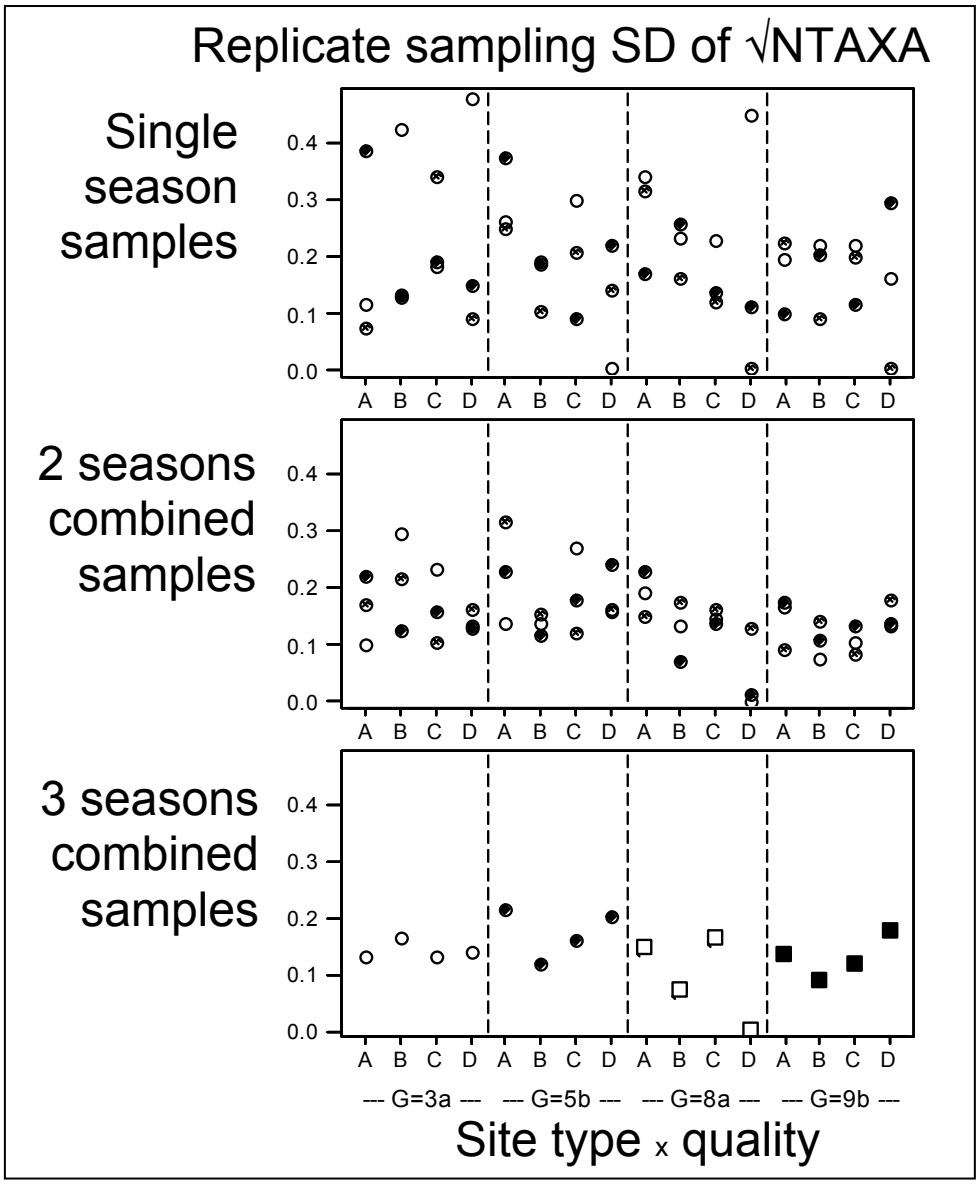


Figure 3.2 Replicate sampling SD of the square root (√) of NTAXA estimated for each of 16 BAMS sites based on (a) single season samples (o=spring, ⊗=summer, • =autumn values), (b) two-season combined samples (o=spring/summer, ⊗=spring/autumn,• = summer/autumn) and (c) three-season combined samples
Note: Sites are classified by ecological quality grade (A,B,C,D), and by RIVPACS II site group (3a, 5b, 8a or 9b) using vertical dotted lines (and also using different symbols for (c)).

3.3.2 Inter-operator effects

The effect of using different biologists on the variability in the observed BMWP index values was assessed by comparing the sampling standard deviation SD_{13}, where samples 1 and 3 are taken by the same person, with the sampling standard deviation SD_{12}, where samples 1 and 2 are taken by different biologists. In particular, if X_{ijk} is the value for sample k in season j at site i then

$$SD_{13} = \{\sum_{i=1}^{16}\sum_{j=1}^{3}(D_{ij3})^2 / 96\}^{0.5} \text{ and } SD_{12} = \{\sum_{i=1}^{16}\sum_{j=1}^{3}(D_{ij2})^2 / 96\}^{0.5} ,$$

where $D_{ijk} = \left| X_{ij1} - X_{ijk} \right|$

The percentage of the total sampling SD due to inter-operator differences was estimated by:

$$F_{pers} = 100(SD_{12} - SD_{13})/SD_{12}.$$

The results are given in Table 3.2. The overall SD of replicate values of √NTAXA for a single season, averaged across all seasons and sites, is estimated as SD_{13} = 0.217 when based on replicate samples taken by the same researcher and as SD_{12} = 0.247 when based on replicate samples taken by different biologists. As might be expected, SD_{12} is slightly higher than SD_{13}, but the percentage difference F_{pers} suggests that only about 12 per cent of the overall sampling SD is due to differences between biologists. As F_{pers} is small, most of the sampling variation is due to intrinsic variability in the fauna within the site and therefore variation in observed values between sites (or between years at the same site) is not strongly dependent on whether the same person took all the samples.

Table 3.2 Assessment of inter-operator effects on sampling variation of the square root of the number of taxa (√NTAXA), the square root of BMWP score (√SCORE) and ASPT

	√T	√S	ASPT
SD_O	0.228	0.588	0.249
SD_{13}	0.217	0.559	0.249
SD_{12}	0.247	0.612	0.259
F_{pers}	12%	9%	4%
N_{more}	20	25	20
N_{less}	19	20	24

Notes: SD_O, SD_{13}, SD_{12} denote respectively the average sampling standard deviation based on (i) all three replicate samples, (ii) samples 1 and 3 taken by the same person and (iii) samples 1 and 2 taken by different biologists. F_{pers} = percentage of overall sampling SD due to inter-operator variability. N_{more} and N_{less} are as detailed in the main text.

As a further check, out of 48 possible cases (16 sites by three seasons) the number of cases (N_{more}) where $D_{ij2} > D_{ij3}$ and the number of cases (N_{less}) where $D_{ij2} < D_{ij3}$ were counted. Where there are relatively large inter-operator effects, N_{more} is expected to be much greater than N_{less}. The difference in √NTAXA between two samples taken by the same person was as likely as not to exceed the difference between two samples taken by different biologists for the same site and season ($N_{more} \approx N_{less}$ in Table 3.2).

Thus, inter-operator effect was negligible. So providing biologists are adequately trained in field sampling procedures, the same estimate of sampling variance can be used irrespective of who takes the sample(s).

3.3.3 Overall estimates of replicate sampling SD for NTAXA

We conclude that by transforming sample values for the number of taxa to the square root scale, the amount of sampling variation is independent of both the quality and type of site. As such, the sampling SD can be assumed to be constant ($SD_{\sqrt{NTAXA}}$ in Table 3.3). Similar patterns of results were obtained for the sampling variation of two- and three-season combined samples. However, as might be expected, a smaller but constant standard deviation ($SD_{\sqrt{NTAXA}}$) applies for the square root of the number of taxa in either two- or three-season combined samples, because they are based on more information (Table 3.3).

3.3.4 Replicate sampling variance of BMWP score

Similar relationships between replicate variance and replicate mean were found for the BMWP score index (which is correlated with the NTAXA index). The same square root transformation made the replicate sample variance and SD of the square root of the BMWP score independent of its mean value, site type and site quality. Estimates of average replicate sampling variance obtained by ANOVA across all the site by season combinations are given in Table 3.3.

Table 3.3 Overall mean replicate sample variances of the square root of the number of taxa ($V_{\sqrt{NTAXA}}$), the square root of BMWP score ($V_{\sqrt{NTAXA}}$) and ASPT (V_{ASPT}), with standard errors of mean variance estimate in brackets

	mean replicate sampling variance			mean replicate sampling SD		
Seasons	$V_{\sqrt{NTAXA}}$	$V_{\sqrt{SCORE}}$	V_{ASPT}	$SD_{\sqrt{NTAXA}}$	$SD_{\sqrt{SCORE}}$	SD_{ASPT}
1	0.0519 (0.0078)	0.346 (0.059)	0.0618 (0.0120)	0.228	0.588	0.249
2	0.0269 (0.0030)	0.175 (0.021)	0.0259 (0.0043)	0.164	0.418	0.161
3	0.0211 (0.0030)	0.130 (0.016)	0.0194 (0.0072)	0.145	0.361	0.139

Notes: The mean variance estimates are calculated separately for single and two- and three-season combined samples; $SD_{\sqrt{NTAXA}}$, $SD_{\sqrt{SCORE}}$, SD_{ASPT} are the corresponding estimates of replicate sampling SD.

3.3.5 Replicate sampling variance of ASPT

The replicate sampling SD for observed sample ASPT appears to be independent of both the mean observed ASPT and the observed number of taxa at the site (Figure 3.3). It might be thought that the ASPT value observed for a site would be more variable when the ASPT was based on few taxa. Figure 3.3 d–f shows scatter plots for the sampling SD of ASPT against the average number of taxa on which the ASPT values for that site and season were based. On average, the SD does not tend to decrease systematically as the number of taxa on which it is based increases. However, there is a tendency for the estimates of the SD for ASPT to be much more variable when based on fewer taxa (Figure 3.3 d–f) and when the mean ASPT is low (Figure 3.3 a–c). This is especially true for single season estimates of SD based on only three replicate values. For variation in ASPT, Taylor's Power Law regression slopes (± SE) were 0.26 (± 0.88), 0.28 (± 0.72) and -0.44 (± 1.01) for one, two and three seasons (all r^2 < 2 per cent) respectively.

Thus, there is no consistent tendency for sampling variation in ASPT to depend on either the value of ASPT or the number of taxa. This is because samples with low ASPT values are generally of poor quality with only a few low-scoring taxa. Although ASPT values are more volatile when based on few taxa, this is counter-balanced by the reduced variability in the BMWP scores of those taxa present compared with those at high quality, taxon-rich sites.

There were no detectable consistent differences in the sampling variance of ASPT according to site quality (A, B, C, D) or season(s). There was some evidence that the sampling variance of ASPT was greater for sites in groups 3a and 9b, but these differences were only significant for two-season combined samples. Such inconsistencies and the wide variation in estimates of sampling SD for sites within the same group suggest that the sampling variance of ASPT may be adequately represented by the same constants for all types and qualities of sites. The only variable being whether ASPT is based on single season samples or two- or three-season combined samples (Table 3.3).

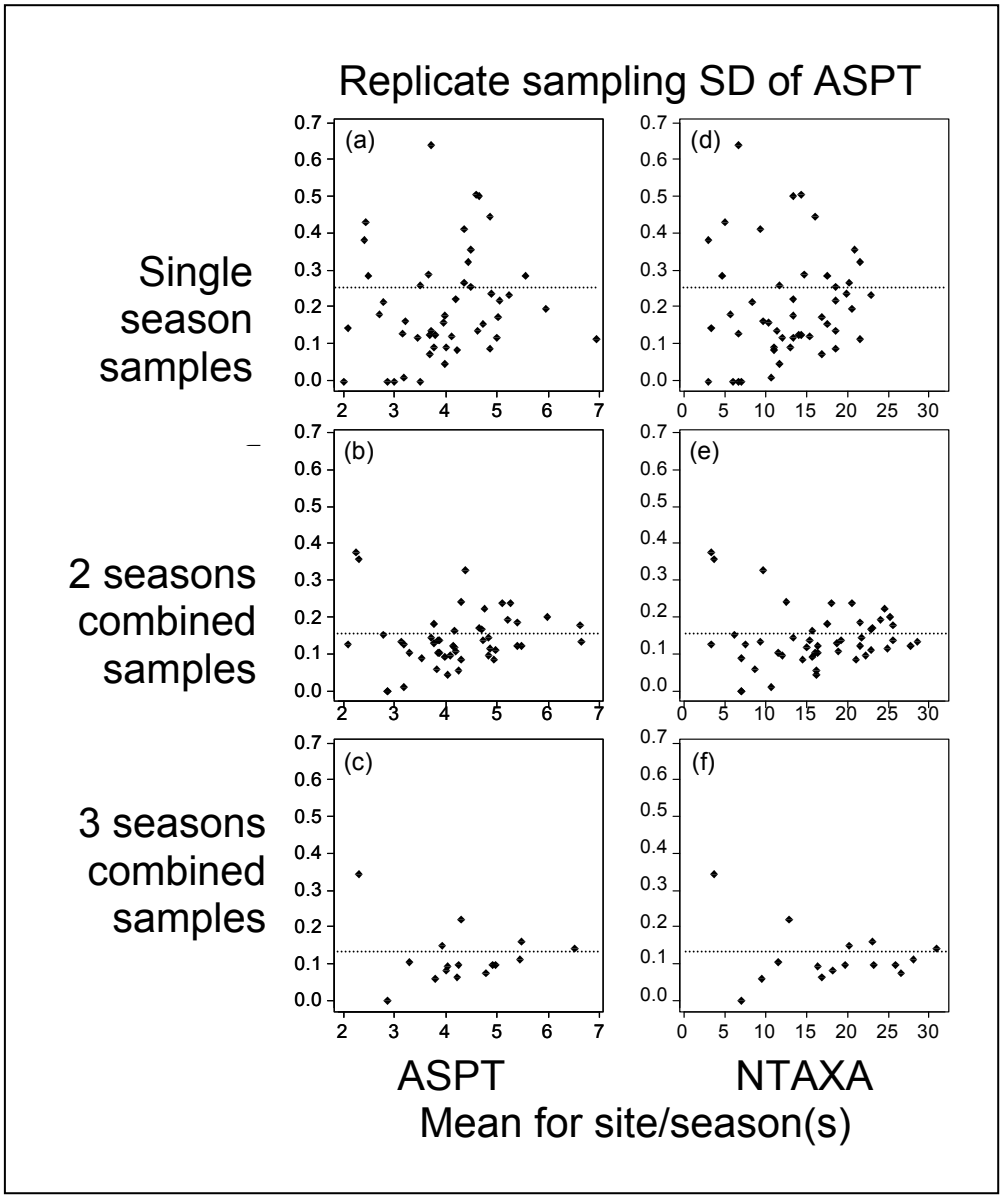


Figure 3.3 Replicate sampling SD of ASPT versus mean ASPT (a)–(c) and mean NTAXA (d)–(f) observed in replicate samples from each of 16 sites for (a) single season samples (spring, summer, or autumn), (b) two-season combined samples and (c) three-season combined samples

The SD of ASPT based on replicate samples taken by different individuals (SD_{12} = 0.259) was only marginally higher than that based on replicate samples taken by the same biologist (SD_{13} = 0.249). The estimated percentage (F_{pers}) of overall sampling variation due to inter-operator sampling effects was only 4 per cent (Table 3.2). The difference in ASPT between samples one and three (taken by the same biologist) was actually greater than the difference between samples taken by different biologists in over half of cases (Table 3.2). This suggests that using different researchers to undertake the sampling in different seasons, years or at different sites has little influence on the value and precision of ASPT, providing that they have been properly trained.

3.3.6 Summary of BAMS replicate sampling estimates

In summary, on the basis of the BAMS dataset analysis, replicate sampling SD of the square root of the NTAXA index, the square root of the BMWP score and untransformed values of the ASPT index can each be estimated by a constant, regardless of the site type or quality. These values only decrease according to whether the indices are based on single season samples or two- or three-season combined samples.

The estimates given in Table 3.3 can be used in uncertainty simulations (Chapter 7) to assess the effect of replicate variability on the uncertainty of EQR values and the confidence of status class based on one or more of these indices. Indeed, the estimates in Table 3.3 are currently used in the RIVPACS III+ (and RPBATCH) software as estimates of sampling variation in the observed values of the original BMWP indices (square root transformed where appropriate).

The detailed statistical analyses of the BAMS replicate samples dataset described above was used to determine the appropriate transformation to be used for each of the original BMWP indices. These transformations were then used as the optimum scale on which to analyse and estimate the other temporal and spatial sources of variance in index values, which are discussed in Chapters 5 and 6. It seems logical that all the other variance components for an index are likely to be more consistent across stream types and periods when their index values are transformed in the same way as was optimal for estimating average replicate sampling variance.

Recent re-analyses of the BAMS dataset in SNIFFER project WFD72C involved both the original BMWP indices and the newly-derived revised weighting Walley-Hawkes BMWP indices, both un-weighted and weighted for abundance forms. Because the original BMWP indices are highly correlated with their revised forms amongst the BAMS samples, the square root transformation was also found to be optimal for the revised BMWP forms of NTAXA and BMWP score.

Replicate sampling variability in other indices, such as Lotic Invertebrate index for Flow Evaluation (LIFE) and Acid-Water Indicator Community (AWIC), were also assessed in the SNIFFER WFD72C project. The patterns were slightly more complex, highlighting the fact that the patterns and consequences of sampling variability in each new index used in bioassessments need to be understood and assessed (see Chapter 7 for further discussion).

4 Sample processing and identification errors, audit and biases

When comparing the observed fauna and observed biotic index values for a monitoring site with the site-specific expected or RC fauna and expected index values, it is crucial that 'like is compared with like'. As such, exactly the same sampling procedures and sample processing methods and efficiency should be used for both the monitoring and reference site samples. This was first highlighted in section 3.1 in the context of the importance of having standardised field sampling methods and as little variation as possible in field sampler efficiency. The field sampling must therefore be based on clear procedures manuals and be conducted by trained staff.

Comparing like with like is also crucial in the context of RIVPACS sample processing back in the laboratory. This is especially the case for sub-sampling, sample sorting, finding and identifying the taxa present, and maybe counting or estimating their abundances. It is important that the same procedures are used to a similar level of efficiency.

4.1 Sub-sampling

Some macroinvertebrate and other bioassessment methods are based on analysing only a sub-sample of the original field sample. The general idea behind collecting more sample material than you can process in the laboratory is to increase the spatial coverage and hence the average of spatial heterogeneity within your field sample. This is done by mixing the sample up thoroughly and then processing only a manageable, more cost-effective sub-sample of all of the material and individuals. However, any such sub-sampling will lead to increased uncertainty.

For RIVPACS samples, the whole sample is sorted and processed to try to find every taxon present. However, from the point of view of estimating family-level abundances, only a fraction of the whole sample may be processed to produce counts of the very common taxa (these counts are then multiplied up to give an estimated abundance for the whole sample). This is a potential source of uncertainty, although it will not affect any indices (such as the original BMWP indices) that are just based on the presence or absence of families within the whole sample.

Other sample methods and sample processing protocols involve only sorting through a sub-sample of the whole sample collected in the field. For example, the STAR-AQEM (STAndardisation of River classifications-Assessment system for the ecological Quality of streams and rivers throughout Europe using benthic Macroinvertebrates) method used in Germany and some other European countries involves spreading a sample out onto a sorting tray comprising a grid of six by five cells and only processing a minimum of five of the 30 cells (Hering *et al.* 2004). Clarke *et al.* (2006b) showed that this laboratory sub-sampling method caused roughly as much variation in the observed values of the metrics for a site as the effects of replicate sampling variation in the field.

In the USA and Australia, a common sample laboratory procedure is to count a fixed number of individuals (200–300), regardless of how many individuals exist in the whole field sample (Ostermiller and Hawkins 2004).

4.2 Sample sorting and identification errors

In sorting the material in a sample from a test site and identifying the taxa, some taxa may be missed or misidentified by less experienced staff. What is crucial, however, is that, as always, 'like is compared with like'. This means that the sorting and identifying of taxa in any observed sample for a monitoring site needs to be done to a comparable level of efficiency and accuracy as the samples from the reference sites used to develop the predictive model (RIVPACS or otherwise) upon which the predictions of the site-specific expected fauna and index values are based.

If monitoring site samples are processed less accurately than the reference samples, then a greater proportion of taxa will be missed. As a result, the O/E ratio for indices involving any form of taxon richness (such as BMWP NTAXA and BMWP score) will be biased under-estimates of the true EQR and ecological quality of the site. This could lead to a systematic under-estimation of the ecological status of the site and water body. Missing or misidentifying an unknown and variable number of taxa also adds to the uncertainty (in addition to bias) in the estimates of EWR values and ecological status.

In the development of UK RIVPACS, all of the reference samples used to develop the RIVPACS model predictions were sorted and the taxa identified by (the same core group of) very experienced FBA/IFE/CEH staff. These staff have, through long experience and testing, been shown to justify their status as experts in UK-wide macroinvertebrate taxonomy. Therefore, it is likely that very few families and species, especially BMWP families, were missed or misidentified in processing the samples for the RIVPACS reference sites.

Whatever the accuracy of current IFE/CEH staff, it is important that the samples from any UK government environment agency monitoring sites are processed to the same level of accuracy, or that auditing procedures are used to assess and quantify the sample processing efficiency of environment agency staff relative to the IFE/CEH experts.

To assess and control for these potential sample processing biases, in the early 1990s the then NRA (which was a forerunner of the Environment Agency) and the IFE set up a quality assurance procedure. This involved IFE/CEH staff re-examining randomly chosen RIVPACS samples from each agency laboratory after initial sorting and identification by agency staff. The UK environment agencies also run their own internal quality assurance scheme to help improve and maintain standards within each laboratory.

Samples for audit by IFE/CEH were selected internally by each of the agencies being monitored. The biologists processing these samples had no prior knowledge of which samples were to be audited. The manner of sample selection, which biologists would be monitored and the number of audit samples from each season were left to the discretion of the agencies, within the limits of the total number of samples that IFE/CEH was contracted to audit.

For each sample to be audited, the UK government environment agency concerned provided CEH with a list of the BMWP families recorded as being present in the sample, a vial containing representative individuals from each found family, and the whole remaining preserved sample. In the audit, the taxa in the supplied vial were re-identified and all differences noted, including both missed and misidentified taxa. IFE/CEH audit staff then identified all the taxa in the supplied remaining sample. Families found by IFE/CEH staff but not recorded by the agency as present are referred to as 'gains', while 'losses' were families recorded as present by the agency but not found by IFE/CEH staff. The 'gains' minus 'losses' represents the net under-

estimation of the number of BMWP families present in the sample, referred to as the sample bias.

This auditing process by CEH has continued up to the present time. Each year for each region and laboratory of each agency (Environment Agency, SEPA and NIEA), CEH provides standard reports of the individual families missed and the families misidentified in each audited sample, together with the net under-estimation (bias) of the number of BMWP families present in the audited samples. The average bias for an agency laboratory, area or region is a measure of the how well it is maintaining or improving RIVPACS sample processing standards. The detailed information on which taxa tend to be most frequently missed or misidentified helps show where additional training is needed.

4.3 Estimation of sample processing biases and implications for uncertainty

Part of the BAMS project remit (Furse *et al.* 1995) was to try to develop procedures to quantify the effects of sample processing errors on biases in the observed values of the BMWP indices, O/E ratios and estimates of uncertainty in ecological quality. To do this, a stratified random subset of NRA samples was selected. These samples had either been audited in 1990 (the fist year of auditing) or 1992 (when efficiency had improved), ensuring that the samples covered a range of classes of recorded taxon richness (1–10, 11–20, 21–30, >30) spread over the three RIVPACS seasons and across all 10 NRA regions.

Overall, this BAMS study found that the NRA staff missed 15.3 per cent of all family occurrences in 209 samples audited in 1990. But in 1992 only 8.3 per cent of all family occurrences were missed by NRA staff for 211 samples audited.

In these early quality audit (QA) years, the families missed by the NRA in over 25 per cent of the samples in which they were present were Dendrocoelidae, Valvatidae, Physidae, Planorbidae, Hydrophilidae, Scirtidae, Psychomyiidae, Hydroptilidae, Goeridae, Lepidostomatidae and Brachycentridae. The most frequently missed taxa (over 20 times amongst 209 samples in 1990) were Hydrobiidae, Lymnaeidae, Planorbidae, Sphaeriidae, Hydrophilidae, Elmidae and Hydroptilidae, but this is partly because these taxa are widespread.

The BAMS study concluded that the number of taxa incorrectly recorded as present in a sample (‘losses’ averaged 0.26 per sample in 1992, or one family per four samples) is negligible compared to the number of taxa missed (‘gains’). However, in all analyses of the impact on the BMWP indices, the net effect or bias (‘gains’ minus ‘losses’) was assessed.

4.3.1 Bias in recorded number of taxa

When auditing began with the 1990 samples, many regions were missing, on average, three or four taxa per sample. In the first spring season, up to eight or nine families were missed in three regions, with 15 taxa missed from one sample.

After national improvements in the NRA's sample processing procedures, the average net under-estimation of the number of taxa in the 1992 samples was reduced to 2.0 or less in all regions. Except for the Thames region, where a lapse in the quality of sample processing in autumn 1992 led to an average of four taxa being missed per autumn sample.

Around the same period, the Water Research Centre (WRc) was commissioned to devise a statistical quality control scheme for sample processing and auditing for an agreed tolerable under-estimation rate of an average of two families per sample (van Dijk 1994). This scheme has continued to be used by the UK government environment agencies in their internal quality control for RIVPACS sample processing.

In the BAMS study of biases, it was assumed that improvements made since the QA start-up year of 1990 would be maintained at around 1992 levels. As such, further analyses and procedures were based on analyses and relationships derived from the 1992 audit data that broadly met the quality control scheme target.

One might expect there to be a tendency for more taxa to be missed in samples containing more taxa. However, the UK environment agencies do not know how many taxa there really are in a sample, only having their own estimates. Therefore, to be of use to the agencies, analyses were undertaken to assess whether there is a relationship between under-estimates in the number of taxa and the agencies' own estimates of the number of taxa.

The average under-estimation of the number of taxa in samples, grouped according to the NRA estimate of the number of taxa in each sample in 1992, is given in Table 4.1. For all classes of the NRA estimated number of taxa (except the class 21–25 taxa), the NRA under-estimated the number of taxa by no more than one taxa in at least 50 per cent of the samples (the median in Table 4.1 is one). This was encouraging, however, because several taxa are missed in a few samples, the statistical mean number missed is higher than one (range 1.0–1.9 in Table 4.1).

Table 4.1 Under-estimation of NTAXA in a sample in relation to the NRA's estimate of NTAXA for samples audited in 1992

NRA estimate of number of taxa in sample	Samples	Under-estimation of number of taxa			
		Mean	SD	Median	Maximum
1–5	4	1.0	1.2	1	2
6–10	21	1.4	1.6	1	5
11–15	32	1.2	2.1	1	7
16–20	52	1.2	1.4	1	5
21–25	62	1.9	2.0	1.5	8
26–30	27	1.5	1.4	1	4
31–38	13	1.5	1.8	1	5

There was no firm evidence that the average under-estimation of the number of taxa was strongly correlated with the NRA's estimate of the total number present. Even where the NRA only recorded 1–5 taxa (n=4 samples), the average under-estimate was still 1.0, compared to 1.5 in samples where the NRA recorded over 25 taxa (n=40 samples).

The under-estimate may be slightly higher than elsewhere in samples where the NRA recorded intermediate taxonomic richness (21–25 taxa). Where the NRA recorded over 30 taxa, the number missed was never more than five taxa. This pattern has some logic to it, in that the NRA is likely to have recorded their very highest values for number of taxa in samples where they did not miss many. There was no statistically significant ($p>0.05$) linear or quadratic relationship between the extent of net under-estimation and the number of recorded taxa.

From these analyses within the BAMS study project, it was concluded that the bias (net under-estimation of the number of taxa present) for an agency sample is not dependent

on the taxonomic richness of the sample. This means that the bias for any particular agency laboratory in any one year can be estimated by the average bias (mean 'gains' minus 'losses') of the externally-audited samples from that laboratory and year. The only exception is if the agency records five or less taxa in a sample, where the recommendation is to assume that the average bias is always 1.0 taxa.

CEH now provides Environment Agency, SEPA and NIEA laboratories, areas and regions with annual QA reports summarising the processing errors for each audited sample and giving the average net gains. These estimates can be utilised as user-supplied input parameters for the sample processing biases provided for in the RIVPACS and RICT software packages assessments of uncertainty in EQRs and status class (see Chapter 7).

The number of taxa missed per sample is not constant, but will vary from sample to sample in an unknown manner for the majority of (non-audited) samples. This represents another source of uncertainty in the observed index values, EQR and status class estimates. The WRc quality control scheme (van Dijk 1994) is based on an assumed statistical Poisson distribution for the number of missed taxa.

Furse *et al.* (1995) assessed the fit of a Poisson distribution to the overall frequency distribution of sample processing biases for all 1992 audited samples, which combines regions with slightly different biases (Poisson mean values in this context). The researchers recommended that a Poisson distribution be assumed for assessing the effect of sample processing errors within any uncertainty simulation software, with the Poisson mean parameter set equal to the appropriate QA estimate of bias for that year and laboratory/area/region.

4.3.2 Bias in ASPT and BMWP Score

A tendency for UK government environment agency biologists to miss certain taxa in a sample will lead to some under-estimation of the observed NTAXA and observed BMWP score (which can only increase with the addition of the scores of missed taxa). However, such sample processing errors may not lead to any general bias in the estimates of the observed ASPT. A sample ASPT value could potentially increase or decrease after correction for the missed taxa, depending on whether the missed taxa have a higher or lower average BMWP score than the average score of the taxa found in the sample by the agency.

To assess this effect, Furse *et al.* (1995) calculated the difference between the ASPT value for a sample based on the audit-corrected taxa list and the ASPT value derived from the taxa recorded as present in the sample by the NRA, for each of approximately 200 audited samples from 1990 and 1992. The difference (audit-corrected ASPT minus original agency ASPT) is referred to as the 'ASPT bias' for a sample, but it could potentially be either positive or negative.

The median ASPT bias was positive (or zero in one case) for every NRA region in 1990, when the number of taxa missed was higher for most regions. In 1992, when the number of missed taxa was generally lower, the median ASPT bias was zero in six of the 10 regions. It only exceeded 0.02 in the Thames region, where there was a lapse in the accuracy of sample processing in the autumn of 1992.

There appeared to be no obvious relationship between the size of the ASPT bias and either the agency's recorded ASPT value or the number of BMWP-scoring taxa in the sample on which the ASPT value was based (Tables 4.2 and 4.3). The only exception might be when the recorded taxa list gives a value for observed ASPT of over 7.0, which is more likely to be an over-estimate of the true sample value (mean ASPT bias equals -0.17 in Table 4.2).

Table 4.2 Under-estimation bias of ASPT for samples audited in 1992 in relation to the NRA recorded value of ASPT for the sample

NRA estimate of ASPT	Samples	Under-estimation of ASPT (bias)				
		Mean	SD	Median	Min.	Max.
≤ 3.0	4	0.07	0.17	0.00	0.00	0.29
3.01–4.00	40	0.07	0.14	0.00	-0.10	0.58
4.01–5.00	55	0.07	0.19	0.00	-0.51	0.48
5.01–6.00	59	0.03	0.15	0.00	-0.37	0.46
6.01–7.00	48	0.02	0.15	0.00	-0.58	0.33
> 7.00	5	-0.17	0.19	-0.10	-0.59	0.00

Table 4.3 Under-estimation bias of ASPT for samples audited in 1992 in relation to the NRA recorded number of BMWP-scoring taxa for the sample

NRA estimate of number of taxa in sample	Samples	Under-estimation of ASPT (bias)				
		Mean	SD	Median	Min.	Max.
1–5	4	-0.05	0.33	0.00	-0.51	0.29
6–10	21	0.05	0.16	0.00	-0.23	0.58
11–15	32	0.00	0.14	0.06	-0.21	0.43
16–20	52	0.04	0.18	0.00	-0.37	0.30
21–25	62	0.06	0.13	0.00	-0.58	0.48
26–30	27	0.07	0.11	0.05	-0.12	0.46
31–38	13	0.04	0.11	0.01	-0.16	0.28

4.4 Procedures to adjust for sample processing errors in observed values of BMWP indices

4.4.1 Single season sample adjustments

A very complicated way to correct for bias would be to take the site-specific RIVPACS expected probabilities for each taxa occurring and select the missing taxa using these probabilities. However, this would only be appropriate for reference quality sites. For poor quality sites, the taxa missed are much more likely to be low BMWP scoring taxa, rather than simply the taxa that were most expected to be present at the site (if it was unstressed). Furse *et al.* (1995) recommended the following practical solution, which

has been incorporated into the RIVPACS and RICT uncertainty simulation software components.

Based on an analysis of the 1992 audited samples, the under-estimation of the BMWP score (U_S) was on average about nine and the under-estimation of NTAXA (U_T) was on average about 1.5. This implies that the overall average BMWP score of missed taxa is about six. However, if the ASPT value (U_A) of the missed taxa (which equals U_S/U_T) is plotted against the number of taxa (N_T) recorded as being present by the NRA, then U_A tends to be less when few taxa are recorded (Figure 4.1). Furse *et al.* (1995) also found that the variance in the ASPT of the missed taxa decreased with the number of missed taxa (U_T) (variance = $2.0/U_T$).

A simulated (U_{Ar}) of the ASPT of the missed taxa for any single season sample from a monitoring site is adequately generated using the best fit linear regression in Figure 4.1. This is combined with a data-based error structure, as follows:

$$U_{Ar} = 4.29 + 0.077\ N_T + Z\ 2/U_{Tr}$$

where U_{Tr} is the simulated (Poisson deviate) number of missed taxa in simulation *r* (any simulated U_{Ar} values outside the BMWP range 1–10 are reset to the limits).

From this equation, the mean ASPT of the missed taxa in a sample is estimated to range from around 4.5, when about five taxa are recorded as present, to over 6.5 when over 30 taxa are recorded.

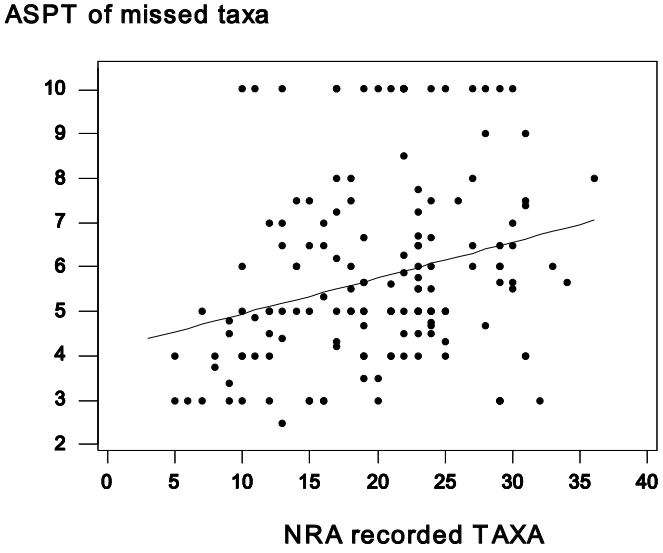


Figure 4.1 The ASPT of the missed taxa plotted against the number of taxa recorded as present by the NRA for 1992 samples with missed taxa (n=154), with best fit regression line

If O_{Tr}, O_{Sr} and O_{Ar} denote the values for the observed values of NTAXA, BMWP score and ASPT for the r^{th} simulation (after allowing for sampling variation), then the r^{th} simulated bias-corrected values (O_{Tbr}, O_{Sbr}, O_{Abr}) of the three indices are

$$O_{Tbr} = O_{Tr} + U_{Tr}$$

$$O_{Sbr} = O_{Sr} + U_{tr} \,.\, U_{Ar}$$

$$O_{Abr} = O_{Sbr} / O_{Tbr}.$$

Thus, we simulate bias-correction terms for both the number of taxa missed and their ASPT, multiply these together to simulate the bias in BMWP score, determine the bias-adjusted simulated values of the BWMP score and NTAXA, and finally use the ratio to obtain a simulated bias-corrected ASPT value. This maintains the internal consistency of simulated values for the three related indices, as the BMWP score should always equal NTAXA multiplied by ASPT.

Combining these values across all regions for 1992 reveals that the missing of taxa in the sample processing appears to lead, on average, to ASPT being underestimated by 0.00–0.04. But the actual effect varies considerably between samples, with a standard deviation of 0.16 (Figure 4.1).

4.4.2 Combined season sample biases and adjustments

Taxa missed or misidentified by UK government environment agency biologists in their laboratory when analysing a sample from a monitoring site in one season may be recorded as being present in a sample taken at the same site in a different season of the same year. Only taxa missed or misidentified in each season's sample contribute to the bias in index values for the combined season sample. This means that the biases in index values estimated from appropriately audited single season samples cannot be simply added together to obtain estimates of combined season sample biases.

The following types of missed taxa are least likely to be 'recovered' in combined season samples.

- Taxa of low local abundance, which are unlikely to be captured in more than one season at a site.
- Taxa that, by virtue of their life cycle, are seasonal in their availability for capture in pond-net samples and hence not likely to be caught in all seasons.
- Taxa that the NRA has most trouble in identifying within a sample and hence tend to miss in every sample.

For national surveys, the UK government environment agencies have always based estimates of site quality on O/E ratios derived from either two-season combined samples (typically spring+autumn) or three-season combined samples. However, because the samples selected for external QA are chosen in a random manner, there are very few occasions where samples from the same site were audited in more than one season of any one year. Therefore, a practical solution was needed.

Based on a detailed analysis of the QA datasets, Furse *et al.* (1995) found that, overall, 45 per cent of the taxa missed by the NRA in single season samples in 1990 were found and recorded as being present in a second season's sample from the same site. The average proportion of taxa occurrences missed in a single season sample in 1990 was 0.153. So if Q is the overall proportion of taxa missed in one season that were present (but not necessarily recorded) in a second season's sample from the same site, then:

$$0.45 = (1 - 0.153)\ Q, \text{ and hence } Q = 0.53.$$

For more typical QA years, such as 1992, where the overall proportion of taxa occurrences missed was only 0.083, the probability of taxa missed in one season's sample being recorded as present in a second season's sample from the same site is estimated by:

$$(1 - 0.083)\ Q = (1 - 0.083)\ 0.53 = 0.49.$$

Therefore, Furse *et al.* (1995) recommended that for 1992 and subsequent years, only 51 per cent of the taxa estimated, or simulated, as missing from any single season's sample should be assumed not to have been recorded as being present in a second season's sample. Hence, only this proportion of taxa influence errors in the two season combined sample values of the BMWP indices for that site.

Overall, 63 per cent of all taxa occurrences in the single-season samples audited in 1990 were found and recorded as being present by NRA staff in samples collected from the same site in at least one of the other two seasons. Therefore, only an estimated 37 per cent of taxa missed in the samples from the individual seasons contributed to the bias and the under-estimation of BMWP index values for the three-season combined samples.

Furse *et al.* (1995) recommended the following procedure to correct for bias in combined season sample BMWP index values. M_1, M_2 and M_3 are the QA-based estimates of average under-estimation in the number of taxa (NTAXA) for spring, summer and autumn of the year and the laboratory/area/region appropriate for a monitoring site. Then the expected average (Poisson) under-estimation of NTAXA for corresponding spring-autumn combined samples is:

$$U_{Tr} = 0.51 (M_1 + M_3);$$

while for three-season combined samples, the expected (Poisson) mean under-estimation of NTAXA is estimated by:

$$U_{Tr} = 0.37 (M_1 + M_2 + M_3).$$

All of these recommended numerical procedures for trying to correct for sample processing errors and biases in monitoring site samples have been incorporated into the simulation algorithms within the RIVPACS and the replacement new RICT software. Chapter 7 of the RIVPACS III+ and RPBATCH User Manual contains precise details of all the equations used in the uncertainty simulation procedures.

4.5 Effects of sample processing errors and biases on other biotic indices

The effects of sample processing errors are complex and will depend on the type of index. Missing taxa will lead to an under-estimate of any index, which always increases if extra taxa are added to the sample on which the index value is based. Examples include measures of taxonomic richness (NTAXA), most (if not all) measures of taxonomic diversity (such as the Shannon-Wiener index) and indices based on the total score of all taxa present (where taxon scores may be based on their perceived tolerance to a particular stress, such as the BMWP score or the Walley-Hawkes revised BMWP score index). However, as seen for BMWP and ASPT, the effect of sample processing errors on biases in the values of other indices is more ambiguous, with errors able to cause such indices to decrease or increase in complex ways.

Examples include any average-score-per-taxon type index where taxa are assigned scores according to their perceived ability to withstand particular stresses. The index is then the average score of the taxa found in the sample, perhaps with taxon scores either weighted (such as Walley-Hawkes weighted ASPT, denoted WHPT) or dependent on taxa abundances (such as LIFE; Extence *et al.*1999).

No analyses have yet been conducted for any indices other than the original BMWP indices. The current SNIFFER WFD72C project (Work Element 5-6 report) assumes that it is reasonable to use exactly the same bias-adjustment procedures developed for

the original BMWP indices (described in section 4.4) for the un-weighted form of the Walley-Hawkes-revised BMWP score, TAXA and ASPT indices.

At present, the general CEH QA scheme does not make any assessment of the extent of sample processing errors in counting or estimating the abundances of individual taxa in the sample. Therefore, there are currently no estimates of the effects of sample processing errors on biases and uncertainty in abundance-based biotic indices. Schemes auditing estimates of taxon abundances and the derived estimates of the impact of sample processing errors on abundance-based metrics are needed if new abundance-related indices, such as WHPT, LIFE or AWIC, are to be included in river WFD assessments.

At present, the RICT software for non-BMWP indices and the STARBUGS (see Clarke and Hering 2006) software for all indices have been programmed to allow the user the option of including sample processing biases on any index. This is done through the addition of an extra stochastic term dictated by the user-supplied estimate of a simple mean and standard deviation of the bias of each index (assuming a normal distribution).

5 Temporal variability – within season and between years

5.1 Requirement to assess temporal variances

In the previous versions of the RIVPACS software (RIVPACS III+ and RPBATCH), all of which were developed prior to 2005, the assessment and simulation of uncertainty in estimates of the observed values of indices only allowed for variation due to replicate sampling variability and sample processing errors. This is discussed in detail in Chapters 3 and 4 respectively.

This limitation was acceptable when the aim was to quantify the uncertainty in the estimates of EQI and quality class for a single sampled site at one point in time. However, the 'Compare' procedure of these previous versions of the RIVPACS software was also used to try to assess whether a real change in quality had occurred at a site at two different points in time. In such cases, the uncertainty should have allowed for potential additional short-term temporal variability in EQI values arising from variability between days and weeks within a (RIVPACS) season. Both the observed and, to a lesser extent, expected values could change within a season, especially following some change in environmental conditions or stress.

It is important to understand that the systematic between-season temporal variation in observed index values does not need to be included in the uncertainty in the observed values and thus the EQI values. This is because between-season variability is incorporated into RIVPACS EQI values by setting season-specific expected index values for each site, where the expected value is based on the single, two- or three-season combined samples corresponding to the observed 'season' index sample.

With the introduction of the WFD, the UK government environment agencies are considering changing from the use of data from a single year to the use of macroinvertebrate data from multiple years for status assessment within monitoring programmes. Specifically, SEPA intends to base its site status classifications on up to three years' worth of sample data, in order to reflect the longer term underlying condition of the biology. For each metric, SEPA will use the average of the EQR values for each of the individual years for which data are available over the three-year period of interest. Thus, class is defined for a three-year period, but does not necessarily require three separate years of data. If only one year's spring and autumn combined sample data were used, it would still give an estimate of the three-year mean condition. Three years' data would, however, give a more precise estimate. Measures of uncertainty and confidence of class are needed for these new measures of three-year average quality, based on an average EQI/EQR for a site. (Spatial variance between sites within a WFD water body also needs to be included, but this is assessed in Chapter 6.)

Real temporal variance implies real differences over time in the average sample biota and the average index values at a site. Within-season temporal variance is measured by the variation in the observed index values between samples taken on different days and weeks within a season. This variance is over and above that expected from the fact that samples taken at different times will also vary, just because of variation between any replicate samples. Doing this necessitates having one or more datasets with biotic index values for a combination of replicate samples and with other samples taken on different

days within any one season for each of a range of sites. The estimation of the various variance components is made using the ANOVA techniques described in section 2.4.

Estimating the variance in index values due to real differences between years within a three-year monitoring period requires some replicate sampling. As well as samples from different days within a season and, of course, samples from different years within any three-year period, all ideally at each of a range of sites.

5.2 Datasets used to estimate temporal variances

The first ever formal attempt to quantify these temporal within-season and between-year variance components for RIVPACS macroinvertebrate sample indices was made by us as part of the recent SNIFFER WFD72C project. Robin Guthrie at SEPA supplied us with two datasets: the 28 sites TAY dataset and a 416 sites SEPA dataset.

5.2.1 TAY dataset: 28 sites from Tay River Purification Board sampled 1988–1997

This dataset was generated by biologists from the Tay River Purification Board (RPB), which is now part of SEPA. The then Tay RPB had a network of 'primary sites', mainly on larger rivers in the Tay catchment (including the River Earn) and various other rivers between the Tay and the North Esk catchment in Angus (Figure 5.1). The biological quality of the sites was generally high or good. However, four sites had some impacted invertebrate faunas and another site was intermittently hard to sample due nearby hydro-electric effects on water levels. In addition, several sites were affected by sporadic sheep dip problems in the mid-1990s.

The sites were sampled between 1988 and 1997. Four replicate samples were taken at each site on each sampling occasion in spring and autumn, and identified to at least BMWP taxonomic level. Although not all sites were sampled in all years, many sites have concurrent runs of data, especially in the five-year period from 1990 to 1994. Using all possible combinations of the four spring and four autumn replicate samples for a site in any one year, 16 spring and autumn combined season samples were generated. These samples were used to derive estimates of the replicate variance in the two-season combined samples for each index and used to determine two-season combined sample inter-year variance components (Table 5.1).

Table 5.1 Components of variability which can be estimated, or for which there is information, within each dataset (indicated by ticks)

Variability component	SD	No. of seasons involved	28 TAY sites	416 SEPA sites	16 BAMS sites	12 NI sites
Replicate sampling	SD_R	1	√		√	
		2	√		√	
		3			√	
Within-season temporal	SD_W	1		√		√
		2				
		3				
Inter-year temporal	SD_Y	1	√	√		
		2	√	√		
		3				

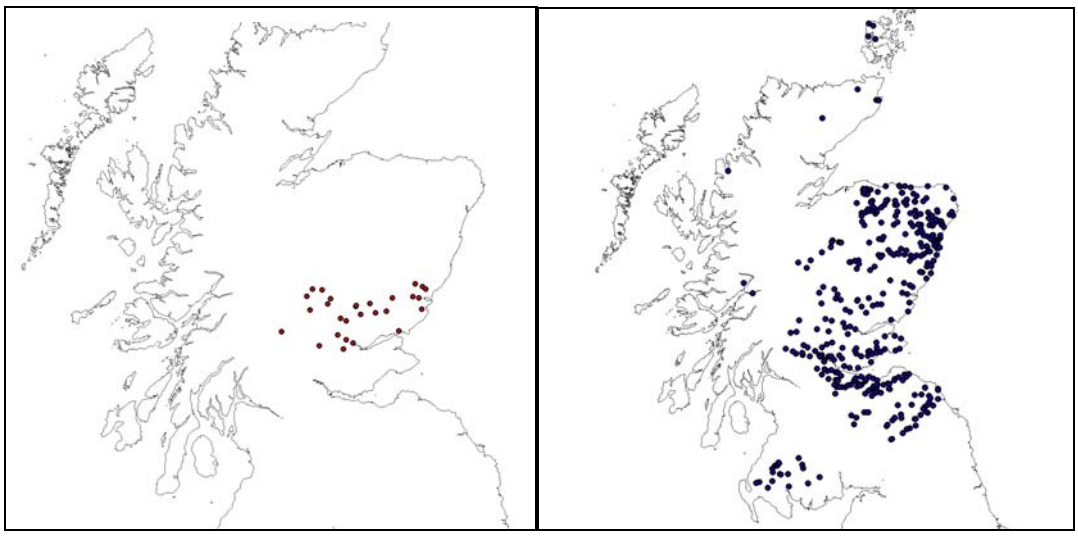


Figure 5.1 Geographic distribution of the sites used to estimate one or more biotic index variance parameters: 28 Tay RPB sites (left), 416 SEPA sites in the East and NE of Scotland (right)

5.2.2 SEPA dataset: 416 SEPA sites

This dataset comprises 416 sites predominantly in the East and North-East of Scotland. These sites cover a wide range of Scottish river types, from very large, oligotrophic rivers such as the Spey through to small, lowland streams in arable areas and rivers in predominantly urban settings (Figure 5.1).

The sites range in quality from the nearly pristine to the very severely degraded. The range of impacts includes organic pressures, hydro-morphological pressures, various toxic pressures, nutrient pressures and acidification.

The dataset was compiled from a range of databases held by the former RPBs and from SEPA's current corporate systems. Guthrie made extensive checks on sample index values and unusual patterns of variation over time, and, after checking with local biologists, he was satisfied with the data quality.

The sites were sampled between 1990 and 2004 in spring, summer and autumn (although summer samples are fewer in number as monitoring over the later part of this period tended to be based primarily on spring and autumn samples alone). From 1990 onwards, the samples were sampled and processed following the now standard SEPA methodology, with the same Analytical Quality Control/audit scheme as used in the 1990 GQA survey. Should bias correction be required, Guthrie has estimated that a figure of 1.7 net gains per sample would be appropriate, as this is consistent with the current overall SEPA performance.

Examining the 416-site dataset reveals that there are 181 occasions where the same site was sampled in the same year and in the same season but on different days. Typically, either two spring samples or two autumn samples were taken (replicate summer samples are much rarer). This data allowed us to derive estimates of the average within-season temporal variance in index values based on single season samples (Table 5.1). But a subset of only four of these 180 site-year combinations was insufficient for deriving reliable estimates of within-season temporal variance of two-season combined sample index values.

Section 5.4.2 explains how within-season temporal variance estimates for two-season and three-season combined samples were derived by inference from the relative size of replicate sampling variance for single, two-season and three-season combined sample index variances.

Only one sample was taken at each site on each sampling date, so this dataset had to be combined with the 'TAY' and/or the BAMS datasets in any ANOVA. This was done to eliminate replicate sampling variability from the observed variation in index values over time.

5.3 Consistency of replicate variability across datasets

5.3.1 Comparison of BAMS and TAY datasets

Replicate sampling variability in the square root of NTAXA values and ASPT for single season samples was found to be broadly similar for the BAMS and TAY datasets. Even though the BAMS sites were selected to cover the full range of site qualities, while the TAY sites tended to be high quality and taxa rich (Figure 5.2). This suggests that it may be acceptable to combine datasets in order to combine various temporal and spatial scales of information on variability. This will allow all the variance components to be estimated simultaneously using statistical ANOVA techniques (Section 2.4).

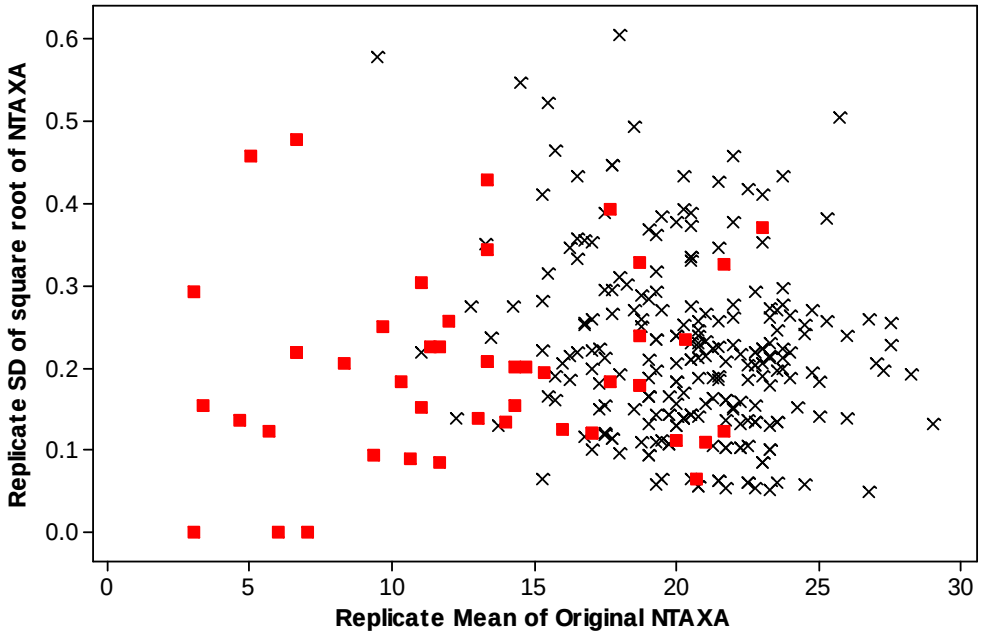
At present, national monitoring survey assessments of the quality of UK river sites for a year are usually based on RIVPACS EQI values for spring and autumn combined season samples. The WFD72C combined dataset analyses provided the first opportunity to check whether replicate sampling variability for such annual assessments was broadly similar across different types of site and region.

Replicate sampling variability in the square root of NTAXA and ASPT for two-season combined samples was similar for the BAMS and TAY datasets. Although the estimates of variability in ASPT for poor quality sites with average sample ASPT values less than four are themselves imprecise, being dependent on the stochastic capture of relatively few taxa (Figure 5.3).

Overall, we concluded that it was valid and appropriate to combine the two datasets and base estimates of replicate sampling standard deviations (SD_R) on a weighted average of the two datasets (Table 5.2). The overall estimates were taken as an average of estimates for the two datasets weighted by the number of sites involved, namely 16 BAMS and 28 TAY sites.

ANOVA estimates based on simply combining all the sample data from the two datasets would be overly-dominated by the TAY datasets, because the same sites were sampled over several years. As the aim was to derive estimates of variability that could be applied to any site, letting each site contribute equally to the overall estimate of typical replicate sampling variability was deemed the best approach.

(a) NTAXA



(b) ASPT

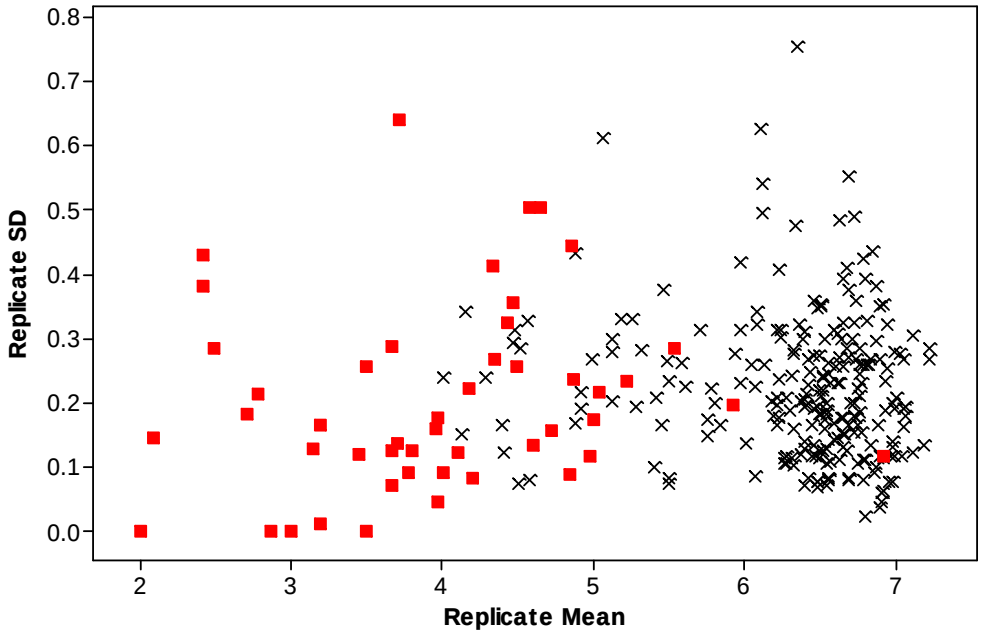
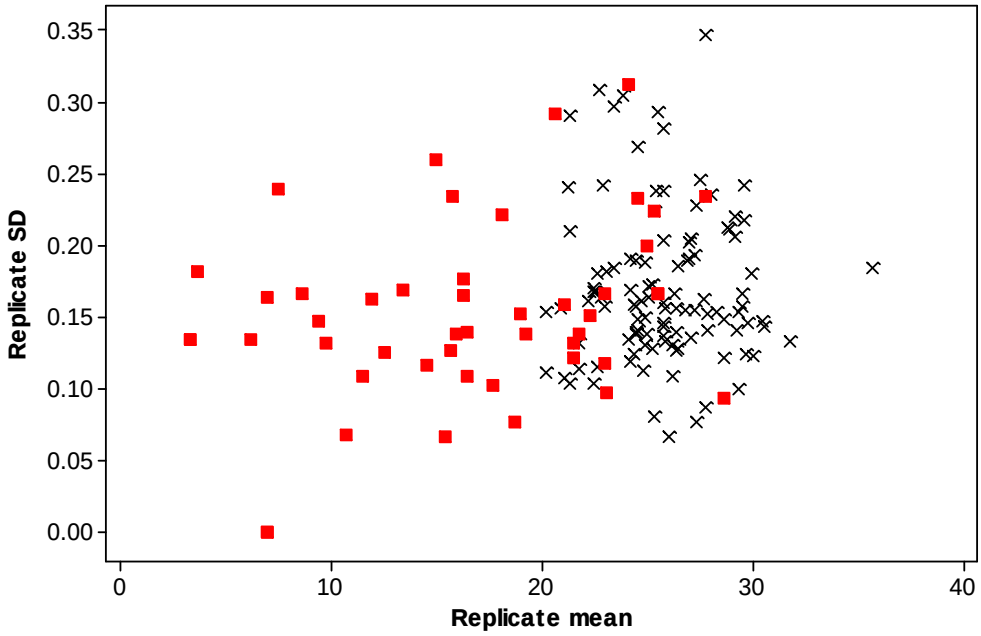


Figure 5.2 Plot of the relationship between replicate sampling SD and mean of the replicate single season sample values for all available combinations of sites and seasons with replicate sampling for the 16 BAMS sites (■) and the 28 Tay sites (x) for the original BMWP

(a) NTAXA 2-season combined



(b) ASPT 2-season combined

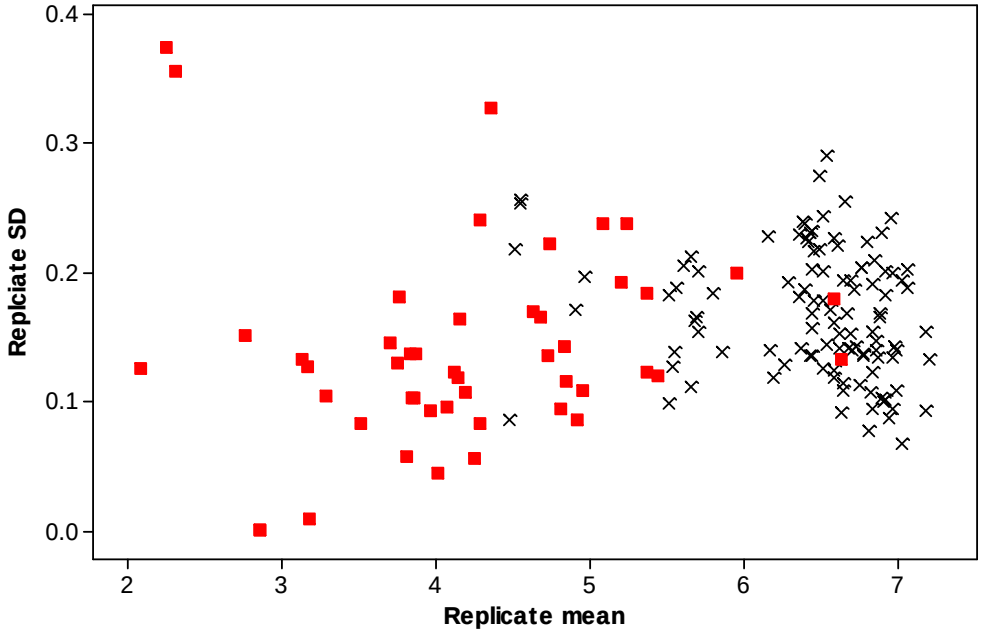


Figure 5.3 Plot of the relationship between replicate sampling SD and mean of the replicate two-season combined sample values for all available combinations of sites and seasons with replicate sampling for the 16 BAMS sites (■) and the 28 Tay sites (x) for the original BMWP

5.3.2 Replicate sampling variability for revised BMWP indices

Future UK river assessments based on macroinvertebrates and RIVPACS may involve using revised versions of the three BMWP indices, as developed by Walley and Hawkes (1997) and refined within the recently completed SNIFFER project WFD72A. The revised BMWP indices are in two forms: one involves an abundance-weighted

scoring system for families and one uses a non-weighted form that depends only on the presence or absence of families.

Estimates of replicate sampling standard deviations in these new indices (square root transformed for the revised NTAXA and BMWP score indices) were derived in the same ways as for the three original BMWP indices, using a weighted average of the estimates for the BAMS and TAY datasets (Table 5.2). This was done to pre-empt the potential change in the use of indices in RIVPACS bioassessments and to prepare for assessing uncertainty using the new indices and the derived EQRs,

The original NTAXA and BMWP score indices and their revised counterparts are highly correlated (within-site between-replicate correlations are all >0.92). This means that the same patterns of replicate variance increasing with replicate mean arise, and that the same square root transformations help to stabilise the variability across different types and quality of site. This allows us to use universally-applicable single estimates of replicate sampling SD (SD_R) for different sites, with the SD_R only decreasing according to whether index values are based on single season, or two- or three-season combined samples.

Equivalent estimates of average replicate sampling standard deviations were also derived for the AWIC (Davy-Bowker *et al.* 2005) and LIFE (Extence *et al.* 1999), as part of the WFD72C project. The estimates are included here for information, but with a caveat that their susceptibility to sampling variation is more complex than for the BMWP indices, depending on both the index (replicate mean) value and hence the type and/or condition of the site.

Table 5.2 Estimates of the replicated sampling standard deviation (SD_R) of indices for either single season, or two- or three-season combined samples based on a weighted average of estimates for the 28 TAY sites dataset and 16 BAMS sites dataset

Index		Transform scale	Number of seasons involved		
			1	2	3
Original BMWP	BMWP score	√	0.639	0.462	0.361
	NTAXA	√	0.238	0.173	0.145
	ASPT		0.250	0.170	0.139
Revised BMWP non-weighted	BMWP score	√	0.645	0.474	0.357
	NTAXA	√	0.243	0.179	0.146
	ASPT		0.235	0.161	0.115
Revised BMWP abundance-weighted	BMWP score	√	0.652	0.484	0.418
	NTAXA	√	0.243	0.179	0.146
	ASPT		0.278	0.201	0.225
AWIC (family level)			0.159	0.112	0.095
LIFE (family level)			0.247	0.173	0.238

Note: only the BAMS dataset has three-season combined replicate samples.

5.3.3 Conclusion on validity of combining datasets

In conclusion, for both the original and revised forms of the BMWP indices, replicate sampling variability is broadly similar across different datasets and types of site within the UK. Therefore, it is reasonable to combine those datasets with replicate sampling information with other datasets on temporal and spatial variability, in order to derive estimates of the full range of variance components.

5.4 Estimates of temporal variance components

Estimating within-season temporal variability for each index requires cases where RIVPACS samples have been taken on different dates within the same RIVPACS season (spring, summer or autumn). Such cases are only available for the 416 site SEPA dataset, where there are 181 situations where two (and in three cases three) samples were taken on different days (and usually months) within the same season of the same year.

Ideally, for these same sites and seasons, replicate samples would also be taken on the same day, so that we could easily 'subtract' away the variability caused by the fact that any two replicate samples will vary to some extent. However, the 416-site SEPA dataset does not have any same-day sample replication. Therefore, to estimate the variance due to real within-season temporal variability, we need to analyse the 416-site SEPA dataset in combination with the other datasets. We could just have combined the SEPA dataset with the TAY dataset, as both datasets comprise sites in Scotland that might be expected to make the sampling variability more similar. However, the 416 SEPA sites cover a much wider geographical and environmental range than the 28 Tay sites (Figure 5.1). For this reason, it was considered best to combine the SEPA dataset with the 16 BAMS sites dataset. Even though, as mentioned before, the 28 Tay sites were sampled in more years and thus carry far greater weight in determining the overall estimate and replicate sampling standard deviation.

In the future, the aim is make assessments of site ecological status based on the average quality over a three-year period. The uncertainty inherent in these estimates when all three years are not sampled will depend on the inter-year variance in index values due to the difference between years in the (unknown) average index values for each year. Therefore, we need to derive an estimate of inter-year variance parameters for three-year periods rather than over all years sampled at each site within the datasets. This was done by coding the years into three-year periods as follows: 1987–89, 1990–92, 1993–95, 1996–98, 1999–2001, 2002–04.

The statistical estimation of parameters was carried out using a hierarchical ANOVA model with the following variance components:

σ_R^2 = replicate sampling variance

σ_W^2 = within-season temporal variance

σ_Y^2 = between-year (within three-year period) temporal variance

+

σ_K^2 = variance due to differences between sites and season combinations

σ_L^2 = long-term inter-period variance.

The last two variance terms (σ_K^2 and σ_L^2) are of less interest, but they need to be included in the model analysing all data from all sites to adjust for and eliminate differences in average index values between all combinations of sites and seasons. For example, separate analyses were made for all single season samples together, while allowing for differences in average values for spring, summer and autumn at each individual site. Such between-season within-year variation needs to be removed in this way because RIVPACS predictions eliminate inter-seasonal differences in average index values by having separate predictions for each season of season combinations. Thus between-season variation is not part of the sampling uncertainty of observed index values or EQIs.

This hierarchical variance model was fitted using the REML directive in the GenStat statistical software package (Genstat Release 8.1, 2005), which treats all of the hierarchical sources of variation as random effects factors.

5.4.1 Single season sample estimates of temporal variances

The estimates for each variance parameter obtained from the fit of the REML to each index model are provided in Table 5.3a. The variance parameters are equal to the square of the equivalent SD parameter.

To assess the relative size of the three variance components determining the total variance of index values in a typical three-year period at a site, the components for replicate variance (σ_R^2), within-seasonal temporal (σ_W^2) and inter-year variance (σ_Y^2) are expressed as a percentage of their sum in Table 5.3b.

Replicate sampling variance generally contributes just under half of the total variance within a three-year period. The exact figure ranges from 38 per cent for the square root of the revised abundance-weighted BMWP score to 45 per cent for the square root of the original BMWP score and NTAXA. Put another way, this means that 55–62 per cent of the total variation in the BMWP index values over a three-year period at any one site is, on average, due to variation over time.

It is useful to express the within-season temporal variability as a percentage (%WT) of the total within-period temporal variance (Table 5.3). This highlights the finding that the variance estimates for short-term, within-season temporal variability are, rather surprisingly, about the same or higher than the longer-term inter-year temporal variance estimates for all BMWP indices.

This raises the issue that any additional samples taken on a later date within the same season may be more likely to have been taken from a site if it was suspected, or known, that there was either some recent problem at the site or the previous sample taken in that season was suspect. Thus, the available data to estimate within-season temporal variance may not completely typical. Moreover, it may tend to over-estimate the typical/average within-season variance, which in turn would lead to some under-estimation of the true inter-year variance components.

Fortunately, later in the WFD72C project we received a new Community Change study dataset from Tommy McDermott at the Environment and Heritage Service (which was replaced by the NIEA in July 2008) containing monthly samples over a period of one year (Feb–Jan) at each of 12 sites in Northern Ireland (NI). From this NI dataset, we extracted a sample from each of the three months in each of the three RIVPACS seasons – spring (Feb–May), summer (June–Aug) and autumn (Sep–Nov) – at each of the 12 sites. Only one sample was taken at each site on each month. It was therefore not possible to separate replicate variance from within-season temporal variability, but only to estimate their combined variance ($\sigma_R^2 + \sigma_W^2$), representing total variance within a season at a site.

We found that the estimates of total within-season variance were actually similar or higher for the NI dataset than for the combined TAY/SEPA/BAMS dataset estimates.

Table 5.3 Estimates of single season sample values (a) variance (σ^2) and (c) SD (=$\sqrt{\sigma^2}$) for replicate sampling (σ_R^2), within-season temporal variability (σ_W^2), inter-year variability (σ_Y^2), and other variance components (inter-period (σ_L^2), site by season (σ_K^2)) based on all data from the BAMS, TAY and SEPA datasets combined; (b) variance components as percentage of total variance ($\sigma_R^2 + \sigma_W^2 + \sigma_Y^2$) within three-year periods

(a) Variance	Index	σ_R^2	σ_W^2	σ_Y^2	σ_L^2	σ_K^2
Original BMWP	√ Score	0.4320	0.2957	0.2746	0.2615	2.5196
	√ NTAXA	0.0578	0.0350	0.0365	0.0291	0.2154
	ASPT	0.0654	0.0596	0.0209	0.0359	0.7859
Revised BMWP non-weighted	√ Score	0.4510	0.3682	0.3074	0.3516	2.8809
	√ NTAXA	0.0608	0.0446	0.0390	0.0391	0.2295
	ASPT	0.0617	0.0658	0.0171	0.0396	0.9748
Revised BMWP abundance-weighted	√ Score	0.4490	0.4273	0.3091	0.3916	3.2213
	√ NTAXA	0.0608	0.0446	0.0390	0.0391	0.2295
	ASPT	0.0722	0.0780	0.0304	0.0589	1.2042
AWIC		0.0269	0.0262	0.0027	0.0078	0.2035
LIFE		0.0446	0.0139	0.0222	0.0132	0.2462
(b) % Variance		$\%\sigma_R^2$	$\%\sigma_W^2$	$\%\sigma_Y^2$	%WT	
Original BMWP	√ Score	43	30	27	52	
	√ NTAXA	45	27	28	49	
	ASPT	45	41	14	74	
Revised BMWP non-weighted	√ Score	40	33	27	54	
	√ NTAXA	42	31	27	53	
	ASPT	43	45	12	79	
Revised BMWP abundance-weighted	√ Score	38	36	26	58	
	√ NTAXA	42	31	27	53	
	ASPT	40	43	17	72	
AWIC		48	47	5	91	
LIFE		55	17	28	39	
(c) SD		SD_R	SD_W	SD_Y	SD_L	SD_K
Original BMWP	√ Score	0.657	0.544	0.524	0.511	1.587
	√ NTAXA	0.240	0.187	0.191	0.171	0.464
	ASPT	0.256	0.244	0.144	0.189	0.886
Revised BMWP non-weighted	√ Score	0.672	0.607	0.554	0.593	1.697
	√ NTAXA	0.247	0.211	0.198	0.198	0.479
	ASPT	0.248	0.257	0.131	0.199	0.987
Revised BMWP abundance-weighted	√ Score	0.670	0.654	0.556	0.626	1.795
	√ NTAXA	0.247	0.211	0.198	0.198	0.479
	ASPT	0.269	0.279	0.174	0.243	1.097
AWIC		0.164	0.162	0.052	0.088	0.451
LIFE		0.211	0.118	0.149	0.115	0.496

Note: %WT = $100\sigma_W^2 / (\sigma_W^2 + \sigma_Y^2)$ = % period temporal variance due to within-season.

If it is assumed that replicate sampling SD is about the same for these 12 NI stream sites as for the average BAMS and TAY sites, then we can conclude that our estimates of within-season temporal SD based on the TAY/SEPA/BAMS combined datasets (Table 5.3c) are probably not biased. They therefore represent the best estimates for use in the new RICT software for assessing uncertainty in average site quality over a three-year monitoring period (see Chapter 7).

5.4.2 Combined season sample estimates of temporal variances

The above analysis can only be conducted properly for all single season samples. This is because sampling on more than one day within a season was only available for more than one season within the same year on four occasions, which is not enough to derive meaningful estimates of σ_W^2 for two-season combined samples.

Therefore, the estimates of the two temporal variance components, σ_W^2 and σ_Y^2, must be derived indirectly. This is done by analysing all of the two-season combined samples from all three datasets combined, and using the REML directive in the GenStat statistical package to fit a hierarchical model with the following variance components.

σ_R^2 = replicate sampling variance

σ_T^2 = total temporal variance (within three-year period) {= $\sigma_W^2 + \sigma_Y^2$}

+

σ_K^2 = variance due to differences between sites and season combinations

σ_L^2 = long-term inter-period variance.

In this case, the available data does not permit direct estimation of the two variance components, σ_W^2 and σ_Y^2}, but only their sum, σ_T^2.

Therefore, we decided it was best to derive values for the two separate components by assuming that, for any particular index, the relative size of these two variances was the same for two-season combined samples as estimated for single season samples (%WT in Table 5.3).

Specifically, from the estimate of σ_T^2 for two-season combined samples for each index, we estimate:

$$\sigma_W^2 = \%WT\ \sigma_T^2 / 100$$

$$\sigma_Y^2 = \sigma_T^2 - \sigma_W^2 .$$

For the three forms of BMWP score and NTAXA, about half (49–58 per cent) of the total temporal variability within three-year periods is, on average, within-season variability. For original and revised forms of ASPT, the equivalents percentages are 72–79 per cent (Table 5.3).

The estimates of σ_W^2 and σ_Y^2, and thus SD_W and SD_Y, derived using the above equations are given in Table 5.4.

The same logic, assumptions and procedures that were used to develop estimates of the two temporal variance components, σ_W^2 and σ_Y^2, for two-season combined sample index values were then used to derive variance estimates for three-season combined samples (Table 5.5).

Table 5.4 Estimates of two-season combined sample parameters for (a) variance (σ^2) and (c) SD (=$\sqrt{\sigma^2}$) for replicate sampling (σ_R^2), within-season temporal variability (σ_W^2), inter-year variability (σ_Y^2), and other variance components (inter-period (σ_L^2), site by season (σ_K^2)) based on all data from the BAMS, TAY and SEPA datasets combined; (b) %σ_R^2 = replicate variance as percentage of total variance ($\sigma_R^2 + \sigma_W^2 + \sigma_Y^2$) within three-year periods

(a) Variance	Index	σ_R^2	σ_T^2		σ_L^2	σ_K^2
Original BMWP	√ Score	0.2260	0.3742		0.2275	3.0147
	√ NTAXA	0.0306	0.0421		0.0262	0.2538
	ASPT	0.0299	0.0487		0.0209	0.7281
Revised BMWP non-weighted	√ Score	0.2430	0.3964		0.3096	3.4454
	√ NTAXA	0.0334	0.0476		0.0368	0.2693
	ASPT	0.0286	0.0435		0.0247	0.9288
Revised BMWP abundance-weighted	√ Score	0.2450	0.4172		0.3552	4.0317
	√ NTAXA	0.0334	0.0476		0.0368	0.2693
	ASPT	0.0369	0.0542		0.0437	1.2701
(b) % Variance		%σ_R^2	σ_W^2	σ_Y^2	%WT – as in Table 13	
Original BMWP	√ Score	38	0.1946	0.1796	52	
	√ NTAXA	42	0.0206	0.0215	49	
	ASPT	38	0.0360	0.0127	74	
Revised BMWP non-weighted	√ Score	38	0.2141	0.1823	54	
	√ NTAXA	41	0.0252	0.0224	53	
	ASPT	40	0.0344	0.0091	79	
Revised BMWP abundance-weighted	√ Score	37	0.2420	0.1752	58	
	√ NTAXA	41	0.0252	0.0224	53	
	ASPT	41	0.0390	0.0152	72	
(c) SD		SD_R	SD_W	SD_Y	SD_L	SD_K
Original BMWP	√ Score	0.475	0.441	0.424	0.477	1.736
	√ NTAXA	0.175	0.144	0.147	0.162	0.504
	ASPT	0.173	0.190	0.112	0.144	0.853
Revised BMWP non-weighted	√ Score	0.493	0.463	0.427	0.556	1.856
	√ NTAXA	0.183	0.159	0.150	0.192	0.519
	ASPT	0.169	0.185	0.096	0.157	0.964
Revised BMWP abundance-weighted	√ Score	0.495	0.492	0.419	0.596	2.008
	√ NTAXA	0.183	0.159	0.150	0.192	0.519
	ASPT	0.192	0.197	0.123	0.209	1.127

The estimates of variance parameters for index values based on three-season combined samples were lower than their equivalent variance estimates based on two-season combined samples (compare Tables 5.4 and 5.5).

From the previous BAMS study and this study (see Chapter 3), the replicate sampling variance of index values was found to be highest for single season samples and lowest for three-season combined samples. This all might be expected, as three season combined samples are based on more information, with more opportunity for taxa not captured in one sample to be found in another. Nonetheless, it is encouraging that the uncertainty parameter estimates are sufficiently precise not to mask any expected trends as the number of samples (one, two or three) involved in determining an index value increases.

Table 5.5 Estimates of three-season combined sample parameters for (a) variance (σ^2) and (c) SD (=$\sqrt{\sigma^2}$) for replicate sampling (σ_R^2), within-season temporal variability (σ_W^2), inter-year variability (σ_Y^2), and other variance components (inter-period (σ_L^2), site by season (σ_K^2)) based on all data from the BAMS, TAY and SEPA datasets combined; (b) %σ_R^2 = replicate variance as percentage of total variance ($\sigma_R^2 + \sigma_W^2 + \sigma_Y^2$) within three-year periods

(a) Variance	Index	σ_R^2	σ_T^2		σ_L^2	σ_K^2
Original BMWP	√ Score	0.1300	0.3437		0.1629	3.7391
	√ NTAXA	0.0209	0.0332		0.0194	0.3169
	ASPT	0.0194	0.0436		0.0092	0.7578
Revised BMWP non-weighted	√ Score	0.1270	0.3898		0.2211	4.2867
	√ NTAXA	0.0210	0.0406		0.0296	0.3370
	ASPT	0.0133	0.0456		0.0086	0.9695
Revised BMWP abundance-weighted	√ Score	0.1740	0.3562		0.2568	5.3482
	√ NTAXA	0.0210	0.0406		0.0296	0.3370
	ASPT	0.0504	0.0214		0.0261	1.5496
(b) % Variance		% σ_R^2	σ_W^2	σ_Y^2	%WT – as in Table 13	
Original BMWP	√ Score	38	0.1787	0.1650	52	
	√ NTAXA	42	0.0163	0.0169	49	
	ASPT	38	0.0322	0.0113	74	
Revised BMWP non-weighted	√ Score	38	0.2105	0.1793	54	
	√ NTAXA	41	0.0215	0.0191	53	
	ASPT	40	0.0360	0.0096	79	
Revised BMWP abundance-weighted	√ Score	37	0.2066	0.1496	58	
	√ NTAXA	41	0.0215	0.0191	53	
	ASPT	41	0.0154	0.0060	72	
(c) SD		SD_R	SD_W	SD_Y	SD_L	SD_K
Original BMWP	√ Score	0.361	0.423	0.406	0.404	1.934
	√ NTAXA	0.145	0.128	0.130	0.139	0.563
	ASPT	0.139	0.180	0.106	0.096	0.871
Revised BMWP non-weighted	√ Score	0.356	0.459	0.423	0.470	2.070
	√ NTAXA	0.145	0.147	0.138	0.172	0.581
	ASPT	0.115	0.190	0.098	0.093	0.985
Revised BMWP abundance-weighted	√ Score	0.417	0.455	0.387	0.507	2.313
	√ NTAXA	0.145	0.147	0.138	0.172	0.581
	ASPT	0.224	0.124	0.077	0.161	1.245

5.5 Recommended estimates of variance parameters

It is recommended that the overall estimates of the standard deviation in index values due to replicate sampling variation (SD_R) be obtained from the weighted average of the estimates for the 16-site BAMS dataset and the 28-site TAY dataset given in Table 5.2.

It is also recommended that the overall estimates of the standard deviation in index values due to within-season temporal variation (SD_W) and between-year temporal variation (SD_Y) be obtained from the estimates based on the combined TAY/SEPA/BAMS datasets for single season samples (Tables 5.3), two-season combined samples (Table 5.4) and three-season combined samples (Table 5.5).

All of these recommended estimates of standard deviation components have been collected together in Table 5.6.

Table 5.6 Recommended estimates of SD parameters of each index for replicate sampling (SD_R), within-season temporal variability (SD_W) and inter-year variability (SD_Y) for (a) single season samples, (b) two-season combined samples and (c) three-season combined samples

		Index	SD_R	SD_W	SD_Y
(a) Single season samples	Original BMWP	√ Score	0.639	0.544	0.524
		√ NTAXA	0.238	0.187	0.191
		ASPT	0.250	0.244	0.144
	Revised BMWP non-weighted	√ Score	0.645	0.607	0.554
		√ NTAXA	0.243	0.211	0.198
		ASPT	0.235	0.257	0.131
	Revised BMWP abundance-weighted	√ Score	0.652	0.654	0.556
		√ NTAXA	0.243	0.211	0.198
		ASPT	0.278	0.279	0.174
(b) Two-season combined samples	Original BMWP	√ Score	0.462	0.441	0.424
		√ NTAXA	0.173	0.144	0.147
		ASPT	0.170	0.190	0.112
	Revised BMWP non-weighted	√ Score	0.474	0.463	0.427
		√ NTAXA	0.179	0.159	0.150
		ASPT	0.161	0.185	0.096
	Revised BMWP abundance-weighted	√ Score	0.484	0.492	0.419
		√ NTAXA	0.179	0.159	0.150
		ASPT	0.201	0.197	0.123
(c) Three-season combined samples	Original BMWP	√ Score	0.361	0.423	0.406
		√ NTAXA	0.145	0.128	0.130
		ASPT	0.139	0.180	0.106
	Revised BMWP non-weighted	√ Score	0.357	0.459	0.423
		√ NTAXA	0.146	0.147	0.138
		ASPT	0.115	0.190	0.098
	Revised BMWP abundance-weighted	√ Score	0.418	0.455	0.387
		√ NTAXA	0.146	0.147	0.138
		ASPT	0.225	0.124	0.077

These estimates can be used as the corresponding uncertainty parameters in the new RICT software. This software has been coded to allow for the individual and combined effect of SD_R, SD_W and SD_Y in simulating estimated uncertainty in EQI and status class assessments based on one or more of these indices. These indices include the current GQA classification option of basing ecological status on the lowest status indicated by EQI_{TAXA} and EQI_{ASPT}.

6 Spatial (and spatio-temporal) variability of sites within water bodies

6.1 Background to sampling sites and lack of spatial replication

In the past, most assessments of rivers by the UK environment agencies and others has been at the RIVPACS site level (metres or tens of metres), even though RIVPACS sampling sites were often selected to be representative of a stretch of river several kilometres long. In particular, since the 1990 NRA RQS and the subsequent national GQA surveys, increasing attention has been given to sub-dividing rivers and catchments as efficiently as possible into relatively homogenous stretches. This is to allow the sampling and monitoring of each and all defined stretches within the practical constraint of limited resources.

Up to now, the ecological quality and status of most river monitoring stretches has been estimated and monitored using RIVPACS macroinvertebrate samples taken from a single, carefully selected site within the whole stretch. The estimated quality at this sampling site is assumed to represent the ecological quality and status throughout the entire stretch.

The advent of the WFD has formalised the requirement to sub-divide catchments and rivers into clearly-defined WFD water bodies for monitoring, reporting and management purposes. In response, over the past few years, the UK environment agencies have developed procedures to sub-divide all catchment and rivers into WFD water bodies. The Environment Agency approach to forming water bodies (discussed at the start of Section 2.1) has meant that many water bodies are now quite large and may include two or more former GQA sampling sites. Meanwhile other water bodies, especially those in upper catchments, may no longer include any former monitoring sites.

To some extent, natural spatial variability in the macroinvertebrate fauna present within a water body is allowed for in RIVPACS bioassessments based on O/E ratios of biotic indices. RIVPACS tries to allow for, and eliminate, the effects of natural variability between sites by standardising the observed fauna at a sampling site by the fauna expected at the site (if at reference site quality), based on its physical and environmental characteristics (as represented in the RIVPACS environmental predictor variables).

However, the majority of these predictor variables are static and measured from maps. They are, of necessity, fairly simple. They are also limited by the need not to involve variables already affected by the environmental stress being assessed (pollution or flow regime), or its consequences. This is because such alterations could lead to mis-setting the ‘target’ expected fauna and biotic index values for the site. In practice, RIVPACS expected values for biotic indices will usually be very similar for all sites within a single water body, partly because water bodies are usually chosen to be relatively environmentally homogeneous compared to the full range of UK river site types.

Because of the historical development of river monitoring based on single sites within river stretches, there has been little data available on, and no statistical analyses of, the typical extent of large-scale spatial variability in RIVPACS sample macroinvertebrate fauna, indices and RIVPACS O/E values between sites within water bodies.

However, many of the newly-defined WFD water bodies are an amalgamation of old monitoring stretches and encompass two or more sites from the previous GQA and other monitoring networks. Thus, there should now be historical (and perhaps continuing) data on spatial variability between RIVPACS sampling sites within a water body.

One such example is a dataset from the Dove catchment in the Environment Agency's Midlands region, which has two or three sites in each of three WFD water bodies that have been sampled over a period of years. This dataset, described in section 6.2, has been analysed for this report to provide an example of how to analyse and estimate the extent of spatial inter-site variance within a water body (section 6.3). (Ideally, estimates of typical between-site within-water body spatial variability should be based on analyses of multi-site data from a much larger number and wider range of water bodies throughout the UK – this is a recommendation for future research.)

6.2 Dove catchment dataset of spatio-temporal variability

This Environment Agency dataset consists of RIVPACS sample data and derived biotic index values for 10 sites within the Dove catchment that have been sampled in all or most years since the early 1980s. Usually, one sample was taken in spring and one in autumn each year. But sometimes only one season was sampled, sometimes all three seasons were sampled in a year and on 16 occasions two samples were taken on different days at the same site in the same season. There is no information on replicate sample variation at these sites and so this will need to be inferred from the BAMS and/or TAY replicated sample datasets. These Dove catchment sites were assumed to be subject to minimal impact and to be of fairly consistent quality over time, such that variability over time and in space was expected to be mainly 'natural'.

Most importantly for this current research, the sites are from different WFD water bodies within the Dove catchment: three sites from the Upper Dove water body, two sites from the Dove water body and three sites from the tributary River Manifold water body (Figure 6.1). This will allow us to estimate variance in index values due to inter-site differences within the same water body

6.2.1 Site-specific expected values of indices

As well as the sample biotic index values was a supplied dataset giving the RIVPACS environmental predictor variables for each site. In addition to the fixed value of the RIVPACS map-derived variables (national grid reference, altitude, slope, distance from source, discharge category), supplied alkalinity values were either constant or varied little and were averaged for each site.

The RIVPACS field-derived variables (stream width and depth and substratum composition) should ideally be measured in each of the RIVPACS seasons (spring, summer and autumn) and then their average values used for any RIVPACS predictions.

Because these variables were not always recorded in each season and year, it was difficult to derive consistent year-specific mean values and, hence, year-specific RIVPACS predictions for each site.

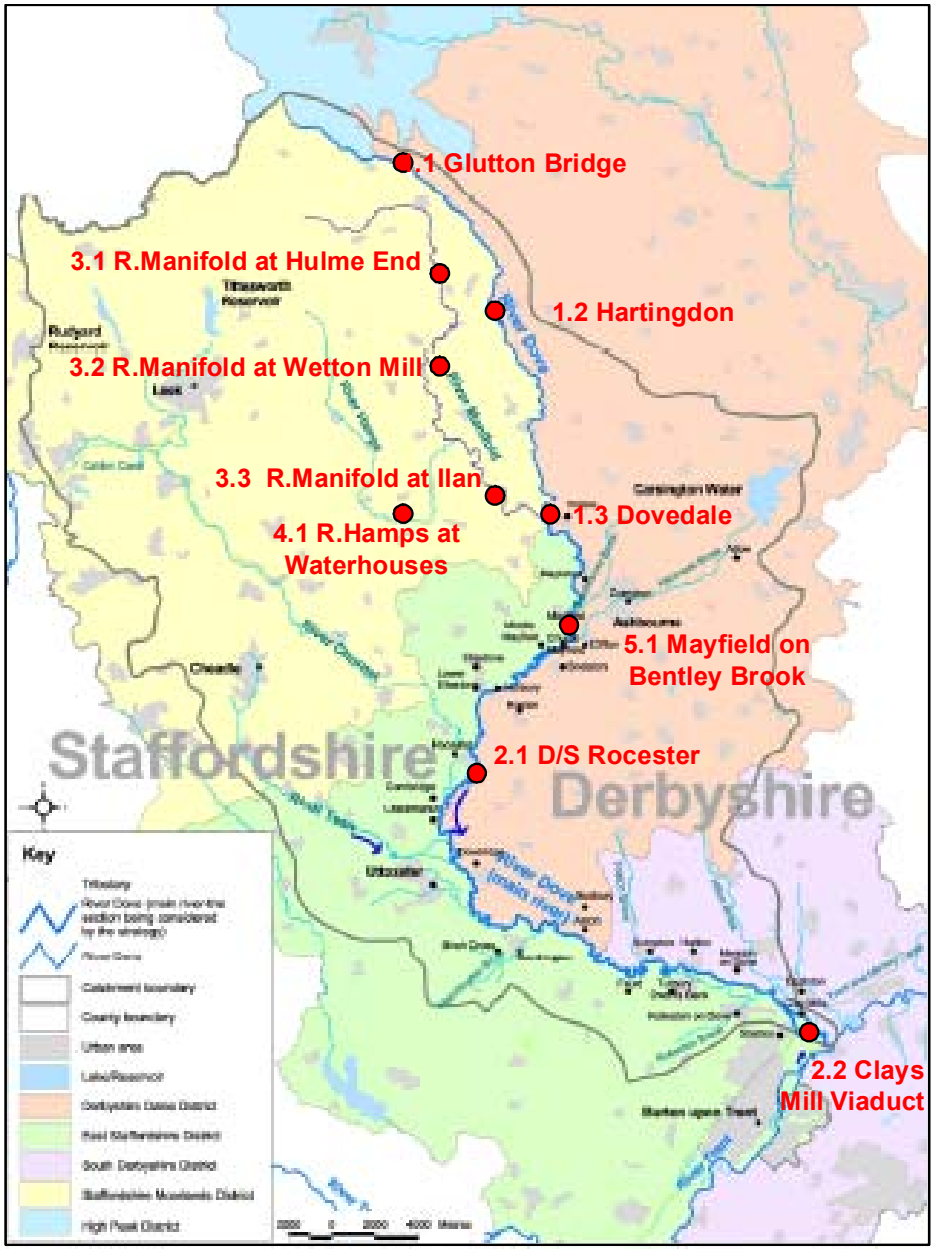


Figure 6.1 Map showing the distribution of 10 study sites over five water bodies (1–5) within the Dove catchment

Table 6.1 RIVPACS expected values of BMWP, NTAXA and ASPT indices for each season (spring, summer, autumn) for each of 10 Dove catchment sites

Water Body	Site	Expected NTAXA			Expected ASPT		
		Spr	Sum	Aut	Spr	Sum	Aut
1.UPPER DOVE	1.1 GLUTTON BRIDGE	23.9	22.2	23.6	6.51	6.20	6.27
	1.2 HARTINGTON	24.0	22.4	23.7	6.14	5.89	5.86
	1.3 DOVEDALE	23.7	22.2	23.4	5.83	5.60	5.55
2. DOVE	2.1 D/S ROCESTER	22.9	21.0	23.0	5.97	5.77	5.72
	2.2 CLAYMILLS VIADUCT	23.3	21.6	23.7	5.88	5.71	5.66
3. MANIFOLD	3.1 HULME END	24.1	22.4	23.8	6.50	6.20	6.26
	3.2 WETTON MILL	23.6	22.1	23.4	6.41	6.10	6.14
	3.3 ILAM	23.0	21.6	22.9	6.05	5.79	5.75
4. HAMPS	4.1 WATERHOUSES	24.0	22.4	23.7	6.44	6.14	6.18
5.1 BENTLEY	5.1 MAYFIELD SK162461	23.4	22.7	23.2	5.64	5.42	5.35

However, visual assessments of within-site variation in the values for these variables and spot-checks with predictions of expected BMWP index values in the Environment Agency's national 1995 GQA database for some of these Dove catchment sites, suggested it would be adequate for this study to average the values of these variables. These average values could then be used in the RIVPACS software to derive single site and season-specific predictions for the expected values of the BMWP indices for each of the 10 sites, irrespective of the year (Table 6.1). (Ironically, using such fixed predictions for each site, as done here, has been a long-term ambition for RIVPACS.)

6.2.2 Observed and O/E index variation over time

Although the Dove catchment sites were assumed to be of relatively consistent quality over time, there is considerable variation in the observed single season sample values for both ASPT and NTAXA over the 20+ years of monitoring data (Table 6.2). Part of this is due to pure inherent replicate sampling variation, which needs to be allowed for.

Table 6.2 Mean and range of observed values of BMWP, NTAXA and ASPT across all single season samples for each of the 10 Dove catchment sites

Water Body	Site	Observed NTAXA			Observed ASPT		
		Mean	Min	Max	Mean	Min	Max
1.UPPER DOVE	1.1 GLUTTON BRIDGE	18.1	11	28	6.13	5.00	6.73
	1.2 HARTINGTON	18.9	10	28	5.57	3.50	6.95
	1.3 DOVEDALE	21.7	16	27	6.41	5.71	7.17
2. DOVE	2.1 D/S ROCESTER	24.8	14	34	5.88	4.83	6.85
	2.2 CLAYMILLS VIADUCT	23.4	13	32	5.60	4.07	6.38
3. MANIFOLD	3.1 HULME END	18.7	8	31	5.87	4.50	6.77
	3.2 WETTON MILL	19.3	11	27	6.25	5.69	6.84
	3.3 ILAM	19.1	13	26	5.95	5.17	6.59
4. HAMPS	4.1 WATERHOUSES	18.7	8	27	5.65	3.75	6.86
5.1 BENTLEY	5.1 MAYFIELD SK162461	20.2	10	33	5.32	4.07	6.41

Having derived site-specific expected values for indices, we can now assess the extent of temporal variability in RIVPACS O/E values for NTAXA and ASPT for each of the 10 sites (Table 6.3, and Figures 6.2 and 6.3). The Dove catchment sites have O/E values ranging from above one to below the GQA 'b'/'c' grades boundary of 0.70, indicating considerable variation over the two decades.

Table 6.3 Mean and range of O/E values of BMWP NTAXA and ASPT across all single season samples for each of 10 Dove catchment sites

Water Body	Site	Observed NTAXA			Observed ASPT		
		Mean	Min	Max	Mean	Min	Max
1.UPPER DOVE	1.1 GLUTTON BRIDGE	0.77	0.47	1.17	0.97	0.80	1.07
	1.2 HARTINGTON	0.80	0.42	1.17	0.93	0.60	1.13
	1.3 DOVEDALE	0.93	0.68	1.14	1.13	1.02	1.24
2. DOVE	2.1 D/S ROCESTER	1.10	0.67	1.48	1.01	0.84	1.20
	2.2 CLAYMILLS VIADUCT	1.01	0.56	1.35	0.97	0.69	1.13
3. MANIFOLD	3.1 HULME END	0.79	0.34	1.29	0.92	0.70	1.05
	3.2 WETTON MILL	0.83	0.50	1.15	1.00	0.93	1.10
	3.3 ILAM	0.84	0.57	1.13	1.01	0.85	1.12
4. HAMPS	4.1 WATERHOUSES	0.79	0.34	1.14	0.90	0.61	1.07
5.1 BENTLEY	5.1 MAYFIELD SK162461	0.87	0.44	1.42	0.98	0.76	1.18

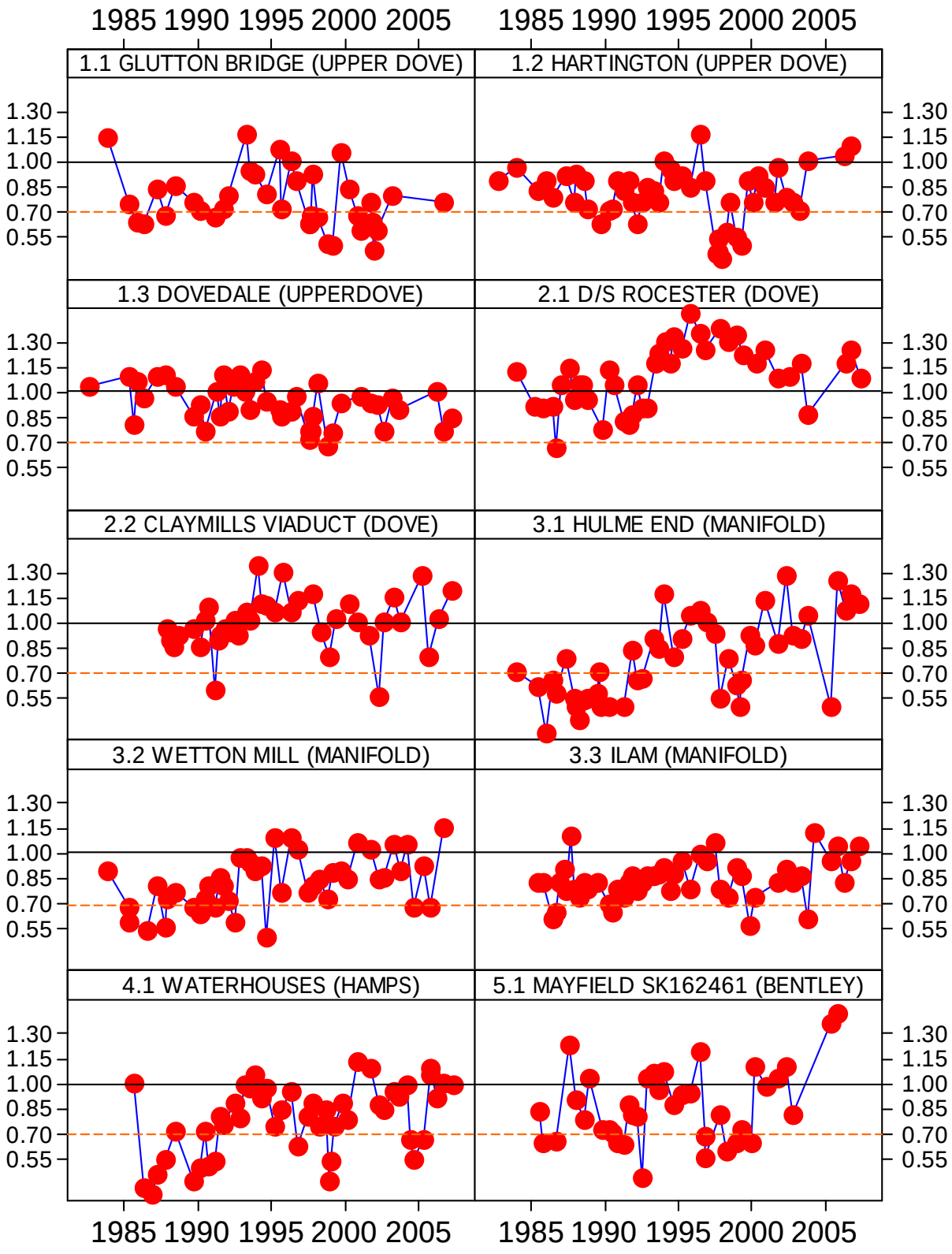


Figure 6.2 Times series plots of variation in O/E TAXA since the 1980s for single season samples from 10 sites in five water bodies within the Dove catchment (name in brackets after site name)

Note: Dashed orange line indicates O/E_{TAXA} boundary between GQA grades 'b' and 'c'.

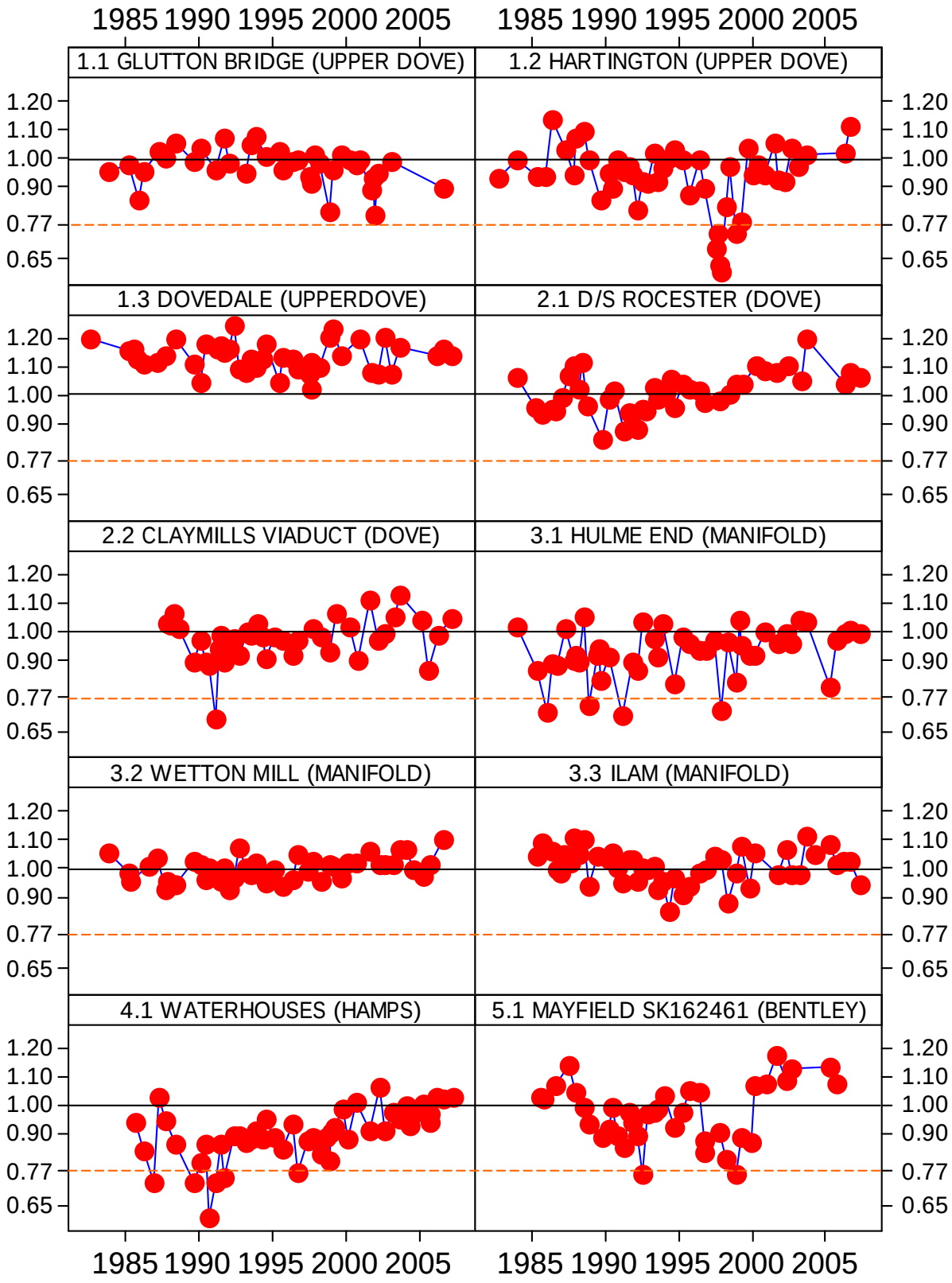


Figure 6.3 Times series plots of variation in O/E ASPT since the 1980s for single season samples from 10 sites in five water bodies within the Dove catchment (name in brackets after site name)

Note: Dashed orange line indicates O/E_{ASPT} boundary between GQA grades 'b' and 'c'.

6.3 Estimation of variance components for Dove dataset

6.3.1 Integrated simultaneous estimation of variance components

The overall variability between samples depends on a range of factors and variance components. For example, variation between single observed values for different sites in the same water body within one season of one year is partly (and perhaps mostly) due to pure replicate sampling variance. There is also an additional within-season temporal variance if the sites were sampled on different days or months within the same season. Thus, the simple standard deviation of the values from different sites within the same water body, season and year does not tell us much about the amount of variation due to real differences between sites within a water body.

Unfortunately, the Dove dataset only has one sample per season for each site. So analysing this dataset on its own does not allow us to distinguish inter-year variance effects from within-season temporal variability and pure replicate sampling variance.

Therefore, the data on the Dove sites was initially combined with the BAMS dataset of replicate sample variability, the TAY dataset with information on replicate and inter-year variability and the SEPA dataset with information on within-season and inter-year temporal variability. This was done in order to estimate all of the potential sources of variance, simultaneously and in a consistent manner, thereby correctly allowing for the effect of each component on other aspects of variability.

These integrated analyses of variance were attempted repeatedly using the ANOVA and REML techniques explained in section 2.4. However, because of the gross imbalance of factors in these combined datasets (not all sites sampled in each season or each year; BAMS sites were single sites per water body and sampled in only one year), the statistical fitting algorithms were unable to converge on a solution and identify the various variance components.

There were also estimation problems when the whole Dove dataset was analysed on its own. However, estimates could be obtained using REML when the Dove dataset was restricted to: one season's samples (spring or autumn); the three water bodies (Upper Dove (1), Dove (2) and Manifold (3)) with more than one site sampled; and the period 1990–2004.

This model estimated the average variance between sites within a water body (allowing for water body differences), variance between years within three-year periods (allowing for period differences) and the remaining residual variation. By subtracting the prior estimates of replicate variance and within-season temporal variance (given in Table 5.3) from this residual variance, estimates of the true site-by-year interaction variance were obtained (Table 6.4). Inter-site variance components differed between seasons.

The site-year interaction variance estimates were always less than the inter-site variance estimates and were zero for ASPT in both spring and autumn. This suggests that different sites within the same water body tended to change in a similar way between years, at least within any (relatively short) three-year monitoring period.

Other simpler approaches were also used to derive estimates of combinations of variances, as detailed in section 6.3.2.

Table 6.4 Estimates of average spatial variance between sites within a water body, temporal variance within a three-year period and their interaction variance, based on the single season samples from sites in the Upper Dove, Dove and Manifold water bodies within the Dove catchment over the period 1990–2004

		From REML analysis of Dove sites data			From Table 14		By subtraction
Season	Index	Sites within water body	Years within 3-year periods	Residual	Reps	Within-season temp	Site by year interaction
Spring	√BMWP	0.4563	0.1546	0.8680	0.4320	0.2957	0.1403
Spring	√NTAXA	0.0389	0.0217	0.0987	0.0578	0.0350	0.0059
Spring	ASPT	0.0706	0.0099	0.1230	0.0654	0.0596	0.0000
Autumn	√BMWP	0.2950	0.1160	0.7500	0.4320	0.2957	0.0223
Autumn	√NTAXA	0.0143	0.0000	0.0822	0.0578	0.0350	0.0000
Autumn	ASPT	0.1040	0.0035	0.1150	0.0654	0.0596	0.0000

Notes: Site-year interaction variance obtained by subtraction (residual variance minus reps variance minus within-season temp variance from Table 5.3); Negative estimates reset to zero.

6.3.2 Variation between sites within a water body for individual years

All Dove catchment samples were grouped by their combination of water body, year and season. One-way ANOVA on these groups was then used to estimate the average simple variance (VARSITES) between observed index values for samples from different sites in the same water body taken in the same season of the same year (Table 6.5). Estimates of the average true variance (σ_S^2) due to differences between sites within the same water body were then obtained by subtracting estimates of the variances due to replicate samples and within-season temporal variance (obtained from Table 5.3) from the estimate of VARSITES (Table 6.5).

Table 6.5 Average variance (VARSITES) between observed index values from different sites in the same water body taken in the same season of the same year

Index	VARSITES	Reps	Within-season temp	σ_S^2
√BMWP	1.4185	0.4320	0.2957	0.6909
√NTAXA	0.1388	0.0578	0.0350	0.0460
ASPT	0.2637	0.0654	0.0596	0.1387

Note: σ_S^2 = estimate of true variance due to site differences (equals VARSITES minus reps variance minus within-season temp variance from Table 5.3), assuming sites were (generally) sampled on different days.

The indirect estimates of between-site variance (σ_S^2) in Table 6.5 were obtained by subtracting lower-order variance components. They are higher than the corresponding more direct estimates of variances between sites within a water body in Table 6.4 for both spring and autumn samples for each of the three indices √BMWP, √NTAXA and ASPT. Thus the choice or availability of samples, sites, years and water bodies in the data used in ANOVA components all influence the estimates and inter-dependence of estimates of the various variance terms involved in the overall uncertainty of sample EQR values. This choice or availability also influences the bio-assessment for either a site or a water body over a single season or year, or multi-year period.

More reliable estimates of the average or typical spatial variance between sites within a water body are needed.

Therefore, it is recommended that further consideration and effort is made within the various UK government environment agencies to consider whether and how a single monitoring site can represent a whole water body. When only a single site is monitored, there is no direct information on spatial inter-site variability within that water body. This must therefore be inferred from inter-site variability in other water bodies for which there is such information, so that this source of uncertainty can be included in the overall uncertainty and confidence of class for water bodies with just a single sampling site.

It is recommended that the UK government environment agencies examine their RIVPACS sample databases in the light of the recent re-alignment of sites within WFD water bodies, in order to assess and extract more information on spatial variability between sites within a water body. This information can then be analysed to derive better and more robust estimates of the typical inter-site variance components.

7 Uncertainty of EQRs and confidence of status class

This section will illustrate how the uncertainty associated with any sample EQR value and derived estimate of ecological status class depends on the various sources of variance discussed and quantified in Chapters 3–6. But it also critically depends on the spatial and temporal scale over which the sample EQR value(s) is intended to represent the aquatic ecological condition.

7.1 Effects of spatial and temporal scale of bio-assessment

In this section, we return to the problem of assessing confidence of class and misclassification rates, as highlighted in section 1.3 (Figures 1.4 and 1.5). In particular, consider the case where the assessment is based on the sample value of EQR determined by BMWP NTAXA, for which variation in the observed values of NTAXA was found to be roughly constant and best-assessed on the square root scale as √NTAXA. We used an expected NTAXA value of 22 (and ignored any error in the expected value, as discussed in section 2.3).

From our variance component analyses, reasonable 'best-available' estimates of the various components for single season sample values of √NTAXA are:

σ_R^2	= replicate sampling variance	= 0.0578
σ_W^2	= temporal within-season variance	= 0.0350
σ_Y^2	= between-year (within three-year period) variance	= 0.0365
σ_S^2	= spatial between-site (within water body) variance	= 0.0266
σ_{SY}^2	= site-year interaction variance	= 0.0029

where σ_R^2, σ_W^2 and σ_Y^2 estimates are from Table 5.2 and the σ_S^2 and σ_{SY}^2 estimates are the average of the spring and autumn sample estimates in Table 6.4.

If σ_T^2 denotes the relevant total uncertainty variance in the observed value of √NTAXA, then we can simulate (or generate mathematically) the probability distribution of values of O/E_{TAXA} that could be obtained for any particular true O/E value. From this, we can then calculate the probability of obtaining O/E values for each status class and hence the misclassification rates.

7.1.1 Assessing average quality for single site on day of sampling

If the single sample O/E value is only intended to represent the ecological quality of the sampling site on the day of sampling, then the only uncertainty involved is pure replicate sampling variance:

$$\sigma_T^2 = \sigma_R^2 = 0.0578$$

and the misclassification rate for sites in relation to their true O/E_{TAXA} is shown by the purple line in Figure 7.1.

7.1.2 Assessing average quality for single site over one season or year

If the sample O/E is intended to represent the average quality at the sampling site within the same sampling season (or seasons, for example spring-autumn combined samples) for that year, then the uncertainty variance now includes the within-season temporal variance and increases to:

$$\sigma_T^2 = \sigma_R^2 + \sigma_W^2 = 0.0578 + 0.0350 = 0.0928.$$

The resultant higher misclassification rates are shown by the red line in Figure 7.1.

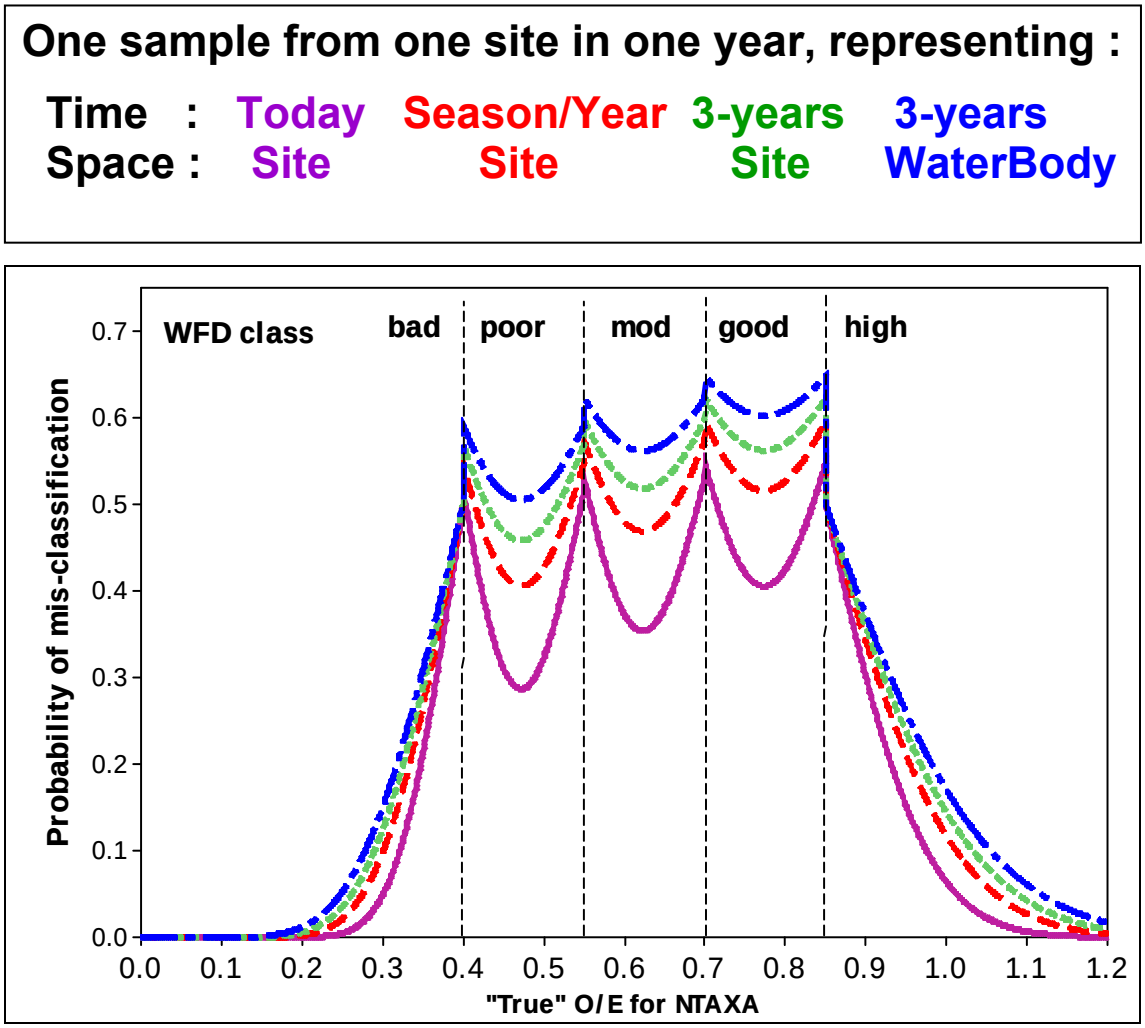


Figure 7.1 Probability of misclassifying a site based on a single season sample O/E for NTAXA in relation to its true O/E and the spatial and temporal scale over which the O/E is intended to represent the average ecological quality

7.1.3 Assessing average quality for single site over a period of years

If the single sample O/E is intended to represent the average quality at the site over a period of N_{YP} years, then the uncertainty and confidence of class also depend on the inter-year (within-period) variance. This gives a total uncertainty variance of

$$\sigma_T^2 = \sigma_R^2 + \sigma_W^2 + \sigma_Y^2(1 - N_Y / N_{YP})$$

where the term $(1-N_Y/N_{YP})$ represents, in statistical terms, the 'finite population correction factor' due to the fact that we have sampled N_Y (in this case one) of the N_{YP} years in which we are interested.

Thus, in the case of monitoring average quality over a N_{YP} = three-year period, we have:

$$\sigma_T^2 = \sigma_R^2 + \sigma_W^2 + \sigma_Y^2(1-1/3) \text{ = 0.0928 + 0.0365 (1-1/3) = 0.1171}$$

while the misclassification rates are shown by the green line in Figure 7.1.

7.1.4 Assessing average quality across a whole water body over a period of years

Finally, if the sample O/E obtained at a single sampling site is intended to represent the average quality over a three-year period across the whole water body in which it lies, then the total uncertainty variance should include an estimate of the spatial variability in index values between sites within a water body. This gives:

$$\sigma_T^2 = \sigma_R^2 + \sigma_W^2 + (\sigma_Y^2 + \sigma_{SY}^2)(1-1/3) + \sigma_S^2$$

$$= 0.1171 + 0.0029(1-1/3) + 0.0266 = 0.1456$$

while the misclassification rates for the example in Figure 7.1 are shown by the blue line.

7.2 Probability of 'moderate or worse' status class

As mentioned in Chapter 1, the current most important concern for European countries trying to implement the WFD is whether each water body is of good or better ecological status. In terms of uncertainty, it is important to have some estimate of the likelihood that the true class of a water body is moderate or worse. Perhaps only if we are for example 95 per cent confident that a water body is of inadequate moderate or worse status would it be justifiable to carry out a costly programme of measures to improve ecological quality within the water body.

Continuing with our illustrative example (with actual data-based variance component estimates), Figure 7.2 shows the probability of the true status being moderate or worse for each possible O/E_{TAXA} value for each of the spatial and temporal scenarios discussed above. In order to have at least 95 per cent confidence that a site/water body is of 'moderate or worse' status, the O/E_{TAXA} value that the sample must not exceed is 0.566 when the assessment is just for that site on the day of sampling. But this value decreases to 0.492 when the assessment is for the average quality of the whole water body over a three-year period (see Figure 7.2).

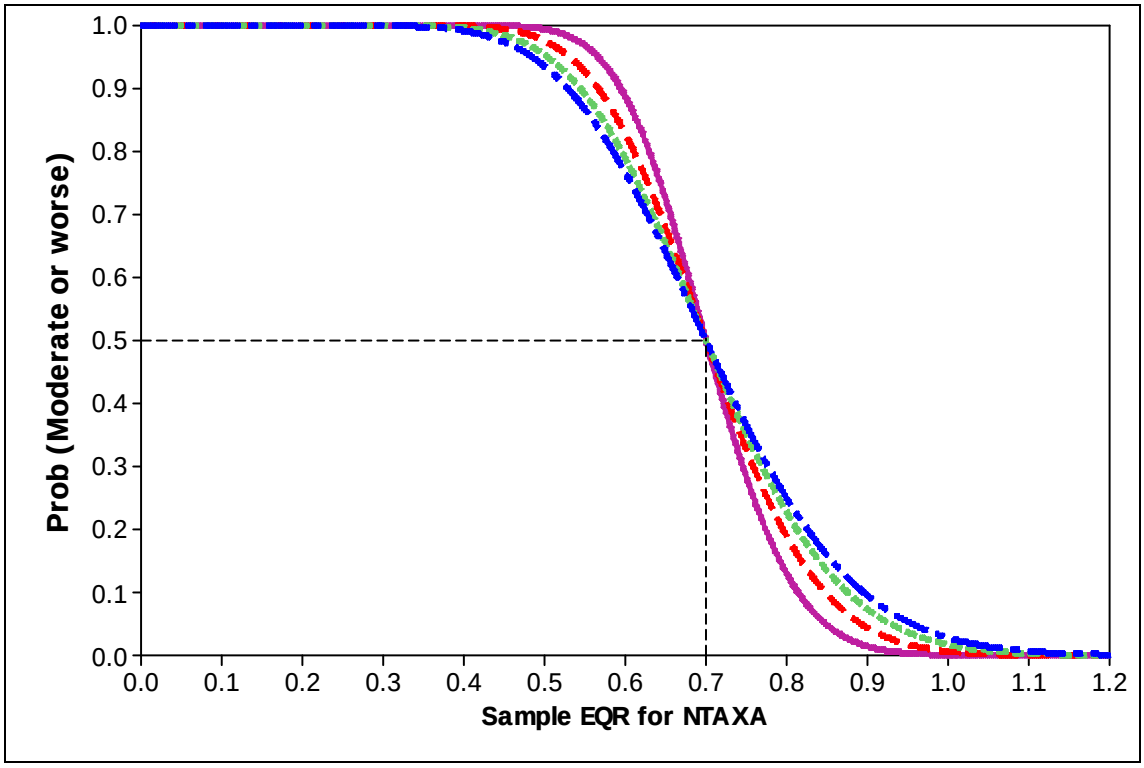


Figure 7.2 Probability of the true status being moderate or worse based on a single season sample O/E for NTAXA in relation to its true O/E and the spatial and temporal scale over which the O/E is intended to represent the average ecological quality

Note: O/E 95% denotes the upper value of O/E_{TAXA} needed to have 95% confidence that a site/water body is of 'moderate or worse' status.

7.3 Optimising sampling and monitoring effort

7.3.1 Number and choice of samples

Breaking down the overall variability in the observed (and O/E or EQR) values of biological metrics is crucial in helping to understand the spatial and temporal scale at which most macroinvertebrate variability occurs.

From our analyses of variance components, we found that roughly half (38–55 per cent) of the total variation in the assessed macroinvertebrate indices occurring at a site over a typical three-year period is due purely to variation between replicate samples taken on the same day (Table 5.3). Thus taking one or more additional replicate samples on a site visit would help to reduce the effect of this portion of the total variance, without incurring any extra costs for additional sampling site visits.

However, if the aim is purely to estimate average quality (average EQR) across a water body over a three-year period, and there were only sufficient resources to take and process three samples in total, then an efficient strategy could be to take one sample from a different sampling site in the water body in each of the three years. The variance of the average of the observed index values would be:

$$\sigma_T^2 = (\sigma_R^2 + \sigma_W^2 + \sigma_S^2)/3$$

This uncertainty variance no longer involves the inter-year variance components (see general formula in section 7.1.3). This is because we have taken a sample from all three years in which we are interested in averaging over. Thus, in sampling terms, we have sampled the whole 'population' of years in which we are interested for this particular assessment. The disadvantage of such a sampling scheme is that we have no direct information on temporal change within any one sampling site.

Figure 7.3 shows the reduction in the probability of misclassification and thus the increased confidence of class of taking one sample from a different site in each of the three years, compared to basing average water body status on only one sample taken in one of the three years. In order to have 95 per cent confidence that the average quality of the water body for the period is of 'moderate or worse' status, the average O/E_{TAXA} need only be 0.586. But a lower value of 0.492 is needed to allow for the uncertainty in EQR associated with a single sample (see Figure 7.2).

In practice, estimates of average quality for a water body over a period may end up being based on a combination of a spring and autumn combined sample EQR in one year and a spring-only sample EQR in the next, because of practical problems in obtaining the autumn sample. This complicates the estimation of uncertainty in the average EQR, as two-season combined sample index values are more precise than their single season sample counterparts. So the overall precision is some average of the variances of the individual year EQRs.

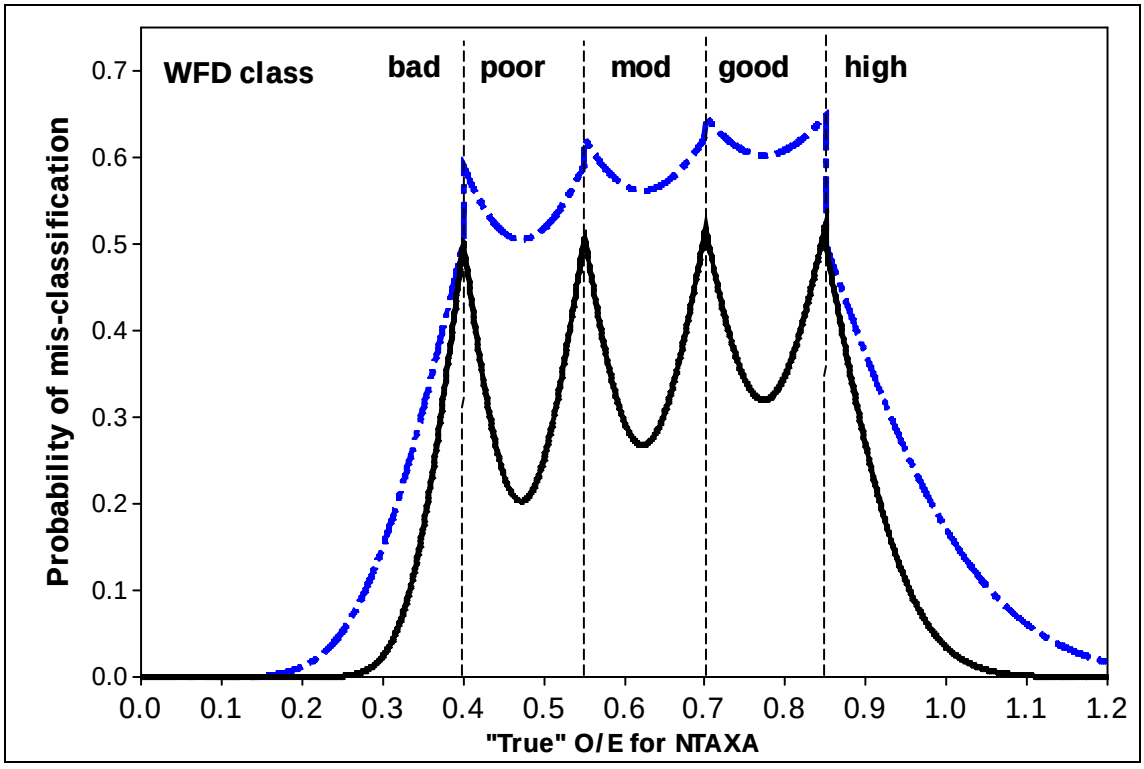


Figure 7.3 Effect of basing average O/E for a water body over a three-year period on the average of a single sample from a different sampling site within the water body in each of the three years (black line) compared to using a single sample from one site in just one year (blue dashed line)
Note: Other details as per Figure 7.1.

7.3.2 All methodological choices affect precision and accuracy

There are usually financial pressures on any environmental monitoring agency to take as few biological samples as possible while still maintaining an adequate level of monitoring of any changes in the level of ecological quality throughout a river network.

With the current practice of monitoring UK rivers using macroinvertebrate samples, time and costs occur at all stages of the assessment process. From getting to/from the sampling site and taking the macroinvertebrate sample, to subsequent laboratory sorting of the sample and taxonomic identification, and finally to processing the derived data through the RIVPACS/RICT analysis and assessment system.

Generally, whatever sampling and sample processing methods are used, throughout the whole process, the overall aim is to maximise the precision and accuracy of any monitoring scheme within the practical constraint of limited resources.

For example, at present taxa are usually only identified to family level for national monitoring purposes, and site assessments are consequently based on family-level metrics such as the BMWP and WHPT indices. This saves time (and involves less costly taxonomic expertise) than identifying to species level. However, it highlights the important potential difference between the precision of an assessment method and its accuracy.

Precision is mostly measured by the amount of variation in biological indices and O/E values, and the uncertainty of status class due to sampling and sample processing variation. However, the accuracy of an assessment method depends not only on its precision (sampling repeatability), but also on how accurately it provides a measure of the true ecological quality of a site or water body.

As a very silly extreme example of high precision but low accuracy, the assessment method could be ‘if a sample contains oligochaeta the site/water body should be classified as good or better status‘. This is a method with high precision, as the same result would be obtained for (almost) all samples. But it is not at all precise at measuring the underlying ecological quality and has no statistical power to detect sites of moderate or worse status.

This simple example reminds us that our choice of macroinvertebrate indices, and how we combine them into multi-metric indices or worst case rules, affects both the precision and the accuracy of the overall bioassessment methodology.

7.4 Links and comparison with VISCOUS approach

Julian Ellis (before retirement from WRc) and Robin Wyatt from the Environment Agency recently developed an EXCEL spreadsheet software tool called VISCOUS. This provides a means for incorporating the effect of spatial variability within a water body into the estimates and uncertainty of estimates for either the mean EQR across the water body or the percentage of the water body below some critical EQR value.

The software works by assuming that the water body can be divided into distinct parts, termed strata, each comprising a known proportion of the total water body area or length. Statistical formulae calculate an estimate for either the average water body EQR and its standard error or the estimated percentage below a critical EQR and SE, both between and within strata. These estimates and their SE are then used (together with the assumption of normal distributions) to estimate the probability that the water

body is below some user-set threshold mean EQR or that the percentage area is below the threshold EQR.

In the VISCOUS spreadsheet tool, Ellis and Wyatt provide solutions for the following example situations of available data:

M = 3 strata and N = 2–4 sites per strata

M = 1 strata and N = 2–4 sites per strata

M = 3 strata and N = 1 site per strata

M = 1 strata and N = 1 site per strata

These example spreadsheets provide a helpful explanation of how VISCOUS works. The software is structured so that the total variation within a water body is split between different sites within a stratum and between different strata. If the user provides input EQR data values for more than one site in at least one stratum, then the software can use the average within-stratum SD to estimate the overall SE for the water body. If only one site has been sampled per stratum, then the user must provide an estimate of the average within-stratum SD for individual EQR values, which is derived *a priori* from external sample datasets. In the RIVPACS III+ and the new RICT software, the estimates of EQI uncertainty for input sites are based on prior data-based estimates of the replicate sampling variance and (RICT only) temporal within-season and inter-year within-period variance components.

However, one needs to use the VISCOUS software with care, because it only explicitly deals with spatial variability within the water body. It takes no account of whether the different sample EQR values have been obtained by sampling different sites on the same day or the same or different sites on different days, seasons or even years. The VISCOUS software has no explicit concept and makes no allowance for variations in time.

VISCOUS can cope with situations where at least one site is sampled in each stratum. If only one stratum within a water body is being sampled, you have to assume it is the only stratum and the software uses a supplied estimate of within-stratum SD as the error term. You must be careful in this situation, however, because no account has been taken of the spatial variability between samples taken from different parts of the site. Consequently, the within-stratum estimate derived from the supplied EQR values, which were all taken from the same part of the water body or even from the same RIVPACS sampling site over time, would not include the unknown spatial variability and thus could under-estimate the uncertainty in the average EQR for the water body.

In the context of macroinvertebrate samples based on the RIVPACS concept of a sampling site (and for assessments based on any other BQE), it is vital to take account of both temporal and spatial variance components. These should be included in the assessment of the uncertainty associated with some measure of the average (or lower percentile) EQR values and status class for a water body over a defined assessment time period (one season, one year or three years).

This explains why we have tried to derive estimates for each of the separate variance components involved in water body assessments, namely replicate (non-spatial, non-temporal), temporal (short-term within season, longer term inter-year) and spatial (between sampling sites), as described in Chapters 3–6 of this report.

However, the VISCOUS tool can still be valid and very useful in its current form, provided the sampling criteria are understood. Specifically, if the user-supplied sample data within a stratum are spread over the whole stratum and time period being assessed, then the overall SD of these sample values will encompass all of the sources of variance. Hence, the simple within-stratum SD will be a valid estimate for

that stratum mean, even though we do not know what is pure replicate variance and what is temporal variance.

Care must be taken when EQR values are only available from one part of the water body or aren't available for all the major years of the assessment period. The new RICT tool includes formulae to cope with situations where sample values may not be available in all the years of a three-year assessment period (see section 7.1.3)

RIVPACS uncertainty assessments are currently based on deriving and using separate estimates of the sampling and other errors for the RIVPACS observed and expected values of individual metrics. It is recommended that additional analyses be conducted to assess the replicate, temporal and spatial variance in O/E (EQI) values directly, as this will make it easier to incorporate these values into EQI/EQR uncertainty assessments. The disadvantage of this approach is that it requires us to have RIVPACS expected values for each metric of each sample, which are used in statistical analyses for estimating variance components. Historical agency datasets often do not have the appropriate environmental predictor variables or expected values data readily available, but adequate estimates can be derived for use in assessing variance components of O/E (EQI) values.

7.5 Combined season or average single season sample O/E

If a RIVPACS macroinvertebrate sample is taken at a site in each of two or three RIVPACS seasons within a year, then an assessment of site quality for the year can be based on either the average or the minimum of single season sample O/E values. Alternatively, the O/E can be obtained by combining the single season samples into a combined season sample, computing the observed metric value and then dividing this by the RIVPACS predictive model site-specific expected value, based on the RIVPACS reference sites sample index values for the same season combination.

The minimum of the single season sample O/E could potentially be used as the measure for the year, but it isn't recommended due to the high uncertainty associated with the minimum of three single season O/E values. The problems and uncertainty implications of using worst case-type rules, such as the minimum of single season O/E values, are highlighted in section 7.6.

So which, therefore, is the best measure to use to represent year quality for a site: average single season O/E or combined season sample O/E?

We first considered this problem back in the mid-1990s within the BAMS project, although we were initially only interested in single year site assessments. In both approaches, the observed value is being compared with the appropriate season or season-combination expected value, so they are both valid ways for estimating quality.

However, they represent different ways of defining and measuring site quality over a period of one year. For the combined season sample O/E will be less sensitive to having much poorer quality in one season (for example the season following some incident).

The Clarke *et al.* (2002) paper summarising the BAMS study on sampling variability includes a section called 'Sampling variation in the average of single season samples'. Clarke *et al.* (2002) showed that two- and three-season combined sample O/E values have slightly lower sampling standard deviations than the average of the two or three individual season O/E values. This is the analysis that led the UK government

environment agencies to continue to use combined season sample O/E rather than switch to the average of the single season O/E values.

That was fine and appropriate for single year assessments. However, with the move to making assessments of average quality over a period of three years, it is probably worth re-considering using the average of all available single season sample O/E values.

Some samples might be missed in some seasons or years, meaning we might be combining a spring-autumn sample O/E in one year with an autumn-only sample O/E in another year. This is not necessarily a major problem for estimating the average O/E, as RIVPACS O/E values do not vary systematically up or down with the number of seasons involved. But obviously the variation in O/E varies with the number of seasons involved and this would complicate the estimate of uncertainty in average O/E.

We need further analyses and consideration of this problem, including comparing the effect of each choice and the relationship between the two choices for O/E estimator.

7.6 Experiences from the European STAR project

The largest ever European research project on macroinvertebrate sampling and methods, and their use in WFD bioassessments, was the EU Fifth Framework STAR project, led by Mike Furse of CEH Dorset over the period 2002–2006 (Furse *et al.* 2006). The main results and findings of the STAR project were published in special issue 566 of the journal *Hydrobiologia* in 2006.

As a major component of the STAR project, Clarke *et al.* (2006a, 2006b) used a carefully designed sampling study to assess the relative susceptibility to sampling variability of a wide range of commonly used European macroinvertebrate biotic metrics and sampling methods. Replicate samples were taken using both a 'national' method and an STAR-AQEM standard method at each of a range of sites covering a range of qualities from high/good to poor/moderate within one to three stream types in each participating partner country (Table 7.1).

In addition to CEH in the UK, the Austrian, German and Greek STAR partners also used the RIVPACS sampling procedures as their 'national' method in order to compare their results with those generated by STAR-AQEM sampling method used by all partners (Table 7.1).

Within the STAR project, the relative precision of metrics and sampling methods was measured by calculating the percentage (P_{samp}) of total variance in the metric values obtained using each method for each WFD system A stream type that could be attributed to replicate sampling variance. High values of P_{samp} indicate that the particular combination of metric and sampling method is highly variable between replicate samples relative to the total variability in metric values between sites for a range of qualities within the same stream type. These specific combinations of metric and sampling method will therefore have little power to detect differences in site quality and status class.

Table 7.1 STAR project: number of sites in each stream type and country with replicate samples obtained using either the RIVPACS or 'national' sampling method (and the STAR-AQEM) in at least one season (from Clarke *et al.* 2006b)

Country	Sampling method	Stream type	Stream type description	Sites	Sites x seasons
Austria	RIVPACS	A05	small-sized, shallow mountain streams	4	6
		A06	small-sized crystalline streams of the ridges of the Central Alps	4	7
Czech Republic	PERLA	C04	small-sized, shallow mountain streams	3	6
		C05	small-sized streams in the Central sub-alpine mountains	3	6
Germany	RIVPACS	D03	medium-sized lowland streams	2	4
		D04	small-sized, shallow mountain streams	2	4
		D06	small-sized Buntsandstein-streams	2	4
France	IBGN	F08	small-sized, shallow headwater streams in Eastern France	6	12
Greece	RIVPACS	H04	small-sized calcareous mountain streams in Western, Central and Southern Greece	6	12
Italy	IBE	I06	small-sized calcareous streams in the Central Apennines	6	11
Denmark	DSFI	K02	medium-sized lowland streams	6	12
Latvia	LVS 240:1999	L02	medium-sized lowland streams	6	12
Poland	National	O02	medium-sized lowland streams	7	13
Portugal	PMP	P04	medium-sized streams in lower mountainous areas of S. Portugal	6	12
Sweden	National	S05	medium-sized lowland streams	3	6
		S06	medium-sized streams on calcareous soils	3	6
UK	RIVPACS	U15	small-sized, shallow lowland streams	3	6
		U23	medium-sized lowland streams	3	6

Notes: Small-sized = 10–100km^2, medium sized = 100–1000km^2, lowland = <200m above sea level.

Table 7.2 gives the average percentage replicate sampling variance (P_{samp}) for each metric averaged across all stream types in the UK, Germany, Austria and Greece, where the RIVPACS (and STAR-AQEM) sampling methods were used.

The Saprobic abundance-based metrics appear to be least susceptible to replicate sampling variability, while replicate sampling variability was higher (P_{samp} = 15–17 per cent) for both 'Number of families' and ASPT in the RIVPACS and STAR-AQEM methods. However, as with all statistics, these average values can be misleading. For the ASPT metric based on the RIVPACS method, P_{samp} was only 5 per cent for the UK STAR stream types and 9 per cent for Austrian stream types. However, it was 19 per cent for Greek stream types and 27 per cent for the sampled German stream types, where the main environmental stress was degradation of stream morphology rather than organic pollution. Thus the total variation in ASPT was relatively small. However, the Saprobic metrics did tend to perform better than ASPT.

Table 7.2 Ordered median values of the average percentage sampling variance (P_{samp}) for each metric across all stream types and countries for the RIVPACS method and for the STAR-AQEM method (from Clarke *et al.* 2006b)

Metric	national/ RIVPACS	STAR-AQEM
Saprobic Index	3	3
German Saprobic new	4	5
Czech Saprobic	4	6
Trait m12 : preferred current<25cm/s	6	12
% Littoral	7	15.5
% EPT (abundance-classes)	7	9
% Rheophilic (abundance-classes)	8	12
Number of EPT taxa	9	15.5
% Shredders	10	10
% EPT taxa	10	18
% Grazers/Scrapers	10.5	16
% Rheophilic	11	12
Trait m2 : >1 cycle per annum	12	10
Trait m1:max size ≤1cm	14	27.5
% Gatherers/Collectors	15	14
% EPT individuals	15	15
RETI	15	12
Diversity SW	16	14
Number of taxa	16.5	15.5
ASPT	16.5	17
Number of families	17.5	15.5
1 –GOLD	17.5	16.5
% Oligochaeta	19	16
Abundance [ind/m²]	21	21.5
IBE	25.5	16.5
Trait m7 : crawler locomotion	26	17.5

In the STAR project, we also tried to compare the relative precision of different 'national' sampling methods by calculating the average percentage replicate sampling variance (P_{samp}) for a method average across all metrics and stream types. As mentioned earlier, the value of P_{samp} can partly depend on the type and range of stress levels operating within the sampled streams. So we also compared the average P_{samp} values for the national and cross-project STAR-AQEM sampling methods, separately for each country (Figure 7.4).

Using this best-available information, we found that the highest average sampling precision was obtained by the Czechs, using both the STAR-AQEM and their own adapted PERLA 'national' sampling and sample processing protocols. (However, in a STAR evening workshop meeting, we did subsequently discover that the various researchers sampling a site first agreed on the relative cover of the different habitats to be sampled, thus eliminating one potential source of variability between replicate samples).

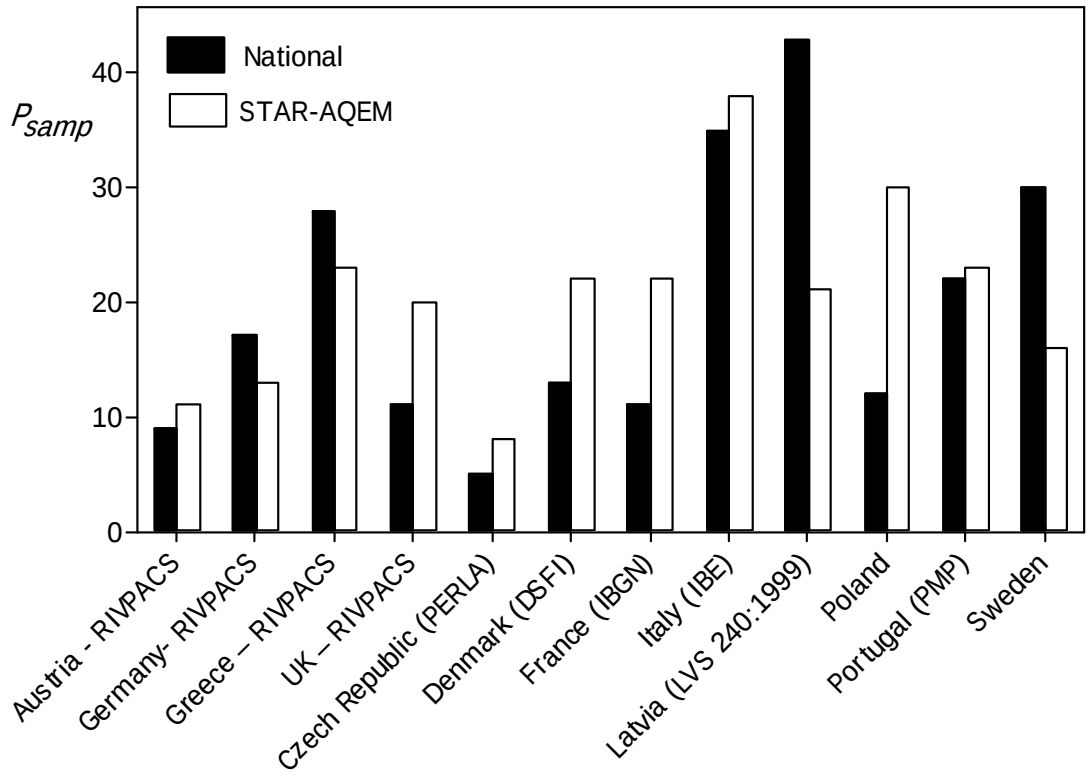


Figure 7.4 Average percentage replicate sampling variance (P_{samp}) over all 26 metrics for the 'national'/RIVPACS and STAR-AQEM methods for each country involved in the STAR project (from Clarke *et al.* 2006b)

The RIVPACS sampling protocol, as used in the UK and Austria, appeared to have one of the lowest average percentage replicate sampling variances (Figure 7.4). This suggests it is at least as precise, in terms of derived biological metrics, as other European national macroinvertebrate sampling (and, equally important, sample processing) methods

The STAR-AQEM multi-habitat sampling protocol, or a variant of it, is commonly used in Germany and the Czech Republic as the standard macroinvertebrate assessment procedure. Clarke *et al.* (2006a) found that the sub-sampling procedures in the STAR-AQEM procedures (whereby a minimum of 700 individuals must be identified and counted) typically accounted for around half of the total variance in metric values between replicate samples. In other words, how a macroinvertebrate sample is processed once it has been collected from the river can be as important a source of variability, error and uncertainty as the natural small-scale spatial variability in macroinvertebrate distribution within a site.

7.7 Uncertainty for multiple metric and worst-case rules

Most of this report and the data analyses have concentrated on the uncertainty associated with a range of individual metrics and the derived EQRs and status classes. Individual metrics are the 'building blocks' of any overall site/water body assessment. In practice, many site and water body assessments are based on a combination of macroinvertebrate metrics and, more generally, on a combination of metrics and indices from more than one WFD biological quality element (BQE), including fish, macroinvertebrates, diatoms, macrophytes and habitat.

When the results of estimating status class based on each of several different metrics (and/or BQEs) are combined, the procedure for defining overall class is often to take the worst class from each of the individual metrics (and/or BQEs). This is known as the 'one-out all-out' or 'worst case' rule.

7.7.1 Worst case rule – implications for uncertainty

A form of worst case rule is also involved when the status class of a water body is set as the worst of the classes estimated at each of several sampling sites within the water body. There is a current discussion within the Environment Agency as to whether to use worst EQR (and class) or average EQR, or maybe to use the class of the worst 15 per cent length of river within the water body (although this is equivalent to worst case if, as is usual, there are less than seven sampling sites within a water body).

However, using any type of worst case rule (either across metrics or space) has implications for the uncertainty and confidence of class associated with the estimate of overall status class for a site or water body.

Ellis (2007) considers two different approaches for defining worst class based on the EQR values and confidence of class probabilities for two or more metrics for a site. One approach (referred to by Ellis 2007 as 'Interpretation II') is equivalent to the approach used in RIVPACS III+ and the new RICT tool. It is based on simulating uncertainty in each metric's EQR and then applying the worst case rule to the status class of the metrics for each simulation in turn. This is equivalent to the RIVPACS/RICT MINTA rule of using the worst of the classes based on O/E_{TAXA} and O/E_{ASPT}, which is currently used for UK river assessments.

The general problem with using worst case rules is best illustrated by an example in which the overall assessment for a site is based on the lowest status class of the EQR values of M metrics, where M could be one, two, three or four. For simplicity, assume that the critical good/moderate boundary of each metric's EQR is set to 0.7. Then taking the worst class of each metric is the same as taking the class of the lowest EQR. It is assumed that the overall sampling variance for each metric is the same, and equivalent to that described in section 7.1 for NTAXA of single season samples intended to represent the average for a water body over a three-year period. This gives each EQR and sampling SD of 0.135. Furthermore, the sampling variability of metrics is assumed, for illustrative purposes, to be independent.

Figure 7.5 shows the distribution of potential values for the minimum of M metric EQR values for a site where the true average value of each metric is on the good/moderate boundary at 0.7. With a single (M=1) metric, half the sample values will be greater than 0.7 and half will be less than 0.7, so the probability of being classed as 'moderate or worse' is 50 per cent.

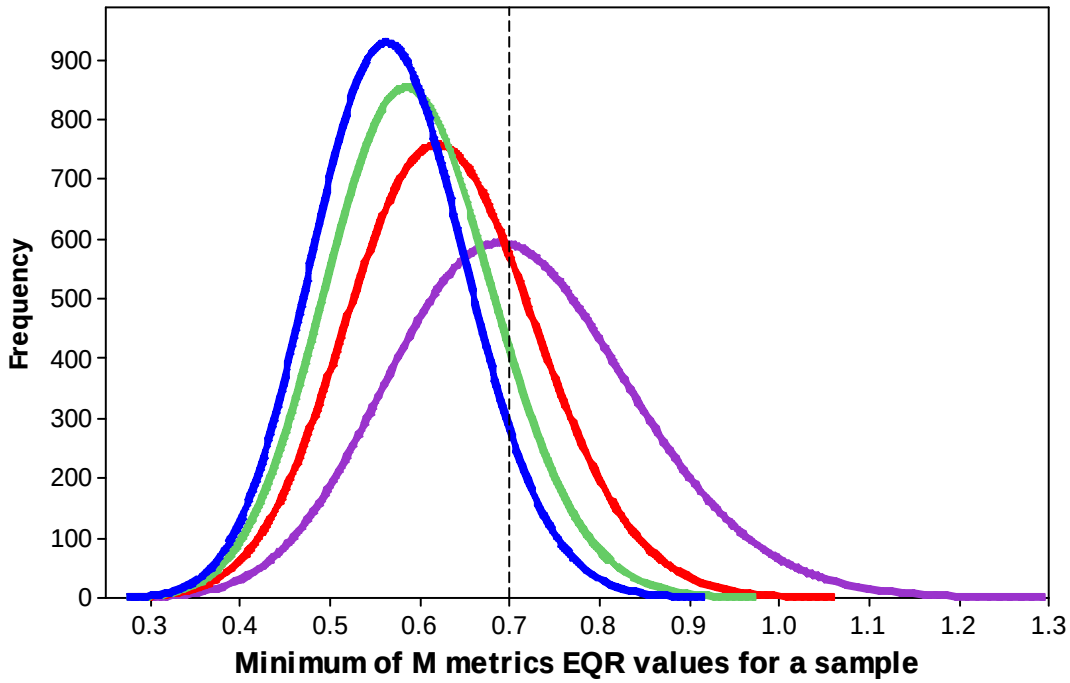


Figure 7.5 Distribution of the minimum of M metrics EQR values for a sample where each metric EQR has a mean of 0.7 and a sampling SD of 0.135
Note: M = 1 (purple line), 2 (red), 3 (green) or 4 (blue).

As the overall assessment is based on an increasing number of metrics, the minimum of their EQR values for a sample and site/water body tends to decrease. In particular, when M equals one, two, three or four metrics, the mean value of the minimum EQR for a sample is 0.70, 0.63, 0.59 and 0.57 respectively (Figure 7.5)

This idea can be extended to derive the distribution of the minimum of M metric EQRs when each EQR has the same true mean value Q, where Q varies between 0 and 1. For each value of Q, RIVPACS or STARBUGS software stochastic uncertainty simulations (or otherwise, for example MINITAB) can be used to calculate the probability that the worst class based on M metrics would be 'moderate or worse' given a good/moderate boundary EQR value of 0.7 for each metric (Figure 7.6).

When the true average EQR for a single metric is around the good/moderate boundary, then, as expected, the probability of 'moderate or worse' is 50 per cent. However, as the overall assessment is based on the worst class of two, three or four metrics, the probability of being classed as 'moderate or worse' when each metric's EQR is on average on the good/moderate boundary increases to 75 per cent, 87.5 per cent and 93.75 per cent respectively (Figure 7.6). This is because the probability of all M metrics (M=1–4) being above the good/moderate boundary is 0.5 raised to the power M, namely 0.5, 0.25, 0.125 and 0.0625.

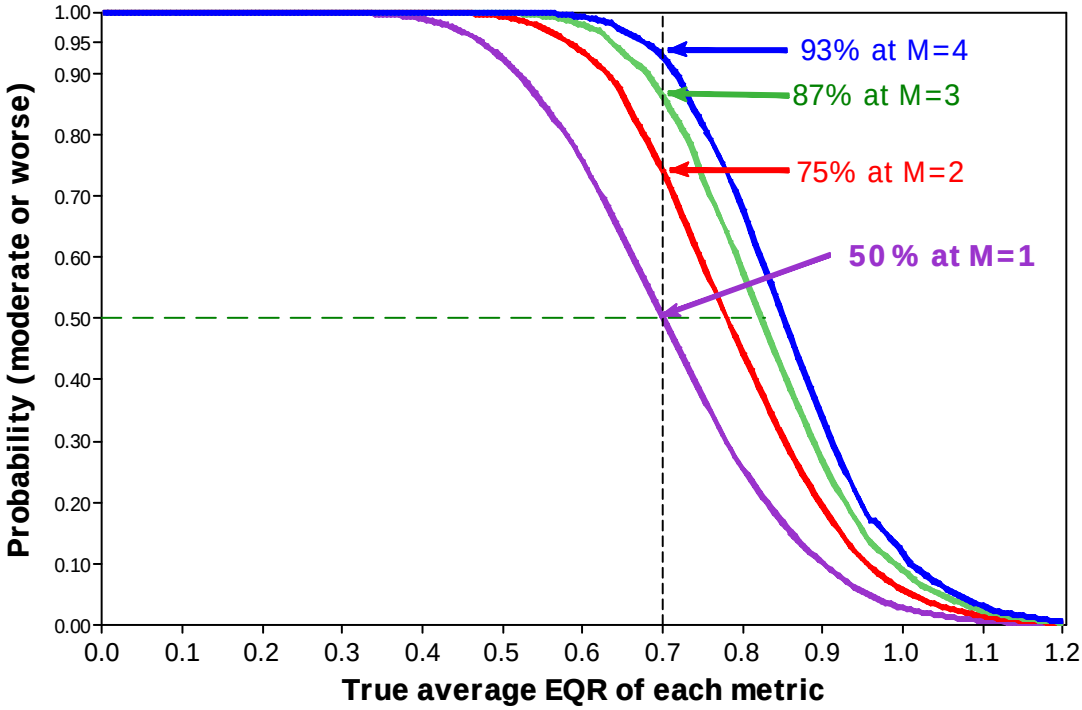


Figure 7.6 Probability of being classed as 'moderate or worse' status when using the worst class rule, where each metric EQR has a SD of 0.135 and the same true average value

Note: M = 1 (purple line), 2 (red), 3 (green) or 4 (blue).

Thus, as more metrics are involved in a worst case rule, the risk of type 1 errors of status misclassification increases (where a type 1 error means concluding that a site/water body is impacted when it is not).

One possible solution would be to try to apply a correction factor to the status class limits in a way that is analogous to the Bonferroni correction used when there are multiple statistics tests for differences in ANOVA. The status class limits for each metric could be set higher to allow for the effects of sampling variability (of prior estimated size) on the worst case rule. In the example given above and shown in Figure 7.6, the site assessment can be taken as the worst EQR (and class) of M=3 metrics, each of which has the good/moderate boundary set to 0.70. To maintain the individual metric's 50 per cent type 1 error for sites with EQR values on the good/moderate boundary, the boundary would need to be set higher, at around 0.82, in order to have a 50 per cent sampling chance that the three-metric worst case rule would assign such 'borderline' sites to the lower 'moderate or worse' class.

In practice, each metric will tend to have different EQR class limits, but this example demonstrates the general principle. What would seem strange, even though logical, would be to have different class limits for a metric depending on how many other metrics it was combined within in a worst case rule.

The UK government environment agencies cannot afford to designate incorrectly too many water bodies as of 'moderate or worse' status. Such status requires costly measures to be put in place by the agencies or businesses to rectify the apparent problems.

This discussion highlights the problems associated with multi-metric and multi-BQE worst case rules due to sampling variability. It also suggests why other forms of multi-metric index that involve the weighted-averaging of individual metric EQR values for a site, or spatial averaging across sample sites within a water body, may avoid some of

these problems. This is partly why we have concentrated this study on estimating the average water quality over a period of time across a water body.

7.7.2 Multi-metric averages

The variance and correlations between metrics – across replicates, space and time, and between sites – will influence the effectiveness, precision and accuracy of any multi-metric index. If two indices are very highly correlated (such as NTAXA and BMWP score), then they will always tend to give the same estimate of class for any sample (assuming class limits have been set equivalently). As a result, their single response to stress will count double in the overall metric. If the overall rule was to use the median EQR of three metrics, two of which are very highly correlated, the result would be the class of these two metrics.

Clarke and Hering (2006) point out that adding an extra metric to a multi-metric index (MMI) could increase the MMI variance and reduce its precision if the extra metric is relatively more susceptible to sampling variation. Equally importantly, from an ecological rather than a purely statistical viewpoint, every change of metric and/or the class limits in an MMI changes the biological requirements for each status class. Obviously, care should be taken in adding new metrics, especially in terms of what stresses they are expected to respond to, the strength and form of their response, and their relative sampling variability and precision.

In the UK, the development of the new RICT software aims to generalise the existing RIVPACS III+ software by making it easier to update the overall classification system to include new metrics. It already includes the ability to utilise expected values of the Walley-Hawkes revised BMWP indices (denoted WPPT), and the family-level LIFE and AWIC indices.

7.8 Recommendations

In conducting this review of past studies and new analyses of temporal and spatial datasets to assess uncertainty of river ecological quality across a whole water body, we have tried to cover all of the factors and sources of uncertainty that can be involved and may need to be considered and quantified. This has led to the discovery of gaps in our knowledge and highlighted potential improvements to the assessment methodology that still need to be addressed. These are summarised in the following recommendations (in no particular order) for further work.

1. The new RICT software should include specific estimates and information on the confidence of failing to achieve good or better status in addition to the confidence of belonging to individual WFD status classes.
2. There is a need to collate and analyse a much larger dataset of spatial variability between sampling sites within the new WFD water bodies, ideally with temporal and replicate sampling information on at least a subset of the same sites. This should allow improved estimates of the scale of spatial heterogeneity within rivers.
3. Methods giving fixed RIVPACS predictions for each site should be developed. These should be based on either temporally-invariant Geographic Information System/map-based site and catchment environmental variables or long-term (five-year) average environmental variables. This would provide O/E values for every sample and allow direct assessment of O/E variance components.
4. Some environmental parameters can be affected by flow and so current predictions can miss the impact of abstraction. There is a need to develop predictions that are not influenced by flow or new rules for using such data for WFD predictions (because flow pressures are to be considered).
5. There should be further analyses of RIVPACS sample audit data to derive and incorporate direct estimates of sample processing errors and biases in other indices (in addition to NTAXA) into the RICT software for assessing confidence of class.
6. The merit of using the average of single season sample O/E values as a measure of water body quality over a period should be reconsidered, and contrasted with the current combined season sample approach.
7. Statistical methods to cope with any actual temporal and spatial mix of samples should be developed and these methods incorporated into either the RICT software or an extended version of the VISCOUS-type software tool.
8. A standardised sampling approach for assessing non-wadeable rivers (based on Environment Agency/NS-share/CEH ‘deep rivers’ research) should be developed and a BAMS-like study to quantify uncertainty conducted.

References

ARMITAGE, P.D., MOSS, D., WRIGHT, J.F. AND FURSE, M.T., 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.

ARMITAGE, P.D., 2000. The potential of RIVPACS for predicting the effects of environmental change. *In:* J.F. Wright, D.W. Sutcliffe and M.T. Furse, eds. *Assessing the biological quality of fresh waters: RIVPACS and other techniques*. Ambleside: Freshwater Biological Association, 93–112.

CLARKE, R.T., FURSE, M.T., WRIGHT, J.F. AND MOSS, D., 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., FURSE, M.T., GUNN, R.J.M., WINDER J.M AND WRIGHT, J.F., 2002. Sampling variation in macroinvertebrate data and implications for river quality indices. *Freshwater Biology*, 47, 1735–1751.

CLARKE, R.T., WRIGHT, J.F. AND FURSE, M.T., 2003. RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. *Ecological Modelling*, 160, 219–233.

CLARKE, R. T. AND HERING, D., 2006. Errors and uncertainty in bioassessment methods – major results and conclusions from the STAR project and their application using STARBUGS. *Hydrobiologia*, 566: 433–439.

CLARKE, R.T., LORENZ, A., SANDIN, L., SCHMIDT-KLOIBER, A., STRACKBEIN, J., KNEEBONE, N.T AND HAASE, P., 2006a. Effects of sampling and sub-sampling variation using the STAR-AQEM sampling protocol on the precision of macroinvertebrate metrics. *Hydrobiologia*, 566, 441–459.

CLARKE, R.T., DAVY-BOWKER, J., SANDIN, L., FRIBERG, N., JOHNSON, R.K. AND BIS, B., 2006b. 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.

DAVY-BOWKER, J., MURPHY, J.F., RUTT, G.P., STEEL, J.E.C. AND FURSE, M.T., 2005. The development and testing of a macroinvertebrate biotic index for detecting the impact of acidity on streams. *Archiv für Hydrobiologie*, 163, 383–403.

ELLIS, J.E., 2007. *Combining multiple quality elements and defining spatial rules for WFD classification*. Environment Agency Science Project Number SC060044, Product Code GEHO0807BMXZ-E-E.

FURSE, M.T., CLARKE, R.T., WINDER, J.M., SYMES, K.L., BLACKBURN, J.H., GRIEVE, N.J. AND GUNN, R.J.M., 1995. *Biological assessment methods: controlling the quality of biological data. Package 1: The variability of data used for assessing the biological condition of rivers.* Bristol: National Rivers Authority, (R&D Note 412).

HERING, D., MOOG, O., SANDIN, L. AND VERDONSCHOT, P. F. M., 2004. Overview and application of the AQEM assessment system. *Hydrobiologia*, 516, 1–21.

MURRAY-BLIGH, J., 1997. Procedures for collecting and analysing macro-invertebrate samples. In: *Quality management systems for environmental monitoring: biological techniques*. Bristol: Environment Agency, (BT001).

OSTERMILLER, J. D. AND HAWKINS, C. P., 2004. Effects of sampling error on bioassessments of stream ecosystems: application to RIVPACS-type models. *Journal of the North American Benthological Society*, 23, 363–382.

VAN DIJK, P., 1994. *Analytical quality control for macroinvertebrate enumeration*. Bristol: National Rivers Authority, (R&D Note 331).

WALLEY, W.J. AND FONTAMA, V.N., 2000. New approaches to river quality classification based upon artificial intelligence. *In:* J.F. Wright, D.W. Sutcliffe and M.T. Furse, eds. *Assessing the biological quality of fresh waters: RIVPACS and other techniques*. Ambleside: Freshwater Biological Association, 263–279.

WALLEY, W.J. AND FONTAMA, V.N., 1998. Neural network predictors of average score per taxon and number of families at unpolluted river sites in Great Britain. *Water Research*, 32(3), 613–622.

WRIGHT J.F., 2000. An introduction to RIVPACS. *In:* J.F. Wright, D.W. Sutcliffe and M.T. Furse, eds. *Assessing the biological quality of fresh waters: RIVPACS and other techniques*. Ambleside: Freshwater Biological Association, 1–24.

WRIGHT J.F., MOSS, D., ARMITAGE, P.D. AND FURSE, M.T., 1984. A preliminary classification of running-water sites in Great Britain based on macroinvertebrate species and the prediction of community type using environmental data. *Freshwater Biology*, 14, 221–256.

WRIGHT, J.F, SUTCLIFFE, D.W. AND FURSE, M .T., eds, 2000. *Assessing the biological quality of fresh waters: RIVPACS and other techniques*. Ambleside: Freshwater Biological Association.

Abbreviations

ANOVA	Analysis of Variance
AQC	Analytical Quality Control
AQEM	Assessment system for the ecological Quality of streams and rivers throughout Europe using menthic Macroinvertebrates
ASPT	Average Score Per Taxon
AWIC	Acid Water Indicator Community
BAMS	Biological Assessment Methods
BQEs	Biological Quality Elements
BMWP	Biological Monitoring Working Party
CAVE	Combines Appropriate Variance Estimates
CEH	Centre for Ecology & Hydrology
EQI	Environmental Quality Index
EQR	Ecological Quality Ratio
FAME	Development, Evaluation and Implementation of a Standardised Fish-based Assessment Method for the Ecological Status of European Rivers
GQA	General Quality Assessment
IFE	Institute of Freshwater Ecology
LIFE	Lotic Invertebrate index for Flow Evaluation
MDA	Multiple Discriminant Analysis
MINTA	Minimum of status classes based on O/E for TAXA and ASPT
MMI	Multi-metric index
NI	Northern Ireland
NIEA	Northern Ireland Environment Agency
NRA	National Rivers Authority
NS Share	North South Shared Aquatic Resource
O/E	Observed/Expected
QA	Quality Audit
RC	Reference Condition
REML	Residual/Restricted Maximum Likelihood
RHS	
RICT	River Invertebrate Classification Tool
RIVPACS	River Invertebrate Prediction And Classification System
RPB	River Purification Board
RQS	River Quality Survey
SD	Standard Deviation
SE	Standard Error
SEPA	Scottish Environment Protection Agency
SNIFFER	Scotland and Northern Ireland Forum for Environmental Research
STAR	STAndardisation of River classifications European project
VISCOUS	Variability In Spatial Component Objectivity Unified Statistically
WFD	EU Water Framework Directive
WHPT	Walley-Hawkes weighted ASPT
WRc	Water Research Centre

Would you like to find out more about us, or about your environment?

Then call us on

08708 506 506* (Mon-Fri 8-6)

email

enquiries@environment-agency.gov.uk

or visit our website

www.environment-agency.gov.uk

incident hotline 0800 80 70 60 (24hrs)
floodline 0845 988 1188

* Approximate call costs: 8p plus 6p per minute (standard landline).
Please note charges will vary across telephone providers

Environment first: This publication is printed on recycled paper.