1996
2006

Environment Agency

enhancing... improving... cleaning... restoring...
changing... tackling... protecting... reducing...
creating a better place... influencing...
inspiring... advising... managing... adapting...

Uncertainty estimation for monitoring results by the WFD biological classification tools

WFD Report

GEHO1006BLOR-E-P

ea/br/e/std/vr10thanni

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: 1844326063

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):
Julian Ellis (WRc), Veronique Adriaenssens (EA)

Dissemination Status:
Public Domain

Keywords:
Water Framework Directive (WFD), water quality, ecological status, uncertainty

Research Contractor:
WRc

Environment Agency's Project Manager:
Veronique Adriaenssens

Product Code: GEHO1006BLOR_E_P

Executive summary

The new European Union (EU) water policy, the Water Framework Directive (WFD), states that all European surface waters achieve good ecological status by 2015. The ecological status is an expression of the quality of the structure and functioning of biological elements associated with surface waters, classified in Annex V. Ecological status should be assessed via a reference condition approach using bio-assessment tools (further called 'classification tools') based on five biological elements (EU 2000) which in the case of rivers are: (1) phytoplankton, (2) macrophytes and phytobenthos, (3) benthic invertebrate fauna and (4) fish. The classification tools need to assess the composition and the abundance of the biological element (as well as the age structure for the fish classification tool) based on the monitoring results.

In order to ensure comparability of such monitoring systems between the different Member States, the results of the systems operated by each Member State shall be expressed as ecological quality ratios (EQRs) for the purposes of classification of ecological status. These ratios shall represent the relationship between the values of the biological parameters observed for a given body of surface water and the values for these parameters in the reference conditions applicable to that body. The ratio shall be expressed as a numerical value between zero and one, with high ecological status represented by values close to one and bad ecological status by values close to zero. Each Member State shall divide the ecological quality ratio scale for their monitoring system for each surface water category into five classes ranging from high to bad ecological status by assigning a numerical value to each of the boundaries between the classes.

The required precision and confidence to be achieved by the monitoring results is mentioned in four contexts:

1. Estimates of the level of confidence and precision of the results provided by the monitoring programmes shall be stated in the river basin management plan.
2. In selecting parameters for biological quality elements Member States need to identify the appropriate taxonomic level required to achieve adequate confidence and precision in the classification of the quality element.
3. Frequencies shall be chosen so as to achieve an acceptable level of confidence and precision.
4. Estimates of the level of confidence and precision of the results provided by the monitoring programmes shall guide the development of cost-effective programmes of measures.

The first three are 'legal' requirements and part of the WFD legislative document. The fourth has been used so far in the UK related to water quality classification.

In this study, the variability of metrics is estimated for a specific classification tool in order to quantify precision and confidence as required by the WFD. We will only develop a method to fulfil the requirements as stipulated under (1). However, once an uncertainty estimation methodology is developed, it can be used to determine (2) to (4) and even help in refining the classification tools in several other aspects (e.g. boundary setting).

Acknowledgements

Thanks to Alan Starkie (EA, Fisheries) and Richard Noble (HIFI, Hull University) for their comments and help in writing this report.

Contents

1 Introduction

1.1 Background

The new European Union (EU) water policy, the Water Framework Directive (WFD), states that all European surface waters achieve good ecological status by 2015. The ecological status is an expression of the quality of the structure and functioning of biological elements associated with surface waters, classified in Annex V. Ecological status should be assessed via a reference condition approach using bio-assessment tools (further called 'classification tools') based on five biological elements (EU 2000) which in the case of rivers are: (1) phytoplankton, (2) macrophytes and phytobenthos, (3) benthic invertebrate fauna and (4) fish. The classification tools need to assess the composition and the abundance of the biological element (as well as the age structure for the fish classification tool) based on the monitoring results.

In order to ensure comparability of such monitoring systems between the different Member States, the results of the systems operated by each Member State shall be expressed as ecological quality ratios (EQRs) for the purposes of classification of ecological status. These ratios shall represent the relationship between the values of the biological parameters observed for a given body of surface water and the values for these parameters in the reference conditions applicable to that body. The ratio shall be expressed as a numerical value between zero and one, with high ecological status represented by values close to one and bad ecological status by values close to zero. Each Member State shall divide the ecological quality ratio scale for their monitoring system for each surface water category into five classes ranging from high to bad ecological status by assigning a numerical value to each of the boundaries between the classes.

The required precision and confidence to be achieved by the monitoring results is mentioned in four contexts:

1. Estimates of the level of confidence and precision of the results provided by the monitoring programmes shall be stated in the river basin management plan.
2. In selecting parameters for biological quality elements Member States need to identify the appropriate taxonomic level required to achieve adequate confidence and precision in the classification of the quality element.
3. Frequencies shall be chosen so as to achieve an acceptable level of confidence and precision.
4. Estimates of the level of confidence and precision of the results provided by the monitoring programmes shall guide the development of cost-effective programmes of measures.

The first three are 'legal' requirements and part of the WFD legislative document. The fourth has been used so far in the UK related to water quality classification.

The panel below provides some basic definitions relating to precision and confidence.

Interpretation of 'precision' and 'confidence'

The degree of assurance provided by a confidence interval is described by ***the confidence coefficient*** *(e.g. 90%, 95%) - often referred to as the* ***confidence level.*** *This is the proportion of occasions, in the long run, on which a calculated confidence interval will actually contain the true (but unknown) quantity being estimated. For example, if we calculate a 90% confidence interval for the mean from monitoring data for each of 40 different sites, we would expect the true site mean to fall within its corresponding confidence interval for about 36 of those 40 sites.*

Precision *is usually defined as the half-width of the confidence interval. For this reason, any statement of precision must be accompanied by a statement of the corresponding confidence level.*

Whenever the words 'precision' and 'confidence' are used together, they should always be interpreted in the ***statistical*** *sense. That is because, as 'precision' is the half-width of the confidence interval, it has no objective meaning unless confidence is defined in the statistical sense.*

However, when the word 'confidence' is used on its own, it often has a ***non-statistical*** *meaning - such as: "We are highly confident that macroinvertebrates are better indicators of organic pollution than fish". Or "We are fairly confident that Site X is representative of average conditions in water body Y". It is important, therefore, to be clear which type of 'confidence' is being referred to whenever the word is mentioned.*

Risk of misclassification
No estimate of quality based on sampling will be exactly equal to the true value in the underlying population (except by a lucky chance). This inevitable element of uncertainty is known as 'sampling error'. Because of sampling error, the estimated EQR for a quality element may cause the water body to be put into a different class from its 'true' class - that is, the class that would be obtained given perfect information for that location and time period. The risk of this happening is known as the ***Risk of Misclassification*** *(RoM).*

In this study, the variability of metrics is estimated for a specific classification tool in order to quantify precision and confidence as required by the WFD. We will only develop a method to fulfil the requirements as stipulated under (1). However, once an uncertainty estimation methodology is developed, it can be used to determine (2) to (4) and even help in refining the classification tools in several other aspects (e.g. boundary setting).

Before we start to explain the methodology as illustrated for the fish classification tool (European Fish Index), we need to set the scope on which sort of systems this methodology is applied.

1.2 Types of classification tools

1.2.1 Unimetric versus multimetric classification tools

The classification tool for a specific biological element can be multimetric - that is, it consists of different metrics which each represents a scoring system for a parameter

indicative for biological quality elements (such as diversity, evenness, etc.). If it only consists of one metric (e.g. RIVPACS) then it is a unimetric tool.

1.2.2 EQRs constrained to lie in [0-1] versus EQRs that may go beyond 1

The EFI fish classification tool (European Fish Index) is an arithmetic mean of ten probabilities (see below), and so no EFI value can ever fall outside the range [0-1]. In contrast, the macroinvertebrate RIVPACS classification tool does allow EQRs to go beyond 1 (although for WFD purposes all scores higher than 1 will be set equal to 1).

1.2.3 Stance taken towards uncertainty in 'Expected' data

In the RIVPACS tool, error arises in measuring the environmental variables at a site, and so there is some statistical uncertainty in the Expected value (for number of taxa, say). Thus there is error due to both the Observed and Expected components of the EQR - and this is taken account of in the simulation analysis carried out by RIVPACS III+.

For other EQRs, however, it may be that the reference condition for a water body is assumed to be known without error, in which case the uncertainty will be due solely to the variability in the Observed data.

1.2.4 The European Fish Index Tool

The analysis given in this paper is confined to the **European Fish Index Tool** (Pont et al., 2006). The EFI (**E**uropean **F**ish **I**ndex) is the Water Framework Directive fish classification tool for rivers developed by the FAME (**F**ish-based **A**ssessment **M**ethod for the **E**cological Status of European Rivers) EC funded project - for which the Environment Agency was the UK Applied partner. It scores single-run electric fishing data from a river survey site on a scale of 0 to 1, using an Index of Biotic Integrity approach with 10 unweighted metrics. These scores equate to the WFD status categories: bad, poor, moderate, good and high as follows:

0.000 – 0.187	= class 5	= bad	(red)
0.187 – 0.279	= class 4	= poor	(orange)
0.279 – 0.449	= class 3	= moderate	(yellow)
0.449 – 0.669	= class 2	= good	(green)
0.669 – 1.000	= class 1	= high	(blue)

To quantify possible deviation from a 'reference condition' for any given site, they first established and validated statistical models describing metric responses to natural environmental variability in the absence of any significant human disturbance. They considered that the residual distributions of these models described the response range of each metric, whatever the natural environmental variability. After testing the sensitivity of these residuals to a gradient of human disturbance, they finally selected 10 metrics,

each lying on a [0-1] scale, that were combined to obtain a European fish assemblage index.

In this study, the uncertainty in the EQR is calculated *directly*-by carrying out an Analysis of Variance on sets of observed EQR values, and not *indirectly* from information about each of the individual metrics (although the latter would in principle be the sounder approach). It follows that, given the 'bundled' nature of the EQR, we do not take into account separate error estimates related to the observed or the expected quality on its own.

2 Variability of Metrics

2.1 Components of variation

Any metric is subject to four broadly different types of variation:

- Spatial variation - differences from site to site within a water body;
- Temporal variation - within year, both seasonal and random, and year-to-year;
- Temporal-Spatial interaction - whereby a particular temporal effect operates differently in some locations than others; and
- Sampling and measurement error.

The extent to which these components of variation have a bearing on water body classification depends crucially on the details of (a) the stipulated sampling methodology, (b) the nature of the class limits (e.g. mean or percentile), and (c) the classification assessment method - including the stance taken towards the burden of proof. These issues are discussed briefly in the next section.

2.2 Relevance to Confidence of Class

Suppose, for example, that the classification is to apply to a *three-year* period, and the limits refer to required *mean* quality. If sampling is to be carried out on just one date, some allowance needs to be made for the possibility that quality was unusually good or poor on that particular occasion; and that requires information on the temporal components of variation. The same applies if a number of samples are to be taken through the three-year assessment period. The seasonal component of variation will be smoothed out by the overall mean, but the random temporal component remains, and some allowance will be needed for this - especially if a Benefit-of-Doubt stance is being taken. In addition, it is important to know how the class limits are to be interpreted. For example, the data assessment would be radically different if the class limits specified the level of quality to be met for, say, *90% or more* of the time rather than *on average*.

Similar issues arise in considering the spatial component of variation. If a single sampling point is being used to assess a water body, this may provide an unduly harsh or favourable picture of mean quality across the whole water body - to an extent that can be quantified knowing the spatial component of variation. The same applies if sampling is to be carried out at several sites. And as with the temporal element, the interpretation to be placed on the class limits has a critical bearing on the outcome. For example, do they specify the level of quality to be met for, say, *90% or more* of the water body, or *on average*?

These are complex issues that we cannot pursue in the present paper. In most of what follows, therefore, we assume that the immediate aim is to determine the Confidence of

Class (CofC) for a single site on a single occasion. We do, however, give some indication of how the approach could be generalised to accommodate a broader assessment over space or time.

In a draft report and an accompanying Excel file (version 11/10/2005) we demonstrated the Risk of Misclassification Approach based on hypothetical EQR (Ecological Quality Ratio) values of a certain water body. We set up an example for surveillance monitoring (all biological elements included) and one for operational monitoring (only pressure elements included). The final classification (ecological status) was obtained applying the one-out all-out approach for biological element classification. We apply a Revised Version of this Risk of Misclassification Approach for the current statistical analysis of EFI data.

2.3 Estimating components of variation

The surest way to gain an understanding of the components of variation affecting any given metric is to carry out a survey designed according to sound statistical principles. For example, the survey design outlined in Figure 1 might be suitable if we wished to estimate the following components of variation:

- between water bodies (of nominally similar type);
- site to site within a water body;
- year to year (possibly varying by site);
- seasonal variation (possibly varying by year or by site);
- local spatial variation; and
- sampling and measurement error.

An excellent account of such an exercise is described by Jones *et al* (2006).

Unfortunately, however, historical data sets associated with ideal programme designs are seldom found, and so in practice - initially at least - we have to make do with less comprehensive data sets. Nevertheless it is often possible to derive at least the main components of variation from historical data sets, and the following section describes one such exercise - for EFI data.

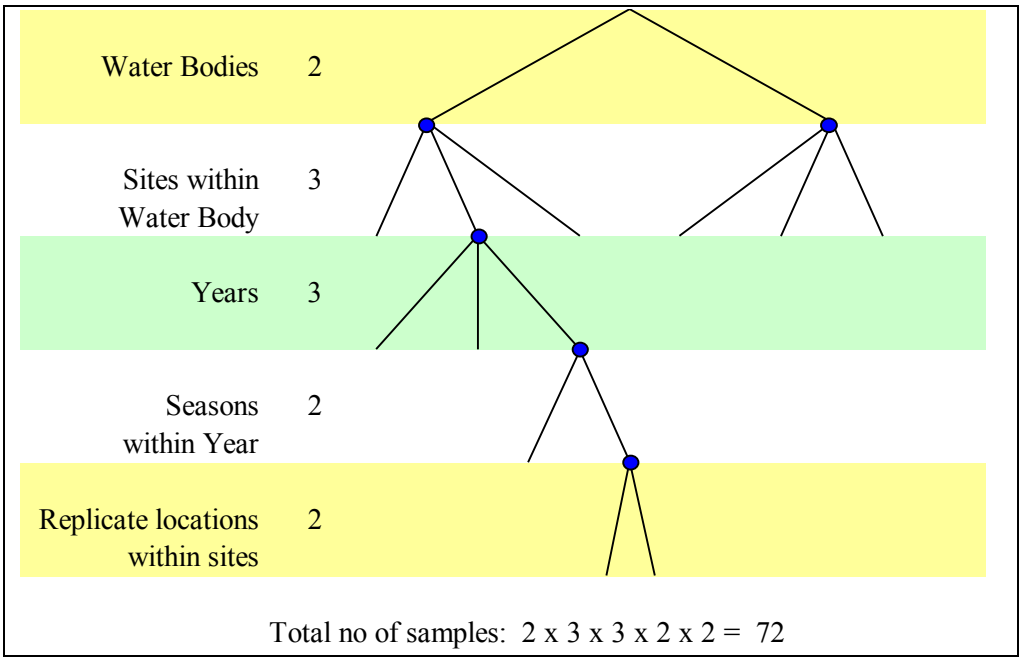


Figure 1 Example of a hierarchical survey design

2.4 Practical example - EFI

2.4.1 Overview of available data

For each of the seven river catchments, the sampling that was carried out formed natural sub-groups from which we can attempt to estimate at least some of the main components of variability, as follows:

On the **Nidd**, same-day sampling was carried out on four neighbouring 100m stretches for each of a number of sites. Thus the within-site variability provides a measure of local spatial variation plus sampling error.

On the **Don**, sampling at any location was carried out within a window of no more than three months. Sampling was carried out at typically 2-4 locations within each broad site area (for example, Shevock1, Shevock2, Shevock3). Thus the within-site variability again provides a measure of local spatial variation plus sampling error, with an additional component of temporal variation.

The sampling for the **Arrow**, **Avon** and **Leadon** was much longer-term in nature, with occasional repeat samples at any one site typically spanning a substantial number of years. For these three rivers, therefore, the within-site variability provides a measure of overall temporal variation (of widely varying duration) plus sampling error. Note that the temporal variation will in general be a combination of *trend* (i.e. a time pattern common to all sites) and *site-specific temporal variations* (known collectively as 'spatial-temporal interaction').

The sampling for the river **Stour** and tributaries was also long term, running from 1981 to 2005. The data was expected to be less prone to trends over time than that for other

rivers, because of the long history of improved water quality in the Stour. This makes it likely that natural temporal variability will have the largest influence on EFI. Thus the within-site variability will provide a measure of random temporal variation plus sampling error.

A sampling programme was conducted in the **Mersey** catchment during 2002 in which samples were taken in May and September at each of 16 sites. The within-site variability will therefore provide an estimate of the seasonal component of EFI, along with random temporal variation plus sampling error.

The **fish sampling procedure** for the Avon, Arrow, Leadon, Stour and Mersey is based on EA 'Fisheries Monitoring Programme Work Instruction 2.1', Electric Fishing in Rivers'. For the Nidd and Don, the fish sampling was part of a scientific study. The length of the river stretch sampled was close to 100m for each of the considered sampling sites.

The proposed **FAME EFI sampling method** will apply electrical fishing in these rivers over a length of 100m as well. Only first-run sampling data will be used for the calculation of the EFI. Hence second- and third-run fish sampling data (where it was available) was not used in this study.

2.4.2 Results - Local variability plus sampling error

As noted above, the variation shown by the repeat samples for the Don and Nidd sites is due to sampling and measurement error, plus local spatial and short-term temporal variation. The within-site standard deviations are plotted against mean EFI in Figure 2 below.

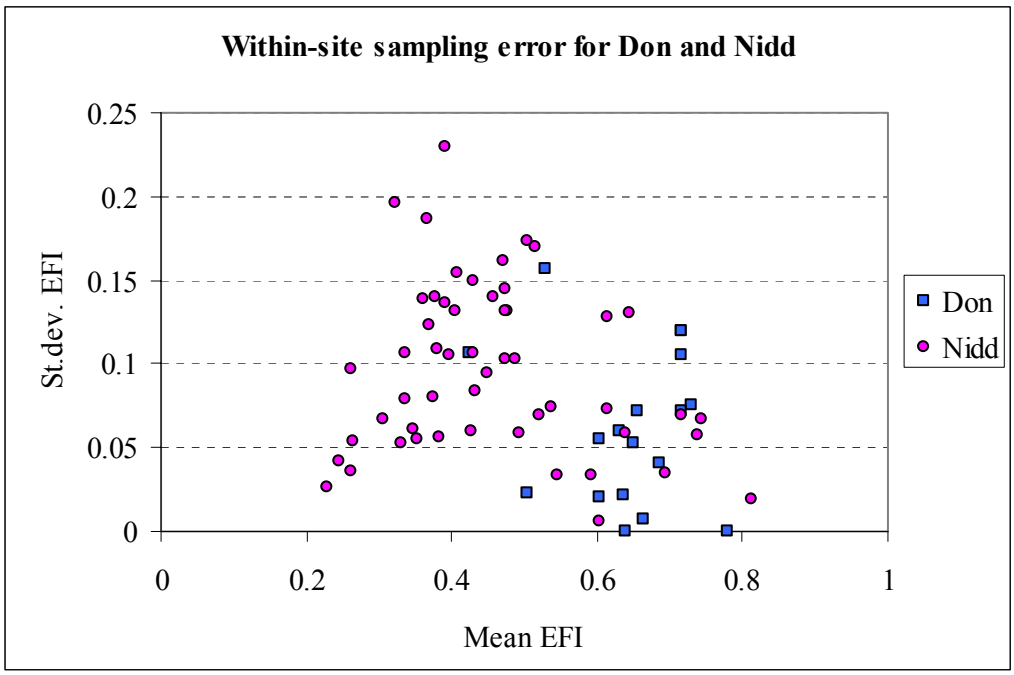


Figure 2 EFI within-site variability for the Don and Nidd

2.4.3 Results - Random temporal variability

Figure 3 shows the within-site standard deviations for (a) the Arrow, Avon and Leadon and (b) the Stour plotted against the number of years spanned by the data. There is a tendency for the standard deviations to increase with years spanned - as indicated by the regression lines - but the effects are relatively slight in comparison with the noise in the data.

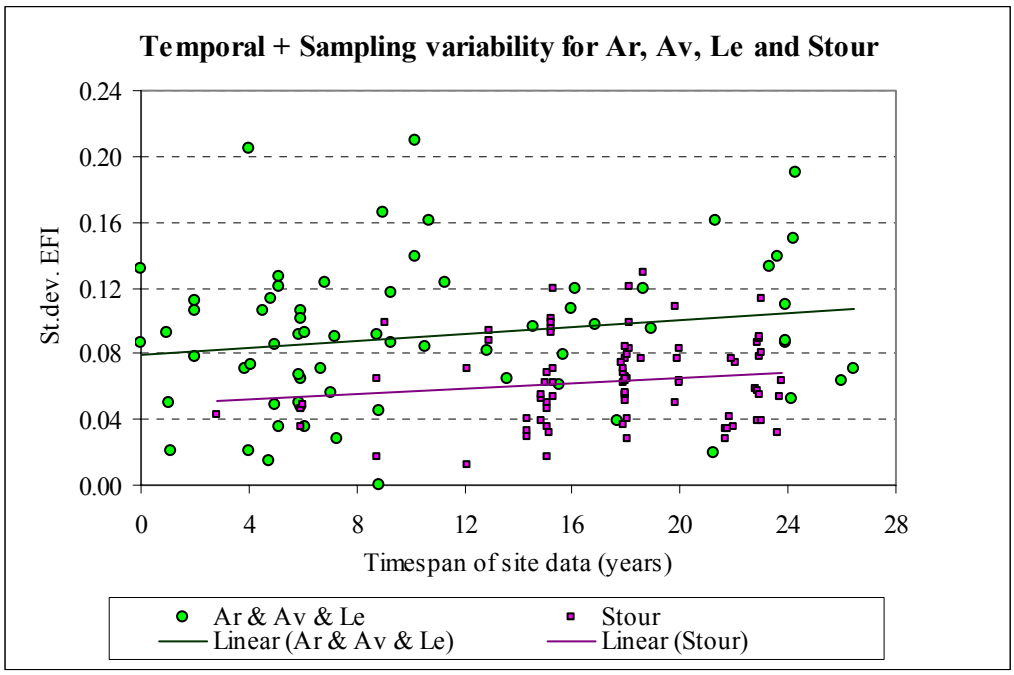


Figure 3 EFI variability for the Arrow, Avon, Leadon and Stour

Figure 4 shows the effect of adding the Don and Nidd data to this plot. These are broadly consistent with the leftmost points for the other four rivers.

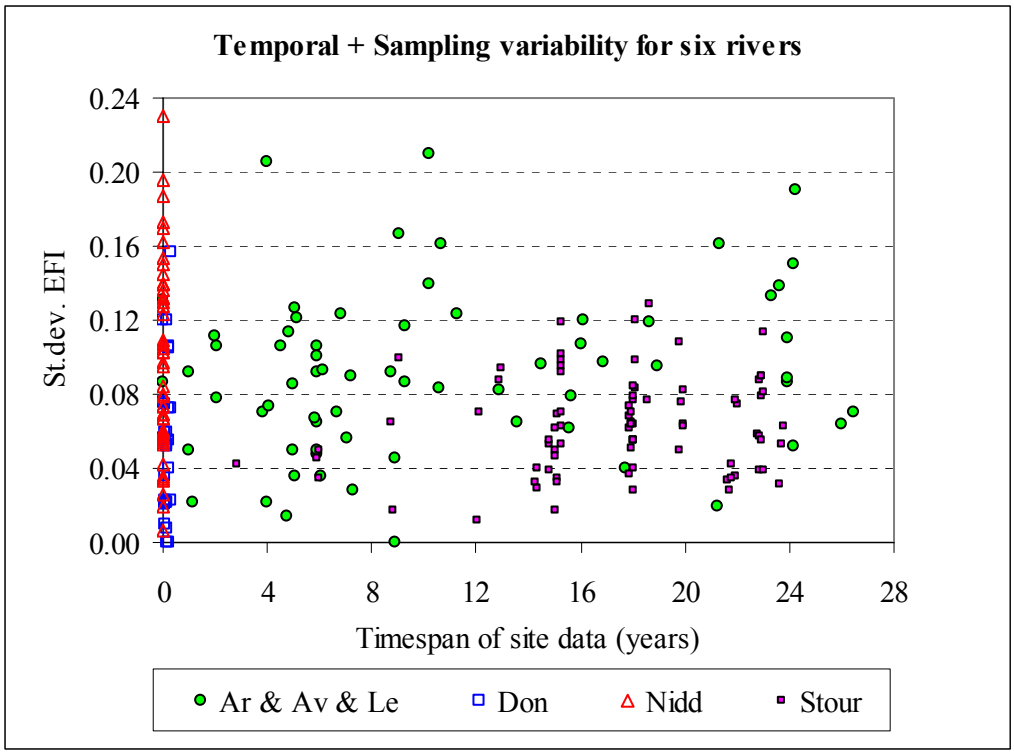


Figure 4 Within-site EFI variability for six rivers

2.4.4 Results - Seasonality

Figure 5 shows a plot of the May and September 2002 EFI values for each of 16 sites in the Mersey catchment. Between May and September, EFI increased at six sites and decreased at the other ten, demonstrating that there is no obvious seasonal effect common to all sites. This is confirmed by the slight mean decrease in EFI of 0.02, which is nowhere near being statistically significant.

This analysis reinforces the general message that seasonality (in the statistical sense indicated above) is not an issue with the EFI tool. This is partly because sampling can only take place in a fairly restricted spring-summer window of the year, which means that there is less opportunity for a seasonal effect to occur than with a tool whose sampling programme spans the full 12 months. The main point, however, is that the temporal variation in EFI at a site within that sampling window is driven by a variety of river- or site-specific conditions (e.g. flow, temperature, vegetation, spawning season, fish mobility) which are likely to swamp any simple seasonal effect. Consequently the main aim in monitoring is to visit each site just once in the year, at a time judged most likely to produce a representative picture of the fish populations and communities in the river.

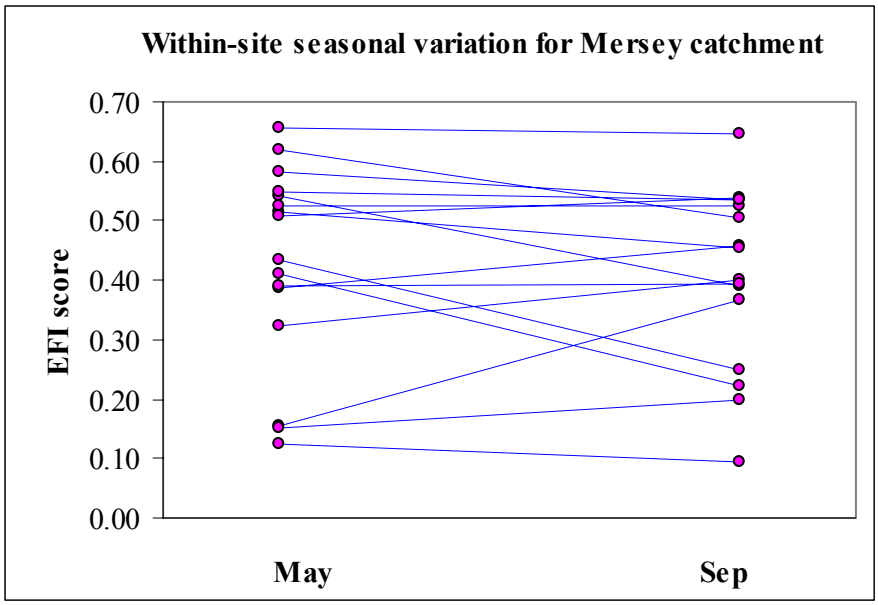


Figure 5 Seasonal variation in EFI values for sites in the Mersey catchment

2.4.5 Results - Spatial variability

There can be a substantial amount of site-to-site variation within a water body. This is illustrated by Figure 5, which shows a plot of site means for the Nidd according to position along the river. The figure also shows 90% confidence intervals for each site mean. It is clear that there is more site-to-site variation - especially in the left-hand water body - than can be explained by the size of the within-site variability.

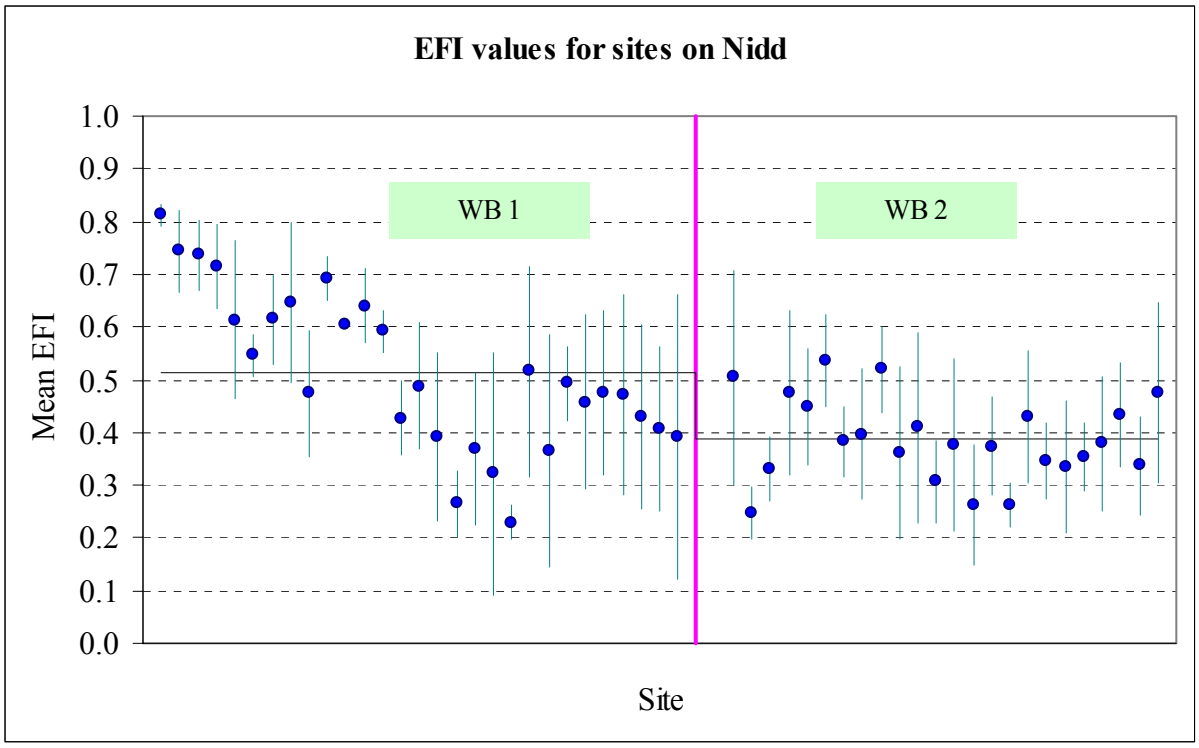


Figure 6 EFI means and 90% confidence intervals for the Nidd

As there is relatively little temporal component of variation in this data, Analysis of Variance (ANoVA) can be used to determine whether or not the between-site variation is statistically significantly greater than would be expected having regard to the within-site variation. If it is, then the excess component can reasonably be attributed to spatial variability. Moreover, this can be further split into *spatial trend* (as represented approximately by a quadratic model) and *random spatial variability*.

The ANoVA results are summarised below. The analysis confirms that the between-site component of variation is highly significant. There is also a highly significant spatial trend for water body 1, accounting for about 68% of the spatial variability. In contrast, the spatial trend component for water body 2 is only 7% of the total, and not statistically significant.

Component	**Water body 1**		**Water body 2**	
	Total SS	% of spatial	Total SS	% of spatial
Spatial trend	1.694	68.4	0.043	7.3
Spatial random	0.783	31.6	0.548	92.7
Total spatial	2.477	100.0	0.591	100.0
Residual	1.146		0.776	
Grand total	3.623		1.367	

2.4.6 Summary of ANoVA results

A more extensive analysis of the Arrow, Avon and Leadon data - both individually and pooled - was carried out using the statistical package GenStat. This broadly confirmed the more informal analyses described above. In particular, there was a statistically significant temporal component for only one of the three rivers - the Leadon. ANoVAs were also carried out for the EFI data for the Don, Nidd, Stour and Mersey. In all cases where the data extended over a number of years, care was taken to exclude any obvious environmental trend from the between-year temporal component by first fitting a linear or quadratic time trend model to the data, and then determining the *additional* component of variance accounted for by fitting a year factor.

A summary of the components of variance obtained by these various approaches is provided in Table 1.

Table 1 Approximate components of variability for EFI data

(a) as Variances

River	Residual	Seasonal	Betw-years	Spatial
Nidd WB1	0.0133			0.0251
Nidd WB2	0.0105			0.0049
Don	0.0050	0.0000		0.0040
Arrow	0.0120		ns	0.0028
Avon	0.0087		ns	0.0096
Leadon	0.0134		0.0015	0.0133
Stour	0.0046	0.0001	0.0003	0.0044
Mersey	0.0057	0.0000		not applic.

(b) as Standard deviations

River	Residual	Seasonal	Betw-years	Spatial
Nidd WB1	0.115			0.158
Nidd WB2	0.102			0.070
Don	0.070	0.000		0.064
Arrow	0.110		ns	0.052
Avon	0.093		ns	0.098
Leadon	0.116		0.039	0.115
Stour	0.068	0.007	0.016	0.067
Mersey	0.075	0.000		not applic.

Note:
1. Shaded cells indicate data sets where the specified variance component does not apply .
2. The 'Residual' term includes the Spatial-Temporal interaction component as well as measurement error.

3 Modelling variability at a site

3.1 Relating standard deviation to mean

Figure 7 reproduces the plot seen earlier (in Figure 2) of EFI within-site standard deviation against EFI mean. (Ignore the curve for the moment.)

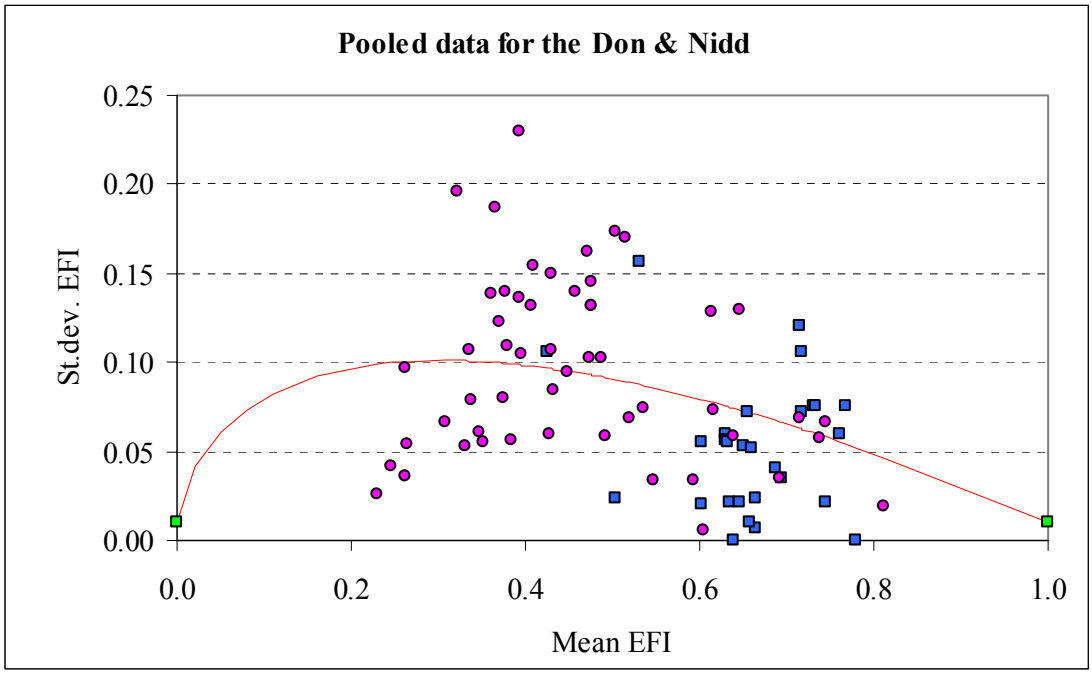


Figure 7 Within-site EFI standard deviation v. mean

At first sight the large amount of scatter may seem discouraging. However, two features should be noted:

1. For any 'vertical' slice through the plot, there is no discernible difference between the two rivers - which is reassuring.
2. A lot of the scatter - especially in the vertical direction - is inevitable because of the high level of statistical uncertainty in standard deviations calculated from small numbers of samples. For example, where a standard deviation s is based on four replicates from a Normally distributed population, the true standard deviation could fall anywhere in the interval (0.6s - 3.7s) with 95% confidence.

In order to develop a general approach for determining CofC, we need a model relating typical EFI standard deviation to EFI mean. One approach that we have developed is to fit a polynomial curve through the data, with the additional constraints that the curve passes through two 'anchor' points at EFI = 0 and EFI = 1. The choice of anchor points

should ideally be made on the basis of actual replicate data sets with means close to 0 and 1. In the absence of that, we have used arbitrary (but plausible) values of 0.01, as shown in Figure 7. The best-fit polynomial using these anchor standard deviations is shown by the red curve.

It is interesting to see what happens when the data for the other four rivers is included in the curve-fitting calculation (see Figure 8). Had there been an appreciable year-to-year component of variation, we would expect to see a general increase in the within-site standard deviations, and hence the curve to move upwards. No such effect is apparent, however; and the curve is actually pulled *down* appreciably in the 0.15 - 0.4 region because of the relatively low variability seen at the Stour sites (see Figure 3).

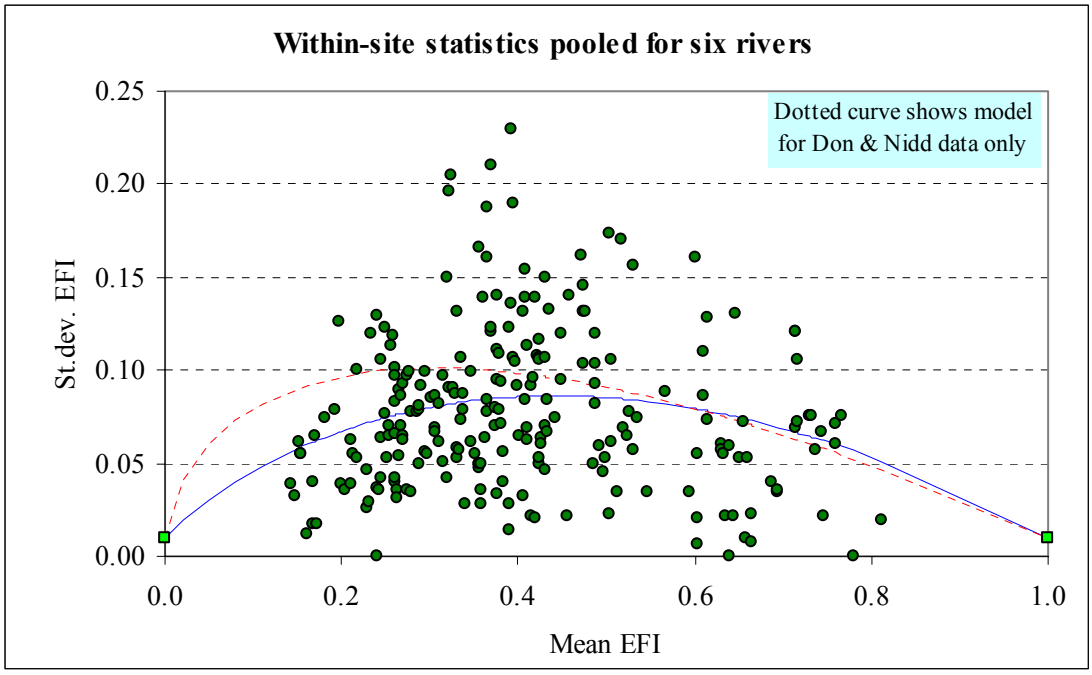


Figure 8 EFI within-site standard deviation v. mean for six rivers

3.2 Statistical distribution of variability

The previous section presented a model for the expected or typical standard deviation in EFI measurements at a site, given the mean EFI. Before we proceed to a consideration of CofC, the final step is to decide on a suitable statistical model for the uncertainty in EFI at a site.

The simplest option is to assume that the EFI uncertainty is Normally distributed around the specified true EFI value, with the predicted standard deviation. However, although this model is quite acceptable for most values of EFI it becomes unsatisfactory at either extreme, because the assumed Normal distribution 'spills' outside the allowed 0-1 range.

For this reason we have adopted the logit transformation, whereby the EFI (x) is transformed to a new variable z given by:

$z = \ln(x/(1-x))$,

where ln denotes 'logs to base e'. As x runs from 0 to 1 the transformed variable z runs from $-\infty$ to $+\infty$, and so there is no longer any risk of spillage. Thus we can safely use the assumption of Normal error in the logit world, and then transform the resulting distribution back into the EFI world.

This is easier to see with the help of a diagram. Figure 9 shows the situation in which the assumed EFI mean and standard deviation are 0.85 and 0.10. Under the simple Normality assumption, an appreciable part of the right-hand tail spills beyond EFI = 1. In contrast, the logit transformation ensures that the error distribution ends asymptotically at 1 (at the expense of a longer left-hand tail so as to achieve the required standard deviation).

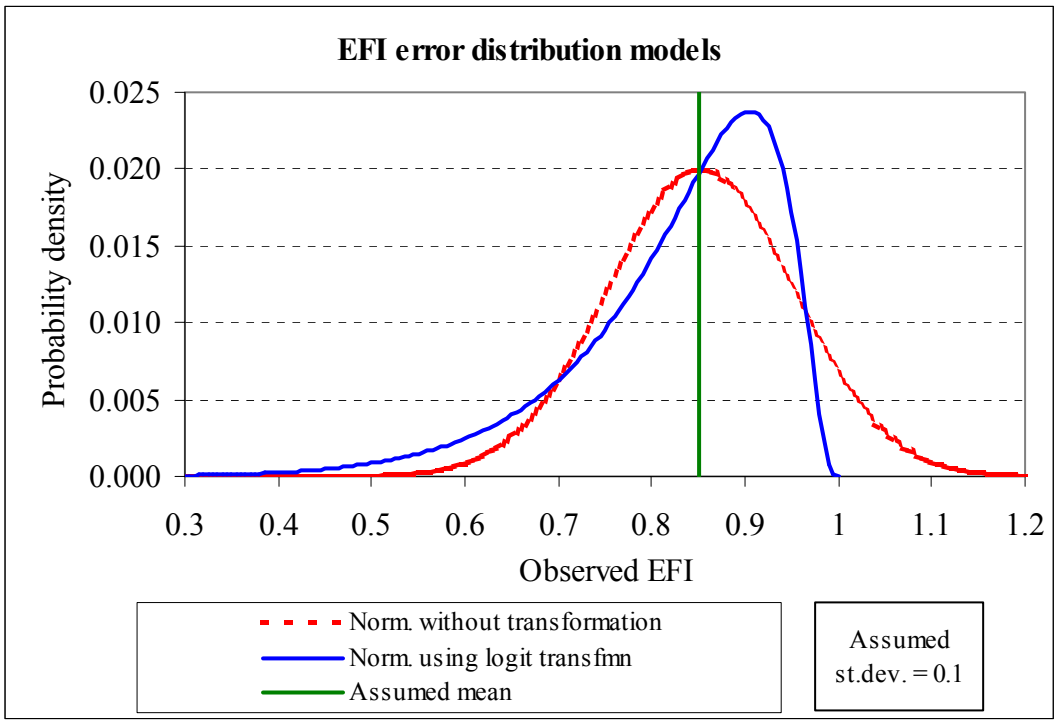


Figure 9 Illustration of the effect of the logit transformation of EFI

4 Confidence of Class

4.1 Forming the appropriate measure of variability

It was noted earlier that the variability relevant to assessing Confidence of Class depends critically on how the status of a water body is to be defined. To illustrate this point, we discuss five hypothetical scenarios below, and quantify each of them using the simple Excel calculation tool CAVE (Combines Appropriate Variance Estimates) illustrated in Table 2. It is important to emphasise that the full specification of each classification tool will in due course include an objective, unambiguous statement of how it is to be applied to a particular water body over space and time. Once that is known, the details can simply be plugged into CAVE to determine the relevant standard error. In the meantime, this exercise illustrates the substantial uncertainties that can be introduced where such a specification has not been made explicit.

Each scenario in Table 2 describes a particular way in which water body status might be defined, coupled with a proposed monitoring programme. The scenarios are as follows:

1. Status is defined by quality on one specific date at one specific location in the water body. One sample is taken at that location on the specified date.

2. Status is defined by mean quality over a 12-month period at one specific location in the water body. One sample is taken at that location at a randomly chosen time during the year.

3. Status is defined by mean quality over a 12-month period over the whole water body. One sample is taken at a random location in the water body at a randomly chosen time during the year.

4. Status is defined as in 3, but now *four* samples are taken at randomly chosen locations and sampling occasions.

The variability associated with the final EFI result is different for each of these scenarios, because it depends on (a) which components of variation may be ignored because of the way status is defined, and (b) how many samples are available to smooth the random components that do apply. CAVE provides a convenient template for working systematically through the components deciding which of them are relevant. The yellow cells indicate values to be supplied by the user, as follows:

- The variability components are specified by the row of seven standard deviation estimates;

- Nreps is the number of replicate samples to be taken at each time/location;

- Nt/s is the number of different times and/or locations to be sampled; and

- The six No/Yes values are set to 1 or 0 depending on whether or not the variance component is relevant in the context of the specified definition.

There is inevitably some degree of arbitrariness in supplying the seven input standard deviations, because it was not possible to identify them all from the historical data sets. The values currently built into CAVE are based on the plausible subdivisions of the Residual and Spatial components shown below. However, it is important to note that the sensitivity of the Confidence of Class outcome to any of these assumptions can easily be tested using the spreadsheet tool shortly to be described.

Variance component from ANoVAs		**How subdivided**
Residual	36%	Local temporal - measurement error & fish mobility
	64%	Site-specific temporal variability
Spatial	32%	Local spatial variability
	21%	Systematic spatial trend
	47%	Random spatial variability within water body

The resulting standard errors for the four scenarios addressed by CAVE are shown in the right-hand column of Table 2. Scenario 1 describes the simplest situation in which a single location is deemed to represent the whole water body (perhaps because it is in the area of the water body experiencing the greatest pressure), and a single sampling occasion is deemed appropriate (perhaps defined by some predetermined condition or time of year). The standard error (SE) is 0.078. This is substantially lower than that for the next two scenarios because only two components of variation are relevant: local sampling error and fish mobility, and local spatial variability.

With Scenario 2, the random temporal, seasonal and temporal-spatial interaction components now have to be included. This has a big impact on the SE, increasing it from 0.078 to 0.119.

Under Scenario 3 the spatial component of variability has to be included as well as the temporal. This further increases the SE from 0.119 to 0.139. (This is not so dramatic as before because the spatial variability is relatively small in relation to the spatial-temporal interaction term, which was introduced in Scenario 2.)

Finally, Scenario 4 shows what happens if three samples are taken at randomly chosen locations and times, rather than the single sample of Scenario 3. This reduces the SE by a factor of $\sqrt{3}$, from 0.139 to 0.080. This is almost exactly the SE under Scenario 1. In other words, the effect of widening the definition of water body status from a single specified location and time (Scenario 1) to the full water body over a year (Scenario 4) is to require a three-fold increase in samples to achieve a comparable level of precision.

Table 2 **How the definition of WB status influences the standard error of monitoring - four illustrations**

CAVE

Combines Appropriate Variance Estimates

Key

SDE	St.dev. due to meas. error and fish mobility
SDInt	St.dev. due to spatial-temporal interaction
SDseas	St.dev. due to seasonal cycle
SDtemp	St.dev. due to random temporal variation
SDlocal	St.dev. due to local spatial variability at a site
SDgrad	St.dev. due to systematic spatial trend along water body
SDspat	St.dev. due to random spatial variability
nReps	No of replicate samples taken at each site/sampling occasion
Nt/s	No of times and/or sites sampled

Component of variance

Scenario	Error	Tmp-Sp	Temporal		Spatial		
	Local	Interactn	Seasonal	Random	Local	System.	Random
	SDE	**SDInt**	**SDseas**	**SDtemp**	**SDlocal**	**SDgrad**	**SDspat**
	0.060	0.080	0.000	0.040	0.050	0.040	0.060
	0.0036	0.0064	0.0000	0.0016	0.0025	0.0016	0.0036
1							
	No/Yes	0	0	0	1	0	0
	Nreps	1					
	Nt/s	1					
2							
	No/Yes	1	1	1	1	0	0
	Nreps	1					
	Nt/s	1					
3							
	No/Yes	1	1	1	1	1	1
	Nreps	1					
	Nt/s	1					
4							
	No/Yes	1	1	1	1	1	1
	Nreps	1					
	Nt/s	3					

Type any desired values into the yellow cells...

Total N	1
Overall Var	0.0061
Overall SE	**0.078**

Total N	1
Overall Var	0.0141
Overall SE	**0.119**

Total N	1
Overall Var	0.0193
Overall SE	**0.139**

Total N	3
Overall Var	0.0064
Overall SE	**0.080**

Note: see text for definition of scenarios

Returning to Scenarios 1 - 3, the important point to emphasise is that, as the SE of the assessment statistic increases, so the Confidence of Class decreases (for any given monitoring outcome). This relationship is quantified in the next section.

4.2 Statistical method

4.2.1 Single site and sampling occasion

Suppose the four intermediate class boundaries are denoted by L_5, L_4, L_3 and L_2 (in the order Bad/Poor $\rightarrow$ Good/High). From our model of standard deviation versus mean EFI, we can determine the standard deviations that would apply when quality were truly at each of those boundaries. Let these be denoted by s_5, s_4, s_3 and s_2. We also assume that the statistical variation in the observed EFI quality at a particular time and place is Normally distributed (after transforming the EFI scale if necessary, as discussed earlier).

Now suppose we observe an EFI value of x. The aim is to determine the levels of confidence we have that the *true* quality (at the time and place of sampling) is respectively in Class 5, 4, 3, 2 and 1. To do this, we first do four calculations. For each class boundary 'i' in turn, we ask the question: What is the probability p_i of observing an EFI of x or better if the true mean quality, μ, were on the L_i boundary? This can be calculated as:

$$p_i = \Pr(X \geq x \text{ given } \mu = L_i) = 1 - \Phi\{(x - \mu)/s_i\},$$

where Φ denotes the cumulative Normal probability.

This probability statement says that $\Pr(X \geq \mu + u.s_i) = p_i$ (where u is the standard Normal deviate corresponding to $1 - p_i$). We can turn this into a confidence statement by inverting it in the customary way, giving:

$\text{Confidence}(\mu \leq x - u.s_i) = 100p_i$.

This enables us to make the following five statements:

- Confidence of class 5 = $100p_5$.
- Confidence of exactly class 4 = $100(p_4 - p_5)$.
- Confidence of exactly class 3 = $100(p_3 - p_4)$.
- Confidence of exactly class 2 = $100(p_2 - p_3)$.
- Confidence of exactly class 1 = $100(1 - p_2)$.

(Note that these five quantities sum to 100%.)

4.2.2 Several sites or sampling occasions

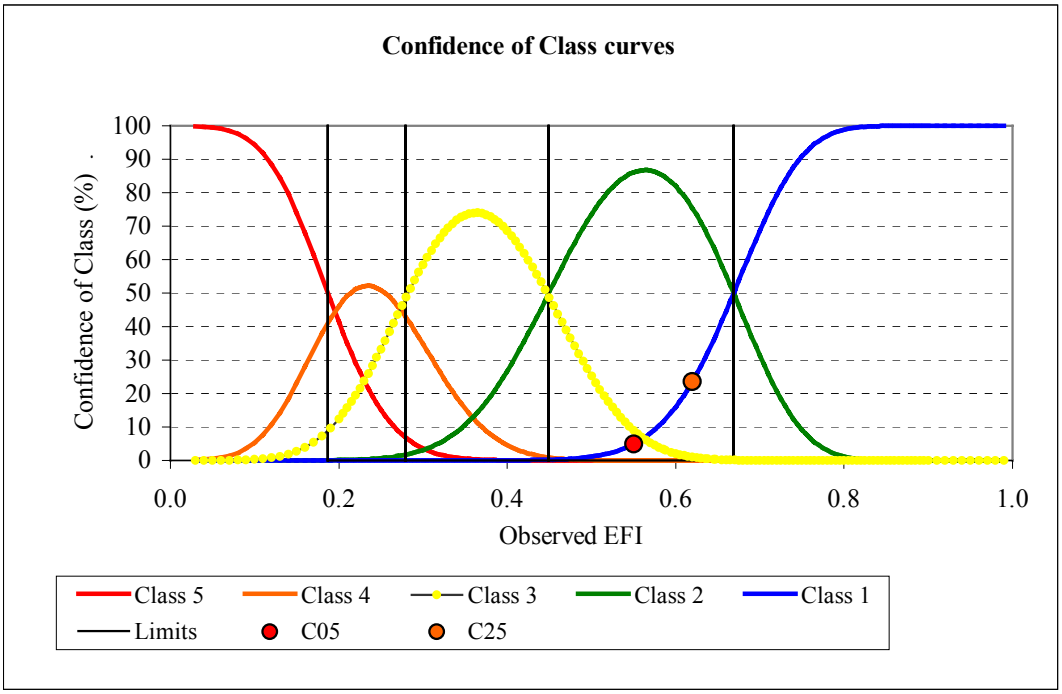
If the required statistic is *mean* EFI calculated across several sites, or for a number of sampling occasions at the same site, the relevant standard error is now a function of several components of variation (as discussed in Section 4.1). Consequently we no longer have a model showing directly how the standard error would change according to how far the true mean EFI was above or below the observed mean.

Nevertheless, we do know that as the true mean EFI moves close to 0 or 1, the standard error must decrease substantially, for otherwise the implied distribution would have too wide a tail to be plausible.

As a pragmatic solution, therefore, we have retained the simple one-sample model of standard deviation versus mean EFI at a site, and scaled it up or down so that the maximum of the curve produces the specified standard error. The CofC calculations then proceed exactly as before.

4.3 CofC for Scenario 1

This section deals with the simple case in which we choose to define water body status on the basis of a single sample at a particular site (Scenario 1 in Table 2). We can calculate the Confidence of Class (CofC) for that specific location and point in time using the method described in Section 4.1.1. The outcome is shown in Figure 10. Consider, for example, when the observed EFI is 0.2. We can see from the red, orange and yellow curves that there are three possible conclusions: the true Class may be Bad (with 50% confidence); Poor (40%); or Moderate (10%).



No of rep. samples:	1

C05	**0.550**
C25	**0.620**

Figure 10 Confidence of Class - Scenario 1

The legend below the figure gives information about two key points on the High curve. Thus, when the observed EFI is 0.55 (the red blob in the figure), we have only 5% confidence that the true class is High. In other words, we can be 95% confident that the true class is worse than High. Similarly, when the observed EFI is 0.62 (the amber blob), we can be 75% confident that the true class is worse than High.

We can recast the information of Figure 10 in a slightly different way, as shown in Figure 11. This plots the risk that a Face-Value interpretation of the observed EFI will put the site into an incorrect class. For example, when the observed EFI is very close to 0.45, namely on the border between Moderate and Good, it is not surprising that there is a 50% risk of making the wrong decision. Conversely when the observed EFI is 0.58, in the middle of the Good class, there is only a 13% risk (i.e. confidence) that the true class is not Good.

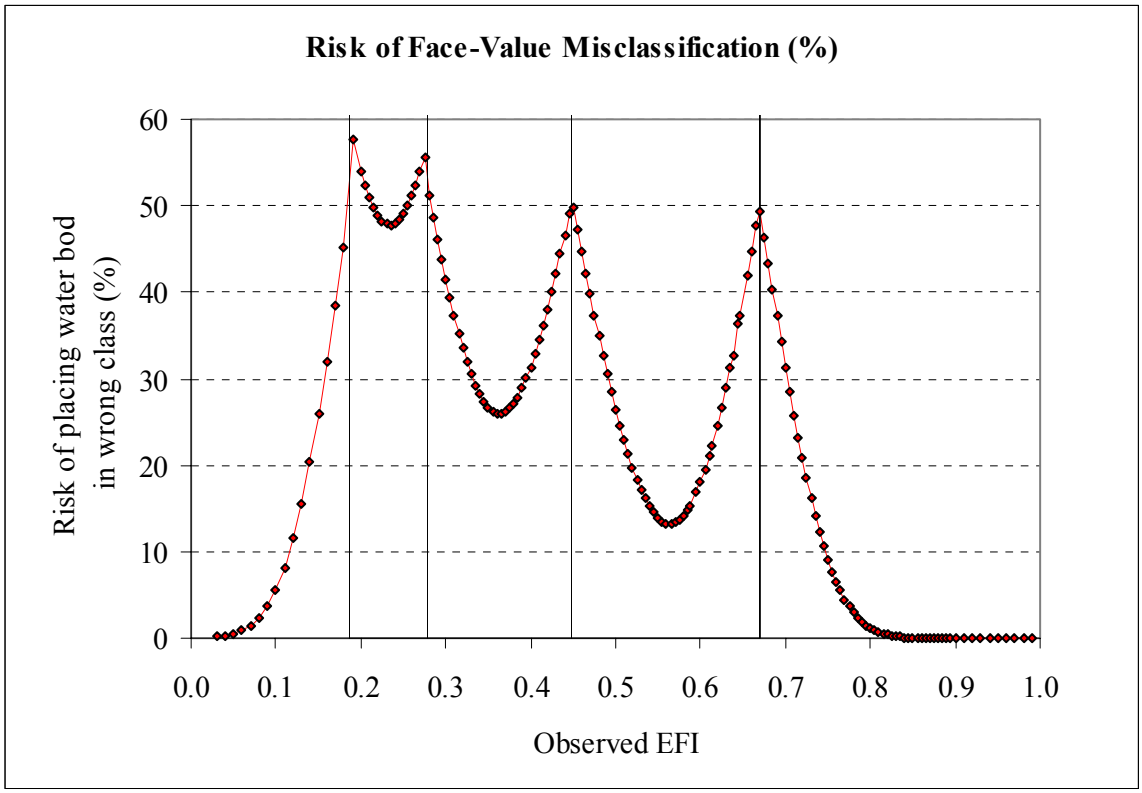
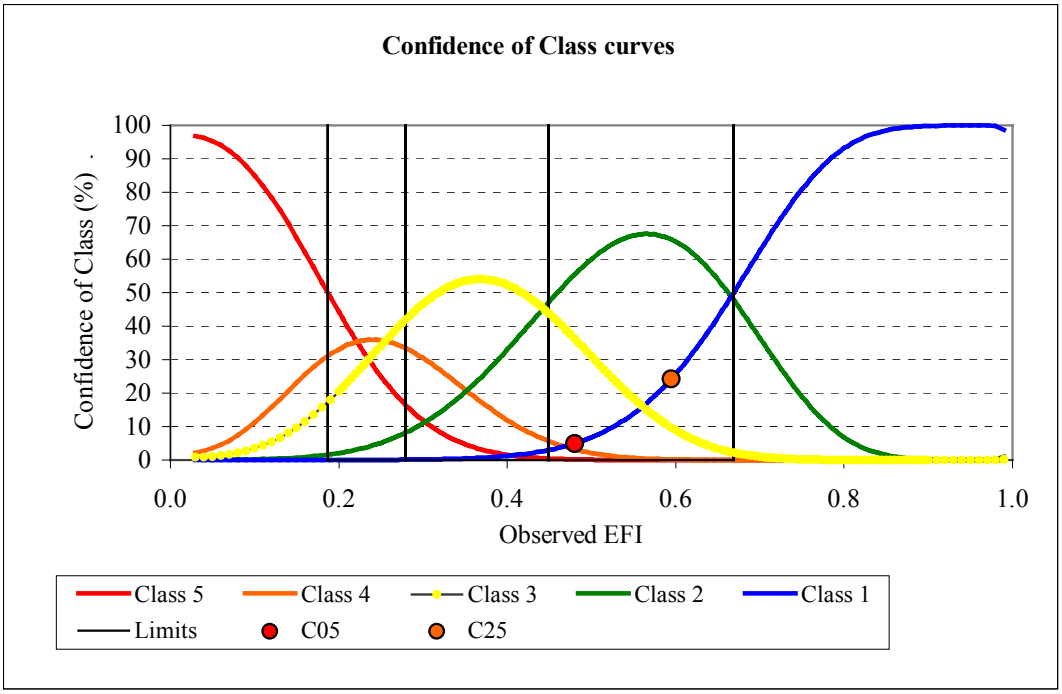


Figure 11 Risk of Face-Value Misclassification for the Figure 10 scheme

4.4 CofC for Scenario 2

Under Scenario 2, we still define water body status by observed quality at a specific location, but now the one sample chosen at random is supposed to represent the whole year. The revised version of the Figure 10 plot is shown in Figure 12. The curves are now much wider than before, indicating the poorer CofC (for any given EFI) brought about by the need to recognise the temporal uncertainty.



No of rep. samples:	1

C05	**0.480**
C25	**0.595**

Figure 12 Confidence of Class – Scenario 2

Figure 13 similarly shows the revised version of Figure 11. The risk of a Face-Value misclassification is still 50% or thereabouts at most of the class boundaries, as before; but for the relatively narrow class 4 the variability is simply too great for effective discrimination and the risk of misclassification rises to 70%.

With Scenario 3, the additional uncertainty introduced by the spatial component is relatively small, and so the CofC results for Scenario 3 are only slightly poorer than those seen here for Scenario 2.

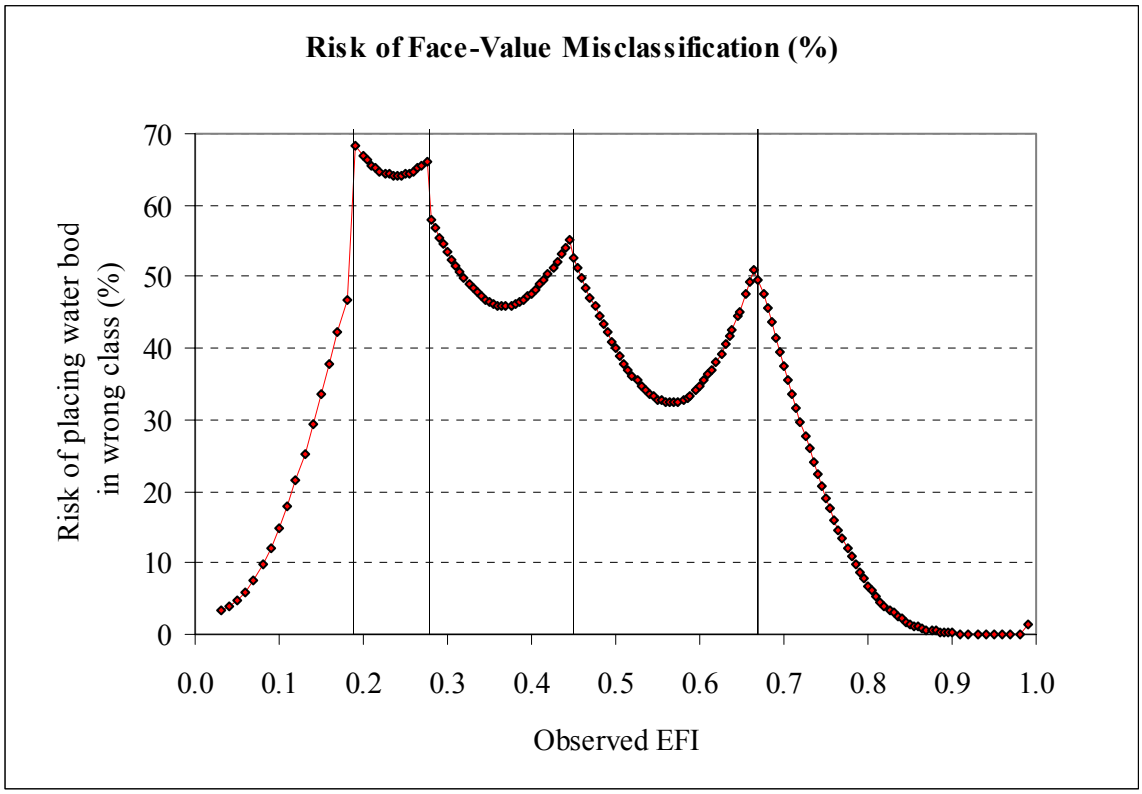


Figure 13 Risk of Face-Value Misclassification for the Figure 12 scheme

4.5 CofC for Scenario 4

Finally, Scenario 4 shows how, even when the definition of water body status requires the classification to reflect both spatial and temporal variability, the precision can be brought to a tolerable level if enough samples are taken. Here, three samples are taken from three different sites at randomly selected times. The resulting Risk of Misclassification plot is shown in Figure 14.

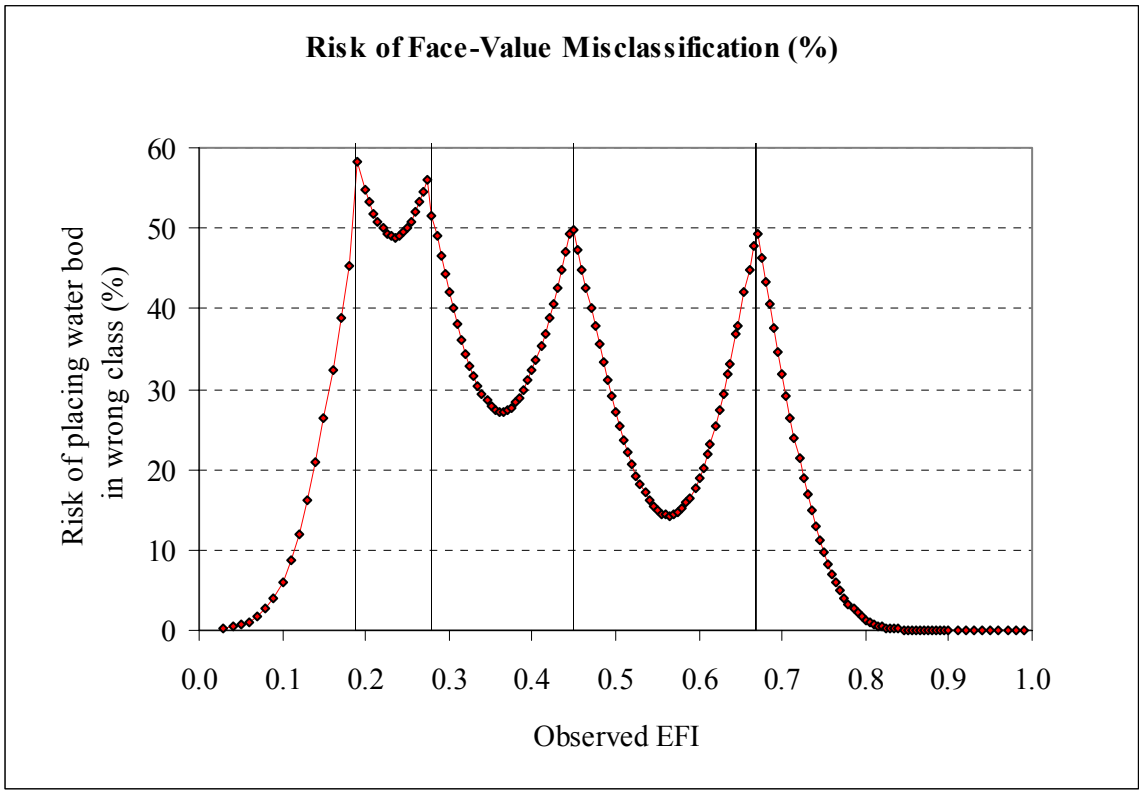


Figure 14 Risk of Face-Value Misclassification for Scenario 4

These examples illustrate the worsening effect on Risk of Misclassification of introducing additional components of variation. This underlines the importance, for any classification tool, of having a clear, unambiguous statement about the basis on which water body status is to be calculated.

References

Jones, I.J., Clarke, R.T., Blackburn, J.H., Gunn, R.J.M., Kneebone, N.T. and Neale, M.W. (2006). Biological quality of lakes: Phase II: Quantifying uncertainty associated with macroinvertebrate sampling methods for lake benthos. Phase II report. R&D Technical Report (13765)

Pont, D., Hugueny, B., Beier, U., Goffauz, D., Melcher, A., Noble, R., Rogers, C., Roset, N. and Schmutz, S. (2006). *Assessing river biotic condition at a continental scale: a European approach using functional metrics and fish assemblages*. Journal of Applied Ecology, **43**, 70-80.

List of abbreviations

ANoVA	Analysis of Variance
CAVE	Statistical tool in Excel - 'Combines Appropriate Variance Estimates'
CofC	Confidence of Class
EFI	European Fish Index
EQR	Ecological Quality Ratio
FAME	Fish-based Assessment Method for the Ecological Status of European Rivers
RIVPACS	River Invertebrate Prediction and Classification System
SE	Standard Error
WB	Water Body
WFD	Water Framework Directive