SNIFFER
SCOTLAND & NORTHERN IRELAND
FORUM FOR ENVIRONMENTAL RESEARCH

Environment
Agency

using science to
create a better place

Combining Multiple Quality Elements and Defining Spatial Rules for WFD Classification

SC060044/SR

ea/br/e/sci/v1

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: : 08708 506 506   Fax: 01454 624409
www.environment-agency.gov.uk

ISBN: 978-1-84432-803-1

The views expressed in this document are not necessarily those of the Environment Agency.

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.

This report is the result of work jointly funded by the Environment Agency and the Scotland and Northern Ireland Forum for Environmental Research (Sniffer).

Author(s):
Ellis J.E.

Dissemination Status:
Public Domain

Keywords:
Water Framework Directive (WFD), Uncertainty, Ecological Status, Classification

Research Contractor:
WRc

Environment Agency's Project Manager:
Veronique Adriaenssens

Collaborator(s):
WRc

Science Project Number: SC060044

Product Code: GEHO0807BMXZ-E-E

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 Group 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

Background / Need

The Water Framework Directive 2000/60/EC (WFD) requires surface waters to be classified through the assessment of ecological status and surface water chemical status. In developing the techniques required to implement this system, the Environment Agency and SNIFFER have collaborated on a number of related R&D projects to investigate the sources of uncertainty in the application of the classification tools and their statistical implications for the classification schemes.

The latest of these has been a project entitled: ‘Uncertainty estimation for monitoring for each of the WFD biological classification tools – Further work on classification, uncertainty and variability aspects‘.

Main objectives / Aims

The broad aim of the present project is to support the decision-making process of the UK TAG Classification Group. More specifically this is to be achieved by delivering statistical advice and recommendations on options for arriving at a water body (WB) classification (a) in WBs where we have multiple sample point data, and (b) where the site is assessed by two or more quality elements (QEs).

To advance the debate, the project team forwarded suggestions to a select group of EA and SEPA representatives who had been asked by the UK TAG Classification Group to bring some views on these classification issues to a workshop organised by UK TAG on 25-26 January 2007. Prior to that workshop, the representatives met on 23 January 2007– the aim of that preliminary meeting being to untangle the technical issues of classification and provide some discussion on the options that needed to be addressed.

This report describes the outcome of those 23 January 2007 discussions.

Conclusions / Recommendations

The conclusions and recommendations fell into four main areas:

Burden of proof and Confidence required

- It was generally accepted that the benefit-of-doubt stance should be adopted when assessing WB status.
- Most WB assessments proceed on the basis that at least 95% confidence is required before a WB can be declared to have failed. As formal agreement has yet to be reached, however, this issue should be forwarded to the UK TAG Classification Group.
- If a monitoring programme has poor statistical power (that is, has a low probability of detecting an unsatisfactory WB), it would be unwise to attempt to improve its performance by relaxing the required confidence level.

Combining Confidence of Class (CofC) information across sites

- Three methods were discussed for combining information across sites: a simple average, a weighted average, and a '% NotGood' approach.

- It was agreed that for Transitional and Coastal (TraC) waters, a '% NotGood' approach was appropriate. This would allow the WB as a whole to be classified as Good provided not more than some specified small percentage of the WB area was worse than Good. The choice of critical percentage (10%, say), is still to be decided.

- For heterogeneous WBs, a weighted average method may be appropriate, with the weights reflecting the known or assumed proportions of the WB in the Good and NotGood categories. However, practical difficulties would often arise in determining appropriate weights that were acceptable to all parties.

- For extreme cases of heterogeneity, the soundest option would be to revise the WB delineation where this was feasible.

- Where the WB was considered spatially homogeneous, it would be appropriate to use the simple average approach.

Combining CofC information for different QEs at a site

- Two principal approaches were discussed: the z-scores method for 'pooling evidence' from a number of QEs all believed to be reflecting a similar pressure; and the one-out, all out ('1oAo') method for determining the 'worst-case outcome' from a collection of QEs.

- It was agreed that, in practice, QEs would virtually always be reflecting different pressures, at least to some extent, and so the correct method to use would be the precautionary 1oAo approach.

Compensating for multiple assessments of a WB

- The greater the number of QEs monitored, the greater is the risk of a false positive - whereby a truly satisfactory WB is judged to have failed. A statistical method ('Bon Ferroni') is available that can compensate for this, but a decision needs to be taken about whether such a step is necessary or desirable.

Contents

Executive Summary **4**

Contents **6**

1 Introduction **7**

1.1 Background 7

1.2 The 23 January 2007 meeting 8

2 Confidence of Class **11**

3 Stance towards Burden of Proof **13**

3.1 The three possible stances 13

3.2 Level of confidence required 13

3.3 Statistical power of a monitoring programme 14

4 Combining CofC for a given Quality Element across sites **15**

4.1 Issues discussed 15

4.2 Possible ways of characterising a WB 16

4.3 Discussion 20

5 Combining CofC for different Quality Elements at a site **21**

5.1 Background 21

5.2 Outline of the two methods 21

5.3 Illustration using lakes data 24

5.4 Controlling the risk of false positives 25

5.5 Discussion 26

6 Worked Examples **28**

7 Conclusions **30**

References & Bibliography **36**

Glossary of terms **37**

List of abbreviations **38**

1 Introduction

1.1 Background

The Water Framework Directive (WFD) requires surface water classification through the assessment of ecological status and surface water chemical status. The classification of surface water bodies as required by the WFD can be illustrated by the diagram shown in Figure 1 (UK TAG, 2005).

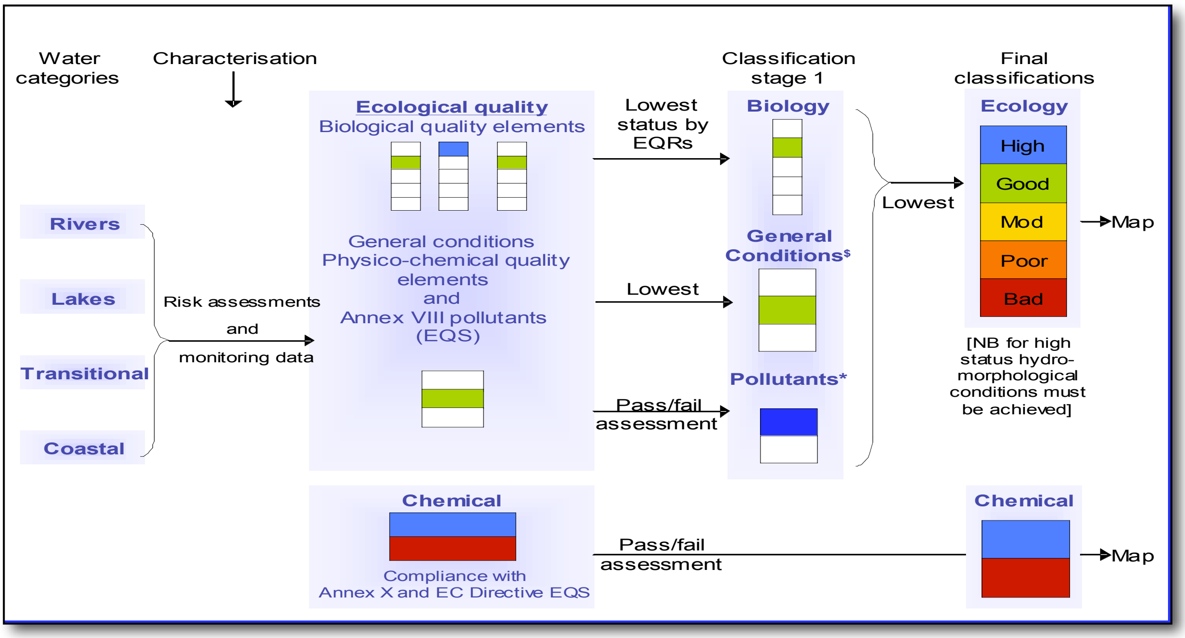


Figure 1 Classification of surface water bodies

As part of the process of developing the techniques required to implement this system, the Environment Agency and SNIFFER have collaborated on related R&D projects to investigate the sources of uncertainty in the application of the classification tools and their statistical implications for the classification schemes.

In a first stage of the project, we developed a statistical methodology that could be applied to ecological quality ratio (EQR) data, and this has been illustrated with the European Fish Index (EFI) data (Environment Agency, 2006). In a second stage a workshop was then organised for all tool developers to obtain an overview of uncertainty estimation methodologies and applications for each of the WFD biological classification tools and to allow further discussion (Environment Agency, 2007).

In a third stage, a project was initiated by the Environment Agency and SNIFFER entitled: 'Uncertainty estimation for monitoring for each of the WFD biological classification tools – Further work on classification, uncertainty and variability aspects'. The aim of this work is to deliver statistical advice and recommendations on (a) options for arriving at a classification in water bodies where we have multiple sample point data, and (b) how the overall classification process can work (i.e. biological and supporting elements). Julian Ellis (WRc) is providing statistical advice on these issues, and the project is intended to support the decision-making process of the UK TAG Classification Group.

To advance the debate on how we might classify a water body based on biological monitoring results from multiple sampling sites and multiple elements, we forwarded suggestions to a select group of EA and SEPA representatives who had been asked by the UK TAG Classification Group to bring some views on these classification issues to a workshop organised by UK TAG on 25-26 January 2007 in Edinburgh. Prior to that workshop, the representatives met on 23 January 2007 in Birmingham – the aim of this preliminary meeting being to untangle the technical issues of classification and provide some discussion on the options we wanted to consider.

This report describes the outcome of those 23 January 2007 discussions, together with some supplementary material provided by WRc.

1.2 The 23 January 2007 meeting

1.2.1 Attendees

The meeting was attended by the following people:

Participant	Title and WFD responsibilities
Veronique Adriaenssens	EA Science Environmental Biology Team - Project manager - WFD uncertainty and variability (SC060044)
Bill Brierley	EA Science Environmental Biology Team - Project manager - WFD uncertainty and variability (SC060044)
Richard Hemsworth	EA Policy Advisor WFD – EMCAR
Dave Jowett	EA Policy Advisor – Marine Task Team chair
Paul Logan	EA WFD programme manager
Geoff Phillips	EA Ecology Technical Team – Lake Task Team chair
Tony Warn	EA Policy manager (water quality)
Julian Ellis	Principal statistician, WRc

Note: Peter Pollard (SEPA, WFD programme) was unfortunately unable to attend.

1.2.2 Purpose of meeting

The purpose of the meeting was to discuss the uncertainty aspects of monitoring results generated by the WFD biological classification tools, with a special focus on (a) spatial considerations, and (b) classification based on multiple biological quality elements.

A fairly detailed agenda was circulated to participants beforehand; this is shown in Table 1. It is worth emphasizing the point made in the Overview section about the intention to leave ‘temporal variation’ off the list of discussion topics. This was explained in a footnote as follows:

> We think it will be unnecessary to spend much time on issues to do with temporal variation - the argument being that tool developers have already worked out the most appropriate way to deal with temporal variation for each Quality Element (QE), and will have properly reflected this in their calculation of the standard error and/or Confidence of Class (CofC) information associated with the EQR value at a site.

The sequence of items in the agenda had been planned to provide the most straightforward route possible through the various issues to be addressed. Similar considerations apply in organising the material covered in this report, and so we have based its structure on main headings 2 to 6 - but with the Worked Examples section moved to the end.

(Note that Heading 7 - Other topics - has been omitted as there was little discussion on these items, and in any case it was decided that the relevant key issues mainly hinged on policy decisions that fell outside the scope of the meeting.)

In some cases the position as set out beforehand - especially in the worked examples - turned out to be incorrect. Nevertheless we have retained these elements of the discussion, as it helps to illustrate the lines of reasoning that were developed during the meeting.

Where helpful, we have also included supplementary conclusions and points of clarification arising from several cycles of discussion between several of the participants in the week following the meeting.

Table 1 Agenda for the meeting

1	**Overview** Background Overall aims The worked examples One item not on the agenda - temporal variation
2	**Confidence of Class** There are five Classes - but in practice, are we usually dealing with just two? (That is, 'Good or High' versus 'Moderate or worse'.) Need to distinguish between: • Confidence of Class (CofC), and • Confidence of 'this Class or better' - and, in particular, Confidence of 'Good or High' (CofGoH).
3	**Stance towards burden of proof** Is 'Benefit of doubt' the accepted stance? Level of confidence required to declare a site as being NotGood: ≥95%?... 50 - 95%? What are the consequences for the 'statistical' power of monitoring programmes? (That is, how poor must a site truly be before there is a high probability of its being declared NotGood?)
4	**Introduction to the worked examples** These illustrate a three-stage hierarchical approach that we are tentatively proposing: 1. for any given QE, combine EQR results across sites by weighted average or a spatial percentile ('Spat%ile'); 2. for several QEs subject to the same Pressure, combine the evidence using the z-scores pooling approach; 3. for several different Pressures, combine the pooled evidence using the one-out, all out (1oAo) approach .
5	**Combining CofC information for a particular QE across sites** What summary parameter should be used for a Water Body? • mean EQR?... or • % of Water Body (WB) below Moderate/Good boundary? If the latter, what would be an acceptable NotGood %? 5%?... 10%? Types of monitoring sites to be accommodated by any general approach: • randomly (or representatively) selected sites; • targeted sites; • a mixture of random and targeted sites; • 'additional' sites selected locally.
6	**Combining CofC information for different QEs at a site** The Aug'06 debate (improving CofC by amalgamating data) Choosing between the 'one-out, all-out' and 'pooling' approaches Cake's Law (You can't have your cake and eat it)
7	**Other topics** • Supporting elements • Hydromorphs • Overall classification

2 Confidence of Class

In the project on uncertainty carried out in the early months of 2006 (Environment Agency, 2006), the focus was on determining CofC for each of the five classes, given an appropriate set of monitoring data for a site. This was accordingly one of the main outputs from the Excel calculation tool that was circulated at around that time.

Since then, however, the focus seemed to have shifted towards a simpler question, namely: Is the site truly above or below the Moderate/Good boundary? In essence, this would mean that the interest was now centred on which of 'two' (aggregated) classes the site falls into, not 'five'.

As it happens, the meeting never explicitly addressed this point, but it was evident from the discussions through the day that this was indeed the main focus. The issues raised can best be introduced with the simple example shown in Table 2. The bold row shows the CofC for each specific class - and so these five numbers add up to 100%. (These are the values that we assume can be calculated for any QE for any given monitoring programme.) In the second row, the CofC values are cumulated moving from right to left to give the confidence of 'this Class or better'. For example, we are 30% confident that the site is Good or High. Lastly, the third row shows the CofC values cumulated from left to right to give the confidence of 'this Class or worse'. For example, we are 10% confident that the site is Poor or Bad.

Note the arithmetic links between these two cumulative confidence measures. Taking the case of the two shaded cells, for example, we see that:

{Confidence of Good or better} = 100% - {Confidence of Moderate or worse}.

Table 2 Confidence of Class illustration

	Bad	Poor	Moderate	Good	High
Confidence of Class	**0%**	**10%**	**60%**	**27%**	**3%**
Conf. of this Class or better	100%	100%	90%	30%	3%
Conf. of this Class or worse	0%	10%	70%	97%	100%

In view of our particular interest in the Moderate/Good boundary, we have defined the following terms:

CofGoH to mean Confidence of 'Good or High'; and **NotGood** to mean 'Moderate or worse'.

Thus, using the principle noted above, Conf(NotGood) = 100% - CofGoH.

In the sections that follow, the discussion will generally be couched in terms of CofGoH - or its counterpart, Conf(NotGood). It is worth noting, however, that exactly the same principles apply for any other class boundary that is used to split the five-class system into two aggregated categories. For example, if we wish to apply the stand-still principle, the lower end of the current class would provide the required criterion.

3 Stance towards Burden of Proof

3.1 The three possible stances

In testing for compliance with a standard, there are three possible stances towards the burden of proof:

- ‘fail-safe’ – whereby observed quality must be somewhat ‘better’ than the standard to allow for the possibility that the true performance is marginally poor, but appears acceptable because of sampling variability;
- ‘benefit-of-doubt’ – whereby observed quality is permitted to be somewhat ‘poorer’ than the standard before a failure is flagged, to allow for the possibility that a truly acceptable performance is being distorted by sampling variability; and
- ‘face-value’ – whereby no allowance is made for the possible effects of sampling error.

Where it is a discharge that is being monitored, and the compliance samples are taken by the regulator, there is a compelling argument in favour of a benefit-of-doubt (‘B-of-D’) stance, on the grounds that the discharger should not be penalised for any bias that may be introduced as a consequence of limitations in the regulator’s effluent monitoring programme. Similarly, the Environment Agency’s Rivers Ecosystem assessment of river quality at a site against the required River Quality Objective also takes a B-of-D approach - the argument being that it is unfair to fail a river (and hence penalise those organisations discharging into it) if the evidence might simply be the result of the regulator’s sampling error.

In the current context of WFD Water Body (‘WB’) assessment, it appears that B-of-D is again the accepted stance. Not everyone, of course, will be happy about giving the benefit of the doubt to the polluter - but this is the economic and political reality of current environmental regulatory policy. Furthermore, the approach brings the important positive benefit that it controls the rate at which monitoring throws up false positives, thereby ensuring that remedial effort is not wasted on sites that are in truth satisfactory. This point is argued very persuasively in a recent paper by Tony Warn - “Classification: dealing with the effect of errors in monitoring” (4 Jan 2007).

3.2 Level of confidence required

Having decided on a B-of-D approach, the next decision to make is what level of confidence is required before declare a site to be NotGood. Prior to the meeting, there seemed to be a fairly widespread view that at least 95% confidence would be required; but there was also some support for a degree of flexibility, with anything between 50 and 95% being required according to the context.

This issue was the subject of considerable discussion during the meeting. As the topic arose particularly in the context of multiple Quality Elements, we defer reporting on this until Section 5. For the moment, however, it may be useful to illustrate, in the context of the CofC data in Table 2, exactly what is implied by requiring at least 95% confidence of failure.

First of all, look at it from the standpoint of ‘Confidence of this class or better’. Moving from right to left, we must give the site the B-of-D as soon as the confidence has climbed above 5%. Thus, we cannot claim that the site is High, as we have only 3% confidence in this; but we can say with better than 5% confidence (30%, in fact) that it is Good or better. The site would therefore be deemed Good.

Equivalently we can assess the CofC data from the standpoint of ‘Confidence of this class or worse’. Moving from left to right, we give the site the B-of-D for as long as the confidence stays below 95%. The final class for which this holds is Moderate. Thus, we cannot claim that the site is Moderate or worse, as we have only 70% confidence in this. In other words, Conf(NotGood) is 70%. As this has not reached the required confidence criterion of 95%, we must therefore accept (as before) that the site may be Good or High.

It may seem that this is going into an unnecessary amount of detail. However, it is easy to get confused over whether one is talking about confidence of ‘failing’ a criterion or confidence of ‘meeting it’ - and the point of this illustration is to demonstrate that these are just two sides of the same coin. A site can be assessed using either measure, and both approaches come to the same thing.

The final point to make here concerns 50% confidence. In essence, requiring only 50% confidence before declaring that a WB is NotGood is the same thing as taking a face-value approach and ignoring sampling error. It would mean that a site which was truly borderline (that is, the true value of its QE was 0.60) would have a 50% risk of being declared NotGood. This does not seem a very desirable state of affairs in the light of the earlier remarks about the practical benefits of controlling the risk of false positives.

The meeting discussion concluded that agreement had yet to be reached on this issue, and that it should be forwarded to UK TAG Classification for further deliberation.

3.3 Statistical power of a monitoring programme

The question posed to the meeting was;

- What are the consequences for the statistical power of monitoring programmes? That is, how poor must a site truly be before there is a high probability of its being declared NotGood?

The discussion around this issue centred on one key concern. With the relatively modest monitoring programmes generally envisaged, there may be cases where it is not actually possible to demonstrate with sufficiently high confidence that a site was NotGood, however poor the true quality was. The view was expressed that, in these circumstances, it might be necessary to relax the confidence criterion. However, others felt that this would be unwise. If the monitoring signals contain too much statistical noise to be very useful, this is an important piece of feedback that should inform the future monitoring, rather than be ‘papered over’ by slackening the assessment criteria.

4 Combining CofC for a given Quality Element across sites

4.1 Issues discussed

A great deal of time was spent by the meeting on this general topic. It is a complex area for two reasons. First, there are several possible ways in which CofC information can be statistically combined across sites to produce a single value characterising the WB. Secondly, the applicability or suitability of each option depends very much on the interactions between three key factors, as outlined in Table 3 below. These are:

1. the number of sites in the WB
2. what is known spatially about the WB
3. how the site locations had been selected.

Table 3 Common spatial scenarios

Number of sites	Spatial nature of WB	Method of site selection
Substantial (say 12 or more)	Homogeneous	Random/representative
	Unknown	Random/representative
	Homogeneous with known impacted area(s)	Random + Targeted
A small number (say 2 - 4)	Homogeneous	Random/representative
	Unknown	Random/representative
	Homogeneous with known impacted area(s)	Random + Targeted

There was much discussion on all these issues, and the main points are summarised in the sections following.

4.2 Possible ways of characterising a WB

4.2.1 Using Mean EQR

If the sites have been chosen representatively, and the aim is to characterise average quality over the WB, then the arithmetic mean EQR could be used. Furthermore, if the WB were believed to be homogeneous, the sampling error could be assumed to be at least approximately Normally distributed, and so it would be straightforward to calculate Confidence of NotGood. The WB would then be failed if this exceeded 95% (say).

4.2.2 Using Weighted Mean EQR

A variant of the above approach might be appropriate if it were known that the WB was not homogeneous, but could be stratified into known ‘good’ and ‘less good’ parts. Such a situation is illustrated in Figure 2 - one of the diagrams used in the meeting to help focus the discussion.

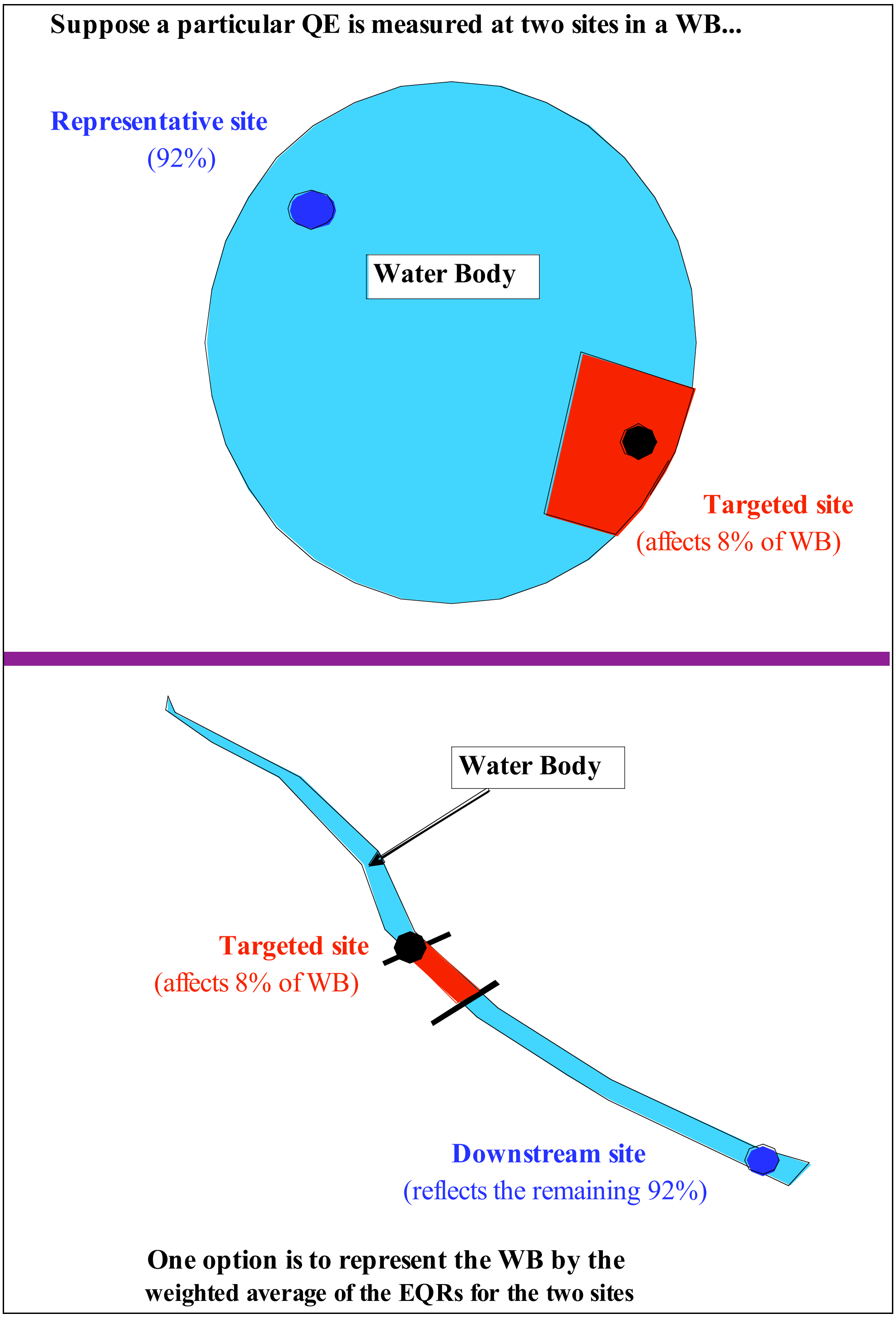

Suppose a particular QE is measured at two sites in a WB...
Representative site
(92%)
Water Body
Targeted site
(affects 8% of WB)
Water Body
Targeted site
(affects 8% of WB)
Downstream site
(reflects the remaining 92%)
One option is to represent the WB by the
weighted average of the EQRs for the two sites

Figure 2 Example of random and targeted sites in a WB

In these circumstances we could plan to have one or more random sites in the 'good' region and one or more targeted sites in the 'less good' region, and then calculate a weighted average of mean EQRs in the two regions. (The figure happens to show only one site per region, but the appearance of the diagram in a more general situation can readily be imagined.)

Of course, this approach does assume that we know the proportional split of the WB into the two regions (92% v. 8% in the example). But even where it is not known, an agreed approximation would be a big improvement on simply applying an unweighted average approach - which, in the example of Figure 2, would be equivalent to making the unrealistic assumption that the two areas each constituted 50% of the WB.

4.2.3 Using % NotGood

The third approach is more stringent, in that it requires nearly all of the WB to meet the Moderate/Good limit - but does allow some specified small percentage to be NotGood. The approach was set out in a paper by Tony Warn entitled “Spatial Considerations in Classification” (15 January 2007).

An example of the situation that might arise is shown in Figure 3 (another diagram used in the meeting). Here we see that a non-parametric estimate of % Good is simply 100 × 10/12, or 83%; and so the estimated % NotGood is 17%.

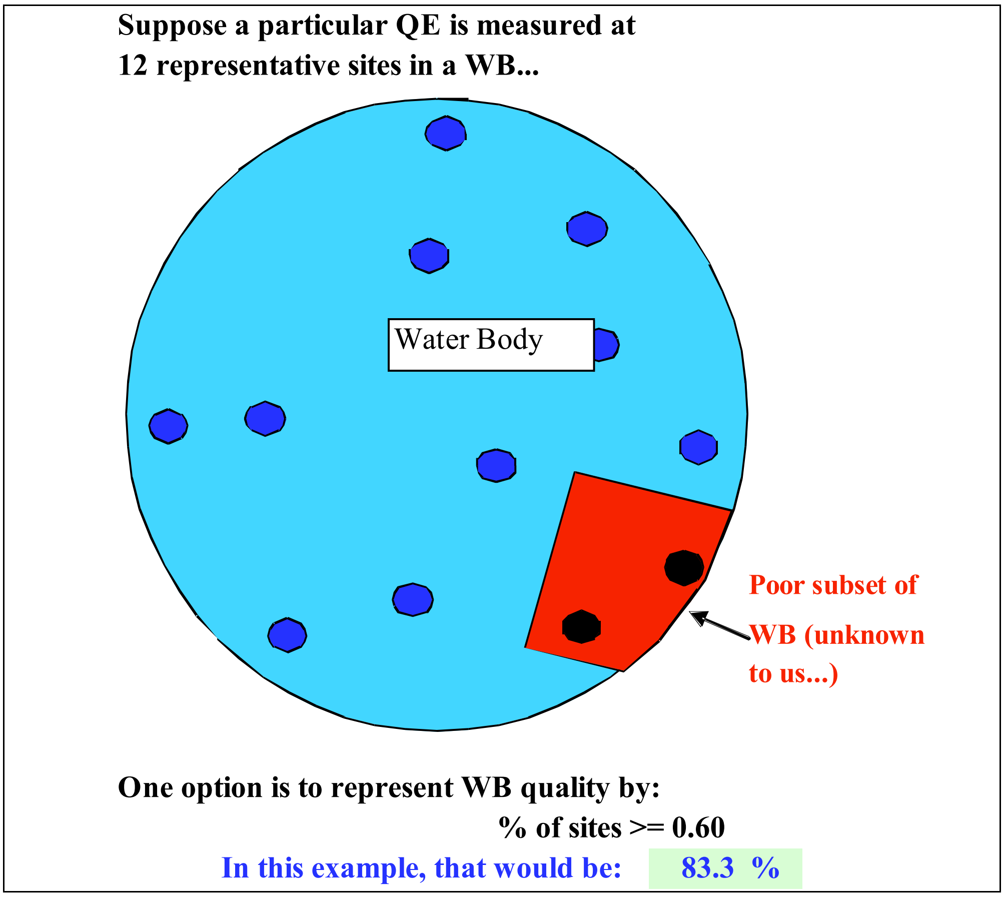
To apply this method, some judgement would be needed as to what constituted an acceptably low threshold. Criteria of 5% and 10% were debated, but the consensus view was that a decision would need to be made by the classification tool developers. (This was further discussed at the UK TAG Jan 25-26 workshop on classification, and Dave Jowett, Peter Pollard and Tony Warn have been asked to resolve this issue).

Once an agreed criterion was in place, however, it would be possible to calculate the confidence with which the estimated % NotGood exceeded this[1].

Clearly the % NotGood approach is more stringent than either of those based on averaging - and there was much discussion on their pros and cons. It was agreed that the approach was appropriate for Transitional and Coastal (TraC) waters, given that they were generally very large. However, it does still depend on the objectives for the WB: is a certain amount of degradation allowed or not? This issue has been discussed further in a recent paper by Dave Jowett and Tony Warn - “Spatial consideration of Ecological Status Assessment in Transitional and Coastal Waters” (14 December 2006).

The approach could also be appropriate for other WB types, such as very long rivers. However, no firm conclusion could be reached about the general circumstances in which this might be appropriate. Admittedly, the opportunity to apply the % NotGood method would arise in only a minority of WBs. However, it was nevertheless important to establish the principle in advance - and it was agreed that this was another decision needing to be referred to a higher body.

[1] Three methods for doing this are proposed by Tony Warn in his paper: two were originally described in an earlier version of his paper, and the third is a refinement of method 2 suggested by Julian Ellis.



Suppose a particular QE is measured at
12 representative sites in a WB...

One option is to represent WB quality by:
% of sites >= 0.60
In this example, that would be: 83.3 %

Figure 3 Example of random sampling locations in a WB

4.2.4 Testing the homogeneity of WB quality

One question raised was how to proceed when the EQR results raised doubts over the supposition that the WB was homogeneous. Several approaches based on testing the homogeneity of the between-site data were discussed. One would involve carrying out an outlier test - but this would be insufficiently powerful, given the likely number of sites. A more robust approach would be to rely on the tool developer having determined a typical 'UK-wide' spatial standard deviation for that particular QE and WB type - S_{typ}, say. We could then calculate the between-sites standard deviation; and if this was statistically significantly greater than S_{typ}, we could declare that the WB was not homogeneous.

The meeting discussed two basic options that could be followed in the event of the WB being demonstrably inhomogeneous:

1. **Revise the WB delineation** There may well be River and TraC WBs that could usefully be split based on evidence from monitoring. This would be reinforced where there were distinct management/quality zones within the WB. However, given the bureaucracy involved in altering the risk assessment maps already lodged with Europe means, one would not want to do this more that once every river basin cycle.

2. **Use a weighted average approach** (as described earlier).

With either option, the need for substantial local knowledge and expert judgement was agreed to be crucial.

4.3 Discussion

The session ended with a general discussion on the following main issues:

Effect of WB area

In determining the number and nature of monitoring sites, the area of the WB is an important consideration, and we do need to look at the spatial context. One key question then is this: at what point do we decide that it is not appropriate to use the average score for monitoring sites in a WB? And which of the alternative options discussed above would then be favoured in what circumstances? One problem with the more explicitly statistical methods - testing for non-homogeneity, or calculating a weighted average - would be the difficulty of producing a standard, unambiguous set of instructions for WB assessment that could be followed by non-specialists.

It was noted that the UK TAG Classification Group would be making proposals in its February meeting.

Types of monitoring site

Any general approach needs to accommodate both:

- sites that are representatively selected from the WB (excluding any designated mixing zones)
- targeted sites.

A mixture of these two types of site was felt unlikely to occur. Even so, if the situation could happen (albeit rarely), there needs to be a protocol agreed in advance for how the resulting data is handled.

5 Combining CofC for different Quality Elements at a site

5.1 Background

The issue of how to combine CofC information for several different QEs has been the subject of discussion for a number of months. A new impetus was given to the debate by a paper written by Geoff Phillips, Peter Pollard and Tony Warn, at the request of the UK TAG Classification Group, entitled "Improving confidence in classification by amalgamating data " (1 August 2006). This triggered a fruitful sequence of email exchanges between the authors and Julian Ellis, and this led to several further papers, including "Combining levels of confidence" (Ellis, 8 August) and a revised version of the earlier UK TAG paper (15 August).

Two principal approaches crystallised from this activity: the **z-scores** method for 'pooling' evidence from a number of QEs all believed to be reflecting a similar pressure; and the **one-out, all out** ('1oAo') method for determining the 'worst-case' outcome from a collection of QEs. Both approaches received a lot of discussion during the meeting. We first outline and illustrate the two approaches below. The subsequent section then summarises the main issues that were discussed.

5.2 Outline of the two methods

5.2.1 Pooling evidence by the z-scores method

The z-scores method relates to the situation where we are testing a particular hypothesis - that the site is truly 'Good or High', say - and we have evidence from several different QEs. It is particularly useful where the evidence from the QEs is suggestive but no one test is conclusive. By pooling the evidence appropriately, we hope that the collective evidence does become conclusive.

Given a number (m) of QEs, the method works as follows:

1. For each of the QEs, take the tail p-value associated with the significance test (p = 1 - Conf/100), and convert this into the equivalent standard Normal deviate. This is the 'z-score'.

2. Calculate z(pooled) as the sum of the m z-scores divided by √m.

3. Finally, convert z(pooled) back to the equivalent tail probability, and hence determine the overall confidence level associated with the hypothesis of interest.

The example shown in Table 4 below - adapted slightly from a lakes example discussed in the meeting - illustrates how the method works. The hypothesis being tested is that the site is ‘Moderate or better’. There is weak evidence for this from QE1, but fairly strong evidence from the other three QEs (between 85% and 92% confidence). When we combine these four pieces of evidence by the z-scores method, we find that the pooled confidence is 97%.

Table 4 Illustration of the z-scores method

		QE1	QE2	QE3	QE4
Conf. of Moderate or better (%):		**51.0**	**85.0**	**92.0**	**90.0**
Corresponding z score:		-0.025	-1.036	-1.405	-1.282
Sum of z scores:	-3.748				
Sum / root(4):	-1.874				
Corresponding Normal probability:	0.0305				
Hence pooled confidence (%):	**97.0**				

It is reassuring to note that, had we chosen to test the converse hypothesis, namely that the site was ‘Poor or worse’, we would have obtained exactly the same strength of conclusion (namely **3.0**% confidence of ‘Poor or worse’). That is because the first row of individual confidence levels would now be 49, 15, 8 and 10%, and their z-scores would be numerically the same as in the above table, but with the opposite sign.

5.2.2 Quantifying the worst case by the one-out, all-out method

A quite different stance from that taken with the z-scores method is the 1oAo method. Here, we declare that the QEs are all designed to respond to different pressures, and so any one QE showing a demonstrably poor performance is sufficient to fail the site.

One slight complication with the 1oAo approach is that there are two variants of the method - which has proved to be an occasional source of confusion in the past. To illustrate these, we start with the example shown in Table 5, where we have the five CofC values for each of two QEs measured at a site (see the yellow shaded cells).

For any QE, we wish to be at least 95% confident before downgrading a site from any presumed class. As discussed in Section 3, that is equivalent to saying that, once we have greater than 5% confidence that the site may be in ‘this Class or better’, we give the site the benefit of the doubt.

So, working from High downwards, we conclude that according to QE1 the site may be ‘Good or better’ according to QE1, but according to QE2 may only be ‘Moderate or better’.

Judging the site by the poorer of the two QEs, therefore, we would classify the site as ‘Moderate’ - and this is what we have called **Interpretation I** of the 1oAo approach.

To quantify the level of confidence associated with this conclusion, the method used by the Environment Agency is to pick the confidence level achieved by the poorest of the

QEs (which will necessarily have been responsible for putting the site in that class in the first place). This can be done for each of the five classes, as shown in the row of the table labelled ‘EA version’. Thus we see that CofGoH is 3%, for example, whilst confidence of ‘Moderate or better’ is 20%.

Table 5 Illustration of the one-out, all-out method

% confidence of *exactly* this Class

	Bad	Poor	Mod	Good	High
QE1	10.0	70.0	14.0	5.0	1.0
QE2	5.0	75.0	17.0	3.0	0.0

% confidence of this Class *or better*

	Bad	Poor	Mod	Good	High
QE1	100.0	90.0	20.0	6.0	1.0
QE2	100.0	95.0	20.0	3.0	0.0

***1oAo* % confidence of this Class or better (EA version)**

Bad	Poor	Mod	Good	High
100.0	90.0	20.0	3.0	0.0

***1oAo* % confidence of this Class or better (WRc version)**

Bad	Poor	Mod	Good	High
100.0	85.5	4.0	0.2	0.0

This approach has the important merit of consistency. That is, it ensures that the overall calculation produces the same B-of-D class (in this case Moderate) as that reached by the process of working through each individual QE, as described earlier. It does, however, have the slight drawback of having no formal statistical justification. For this reason an alternative approach is worth consideration. In essence, this is based on combining the cumulative confidences at any given Class as though they were probabilities. (A justification for this is given in WRc’s August 2006 paper.)

Take, for example, the CofGoH values associated with the two QEs. We see that this is 6% for QE1, and 3% for QE2. Clearly, therefore, it is very unlikely that both QE1 *and* QE2 are truly achieving the status of Good or High - and this is reflected by the joint confidence calculated as 6% × 3%, namely 0.18%. (See the row labelled ‘WRc version’.)

By this approach we see that our confidence of ‘Moderate or better’ is only 4% (i.e. 20% × 20%). This is the same as saying that we are 96% confident that the site ‘fails*’* to achieve Moderate status. We must therefore step down a class and classify the site as ‘Poor’ - and this is what we have called **Interpretation II** of the 1oAo approach.

It is fairly evident why the conflict arises. Under Interpretation I, we are giving the site ‘two’ generous applications of B-of-D before picking the poorer of the two conclusions;

whereas under Interpretation II we are using an intrinsically stricter statistical process to combine the CofCs prior to a ‘single’ application of B-of-D.

There seemed broad agreement amongst meeting participants to endorse the pragmatic method of calculation embodied in Interpretation I.

5.3 Illustration using lakes data

Discussion in this part of the meeting centred around some illustrative CofC results for five lakes: this had been provided by Geoff Phillips immediately prior to the LTT meeting on 9 January 2007 (and discussed in that meeting). There was ‘genuine’ CofC data for Chla and CPET, and ‘estimated’ CofC for macrophyte and diatoms, based on the values of the EQRs in relation to the Class boundaries. The overall CofC results produced by the various statistical methods discussed above are reproduced in Table 6 below.

Table 6 Overall CofC results for five lakes

Water Body ID	**Assessment**	**Overall Conf. of this Class or better**				
	method	**Bad**	**Poor**	**Mod**	**Good**	**High**
Bassenthwaite Lake						
	1oAo (EA)	100.0	100.0	100.0	**55.0**	0.0
	1oAo (WRc)	100.0	100.0	100.0	**37.2**	0.0
	z-scores pooled	100.0	100.0	100.0	**99.9**	0.0
Sunbiggin Tarn						
	1oAo (EA)	100.0	100.0	**51.0**	0.0	0.0
	1oAo (WRc)	100.0	100.0	**44.5**	0.0	0.0
	z-scores pooled	100.0	100.0	**100.0**	0.0	0.0
Rollesby Broad						
	1oAo (EA)	100.0	100.0	**50.0**	0.0	0.0
	1oAo (WRc)	100.0	100.0	**47.5**	0.0	0.0
	z-scores pooled	100.0	100.0	100.0	**27.2**	0.0
Llyn Rhos-ddu						
	1oAo (EA)	100.0	100.0	**50.0**	0.0	0.0
	1oAo (WRc)	100.0	100.0	**47.5**	0.0	0.0
	z-scores pooled	100.0	100.0	100.0	**20.5**	0.0
llyn Helyg						
	1oAo (EA)	100.0	100.0	95.0	**20.0**	0.0
	1oAo (WRc)	100.0	100.0	95.0	**19.0**	0.0
	z-scores pooled	100.0	100.0	100.0	**99.5**	0.0

Note: cells shown in bold indicate the Class resulting from a B-of-D 95% confidence rule

The table illustrates two main points. First, there is little practical difference between the EA and WRc variants of the 1oAo rule. All five lakes are put into the same Class by either option, and the CofC values are not more than a few percentage points apart except for Bassenthwaite Lake.

The second point is that, as we would expect, the z-scores method produces a more favourable conclusion than does 1oAo. In two cases it puts the lake in one Class higher;

and for the other three lakes, the CofC by z-scores is much higher (over 99%) than it is by 1oAo (50% or less).

In the meeting, Geoff Phillips made the interesting observation that these results had persuaded him to go for 1oAo in preference to the z-scores pooling approach. He felt that, if a phosphorus pressure were present, then - depending on the lake - one or more of chlorophyll, diatoms, macrophytes or CPET would be likely to show NotGood status at a confidence of 95% or greater.

He went on to raise the question of whether, in the scenario described above, the status assessment should be modified if fewer than four elements are monitored - and, if so, how? This led to a brief discussion on the ‘multiple comparisons’ problem (as it is generally referred to in statistical circles). Shortly after the meeting he suggested the use of the ‘Bon Ferroni’ correction. This prompted an email debate between him, Tony Warn and Julian Ellis - the main outcomes of which we summarise in the next section.

5.4 Controlling the risk of false positives

The problem

Imagine a site that is exactly borderline in respect of a number of QEs. That is, the true value of each QE - as could in principle be determined by a high-frequency monitoring programme - sits exactly on the Moderate/Good boundary. In statistical parlance, this is the Null Hypothesis. (The ‘Alternative’ Hypothesis is that one or more of the QEs is in truth ‘worse’ than borderline.)

If we take a B-of-D stance at the 95% confidence level (as we have been assuming in much of the foregoing discussion), there will necessarily be a 5% risk that the results from any one QE will accidentally trigger a failure. It follows that, if the QEs are independent (as far as their sampling error is concerned), the risk of a false positive will progressively increase as more QEs are introduced into the assessment. In fact, the risks would be 5.0%, 9.8%, 14.3% 18.5% and 22.6% as the number of QEs increased from 1 though to 5.

This is an example of the ‘multiple comparisons’ problem. The more bites we have at the cherry, the greater is the risk of a false positive.

The Bon Ferroni solution

There are numerous statistical solutions to the multiple comparisons problem, according to the particular circumstances (such as the number and type of comparisons we state beforehand that we wish to make). One simple but effective approach is the ‘Bon Ferroni’ method. This works as follows.

Suppose we wish to operate at an overall confidence level of C = 95%, but intend to apply m separate significance tests. The Bon Ferroni solution is to carry out each of those individual tests at a higher level of confidence, C_{BF}, defined as:
$C_{BF} = 100(C/100)^{(1/m)}$.
For example, if m = 4 (as in the lakes example) and C = 95%, then

$C_{BF} = 100(95/100)^{0.2} = 99.0\%$.
This tells us that, if we test each of the five QEs at the 99% confidence level, and any one of them fails, we can say with 95% confidence that the site as a whole has failed.

A counter argument
There is a counter argument to the use of multiple comparison correction methods such as Bon Ferroni. As Tony Warn put it during some recent email exchanges: "What is to stop us adding in lots of pristine clean quality elements and so cranking down the test applied to the single one that is worse than all the others?"

This is a real concern. Consequently there is considerable merit in taking the following stance:

- declare that each site is to be assessed by just two or three carefully selected QEs;
- nominate that we intend to apply the '95% confidence of failure required' rule to each of them individually; and
- accept that the overall risk of failure under the Null Hypothesis will be more like 10 or 15%, but take the odd additional marginal failure on the chin. (After all, how many sites are actually likely to be exactly borderline?)

5.5 Discussion

Choosing between the 'one-out, all-out' and 'pooling evidence' approaches
Prior to the meeting, the project team had felt that the z-scores approach would be the natural candidate in cases where several QEs all reflected a similar pressure. (Indeed, this point of view was incorporated into the worked examples described in the next section.) However, the general feeling in the meeting was that 1oAo was the appropriate method in all cases - principally because, in practice, QEs would virtually always be reflecting different pressures, at least to some extent.

Pooling metrics within the same QE
Several participants suggested that the z-scores approach might well provide a useful way of combining different metrics within the same QE. However, subsequent reflection indicated that this would not be a valid application. The z-scores method is designed for combining levels of 'statistical confidence' in a particular hypothesis, not for combining the real 'environmental variation' shown by the different elements of a multi-metric QE. That is one of the key responsibilities of the tool developers, whose task is to form whatever arithmetic combination of those components is needed to ensure that the resulting QE responds most closely to the relevant pressures within the WB.

Cake’s Law

The term ‘Cake’s Law’ - stating that, ’You can’t have your cake and eat it”’ was coined during the discussions of August 2006. In the WFD context, it is shorthand for saying that one cannot set out to analyse the data in one way, only to switch subsequently to an alternative statistical method simply because it looks as though it might give a more useful or acceptable answer.

For example, suppose we are assessing compliance with the Moderate/Good boundary on the basis of two QEs. Cake’s Law says that we are **not** in general allowed to apply the following two-pronged rule:

- if Conf(NotGood) is ≥95% for either of the QEs, downgrade the WB (that is**, use 1oAo**);
- otherwise pool the evidence of the Conf(NotGood) values for the two QEs by the **z-scores** method and downgrade the WB if that produces a Conf(NotGood) value ≥95%.

In summary, therefore, any rules that are used for evaluating monitoring data should have been agreed in advance of seeing the data - and this important maxim was endorsed by the meeting.

6 Worked Examples

Shortly before the meeting, the project team developed a collection of 13 hypothetical examples. These provided plausible CofC data of varying degrees of complexity for:

- a number of QEs (between 1 and 4), reflecting
- a number of pressures (between 1 and 3), at
- a number of targeted and/or representative **sites** in the WB (between 1 and 13).

The examples represented Rivers, Lakes, and Transitional & Coastal Waters.

In each case, the aim was to demonstrate an objective procedure for arriving at an overall CofGoH value for the WB. We tentatively proposed a three-stage sequential approach, as illustrated in Figure 4, whereby:

1. for any given QE, EQR results were averaged across sites (by one of the two methods already discussed);
2. for several QEs subject to the same Pressure, the evidence was pooled (using the z-scores pooling approach); and then
3. for several different Pressures, the pooled evidence was combined using a 1oAo approach.

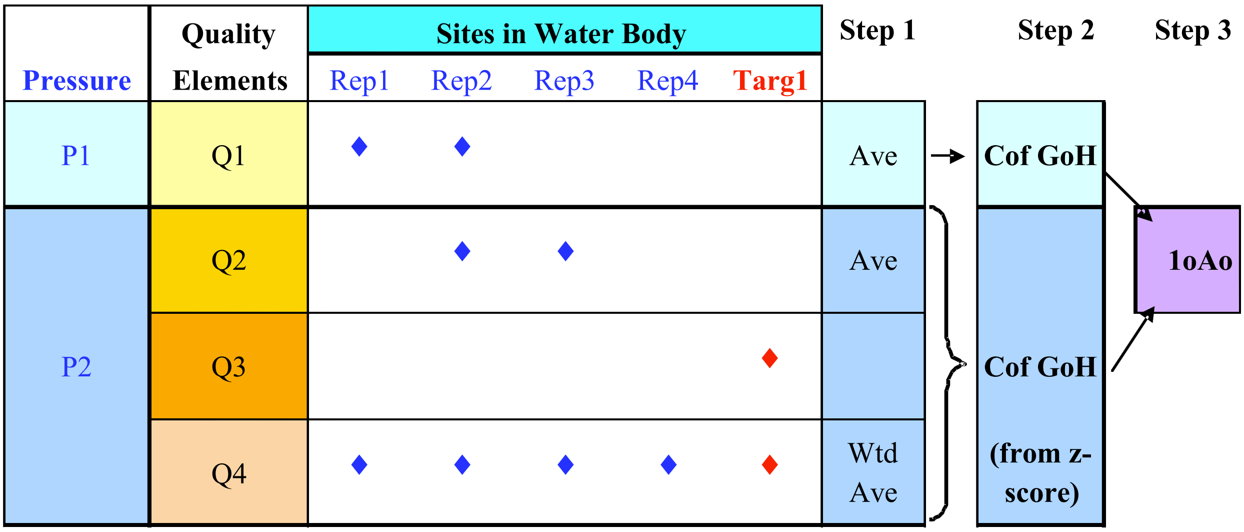


Figure 4 Summary of the sequential approach used in the worked examples

The worked examples are summarised in Table 7. Details of the statistical calculations are discussed in Appendix A.

As noted earlier, the general feeling of the meeting was that stages 1 and 3 of the suggested sequential approach were broadly sound, but that stage 2 (i.e. use of the z-scores method) would seldom be appropriate. We have nevertheless retained the examples in their original form - partly as a record of the material presented to the meeting, but also to show how the z-scores methodology works for those cases where it **is** thought appropriate.

Table 7 Summary of the 13 worked examples

Example	WB			Type of site		Analysis stages in arriving at		
no.	type	QE	Pressure	D/s or Repr.	Targeted	*Conf. of Good or High*		for WB
1	River	Q1	P1	R1		CofGoH		
2	River	Q1	P1	R1		CofGoH		1oAo
		Q2	P2		T2	CofGoH		
3	River	Q1	P1	R1		CofGoH	z-scores	
		Q2	P1	R1		CofGoH		
4	River	Q1	P1	R1	T2	wtd av	z-scores	
		Q2	P1	R1	T2	wtd av		
5	River	Q1	P1	R1		CofGoH	z-scores	
		Q2	P1	R1		CofGoH		
		Q3	P1	R1		CofGoH		
6	River	Q1	P1	R1		CofGoH		1oAo
		Q2	P2	R1		CofGoH	z-scores	
		Q3	P2	R1		CofGoH		
7	River	Q1	P1	R1		CofGoH		1oAo
		Q2	P2	R1	T2	wtd av	z-scores	
		Q3	P2	R1	T2	wtd av		
8	Lake	Q1	P1	R1		CofGoH	z-scores	
		Q2	P1	R1		CofGoH		
		Q3	P1	R1		CofGoH		
		Q4	P1	R1		CofGoH		
9	Lake	Q1	P1	R1		CofGoH	z-scores	1oAo
		Q2	P1	R1		CofGoH		
		Q3	P1	R1		CofGoH		
		Q4	P1 & P2	R1		CofGoH		
10	Lake	Q1	P1	R1		unwtd av		
		Q1	P1	R2				
11	TraC	Q1	P1	R1 - R12		CofGoH*	z-scores	
		Q2	P1	R13		CofGoH		
12	TraC	Q1	P1	R1 - R12		CofGoH*	z-scores	1oAo
		Q2	P1	R13		CofGoH		
		Q3	P2	R1 - R12		CofGoH*	z-scores	
		Q4	P2	R13		CofGoH		
13	TraC	Q1	P1	R13		CofGoH	z-scores	1oAo
		Q3	P1	R13		CofGoH		
		Q2	P1 & P2	R13		CofGoH		
		Q4	P2	R1 - R12		CofGoH*	z-scores	
		Q5	P2	R13		CofGoH		

* using Spat%ile method

7 Conclusions

Burden of proof and Confidence required

- It was generally accepted that the benefit-of-doubt (B-of-D) stance should be adopted when assessing WB status.
- Most WB assessments proceed on the basis that at least 95% confidence is required before a WB can be declared to have failed. However, the meeting discussion concluded that agreement had yet to be reached on this issue, and it should be forwarded to the UK TAG Classification Group for further deliberation.
- If a monitoring programme has poor statistical power (that is, has a low probability of detecting an unsatisfactory WB), it would be unwise to attempt to improve its performance by relaxing the required confidence level.

Combining QEs or CofC information across sites

- Three methods were discussed for combining information across sites: a simple average, a weighted average, and a '% NotGood' approach.
- It was agreed that for TraC waters, a '% NotGood' approach was appropriate. This would allow the WB as a whole to be classified as Good provided not more than some specified small percentage of the WB (by area) was worse than Good. The choice of critical percentage (10%, say), is still to be decided.
- For heterogeneous WBs, a weighted average method may be appropriate, with the weights reflecting the known or assumed proportions of the WB in the Good and NotGood categories. However, practical difficulties would often arise in determining appropriate weights that were acceptable to all parties.
- For extreme cases of heterogeneity, the soundest option would be to revise the WB delineation where this was feasible.
- Where the WB was considered spatially homogeneous, it was agreed that it would be appropriate to use the simple average approach.

Combining CofC information for different QEs at a site

- Two principal approaches were discussed: the z-scores method for 'pooling evidence' from a number of QEs all believed to be reflecting a similar pressure; and the one-out, all out ('1oAo') method for determining the 'worst-case outcome' from a collection of QEs.
- It was agreed that, in practice, QEs would virtually always be reflecting different pressures, at least to some extent, and so the correct method to use would be the precautionary one-out, all out approach.

- It was also agreed that monitoring data should not be used retrospectively to suggest the method of analysis. The rules to be used for combining CofC data should be agreed beforehand and adhered to.

Compensating for multiple assessments of a WB

- The greater the number of QEs monitored, the greater is the risk of a false positive - whereby a truly satisfactory WB is judged to have failed. A statistical method ('Bon Ferroni') is available that can compensate for this, but a decision needs to be taken about whether such a step is necessary or desirable. This issue is currently being debated by email.

Appendix A Details of the worked examples

A1 Introduction

Table 7 in the main text introduces 13 worked examples - seven for Rivers, and three each for Lakes and Transitional & Coastal waters. We have applied a three-stage 'sequential building block' approach to arrive at the final 'Confidence of Good or High' value for the Water Body. These are set out in Section A2 below, and illustrated in the final three columns of the table. Section A3 then provides a detailed commentary for two of the most complicated examples - 7 and 13.

Please note that the input data used in the examples is artificial, and designed solely to illustrate various common scenarios. However, the examples are available in Excel format, upon request, and so anyone interested in substituting their own 'real' or 'what-if?' data can readily do so.

A2 Principles adopted

- Where there is one site with several QEs reflecting the same Pressure, the CofGoHs are pooled by the z-scores method. (This principle may be over-ruled, on the basis of scientific judgement, in favour of 1oAo. However, such decisions should be declared a priori, rather than in an attempt to coax more information out of inconclusive data - in contravention of Cake's Law.)

- Where there is one site with several QEs reflecting different Pressures, the 1o,Ao method is used to determine CofGoH.

- Where there are two or more sites with different Pressures (whether or not the QEs are the same), the 1oAo method is used to determine CofGoH.

- Where there are two randomly selected sites with the same QE and Pressure, the EQRs for the two sites are averaged, and the associated CofGoH calculated.

- Where there are many (say six or more) randomly selected sites with the same QE and Pressure, the EQRs for the two sites are combined by the Spat%ile method, and the associated CofGoH calculated.

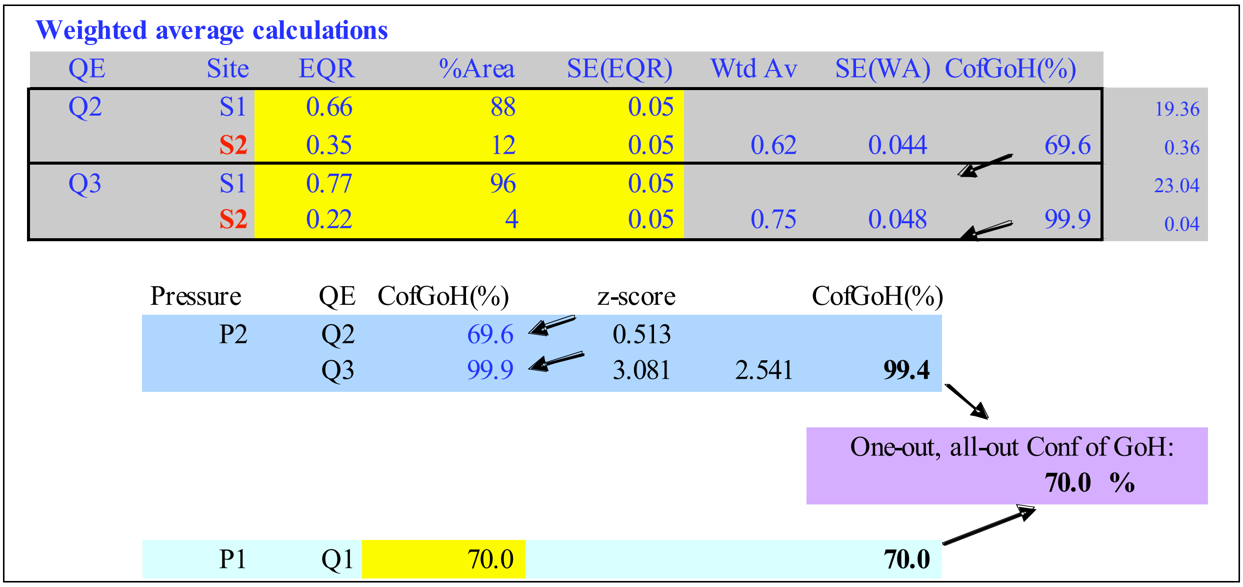
- Where there are 3 - 5 randomly selected sites with the same QE and Pressure, a judgement needs to be made as to whether a mean EQR or an X%ile EQR approach is appropriate. Then use one of the preceding two methods.

- Where there is a mixture of sites, including both 'randomly selected and targeted' sites, with the same QE and Pressure, a weighted average of the EQRs for the sites is calculated. The weights are the percentages of the WB that are believed to be represented by the conditions at the targeted sites. The associated CofGoH is then calculated.

A3 Detailed description of two examples

Note that in both cases, the yellow cells represent user inputs.

A3.1 Example 7



Weighted average calculations

QE	Site	EQR	%Area	SE(EQR)	Wtd Av	SE(WA)	CofGoH(%)	
Q2	S1	0.66	88	0.05				19.36
	S2	0.35	12	0.05	0.62	0.044	69.6	0.36
Q3	S1	0.77	96	0.05				23.04
	S2	0.22	4	0.05	0.75	0.048	99.9	0.04

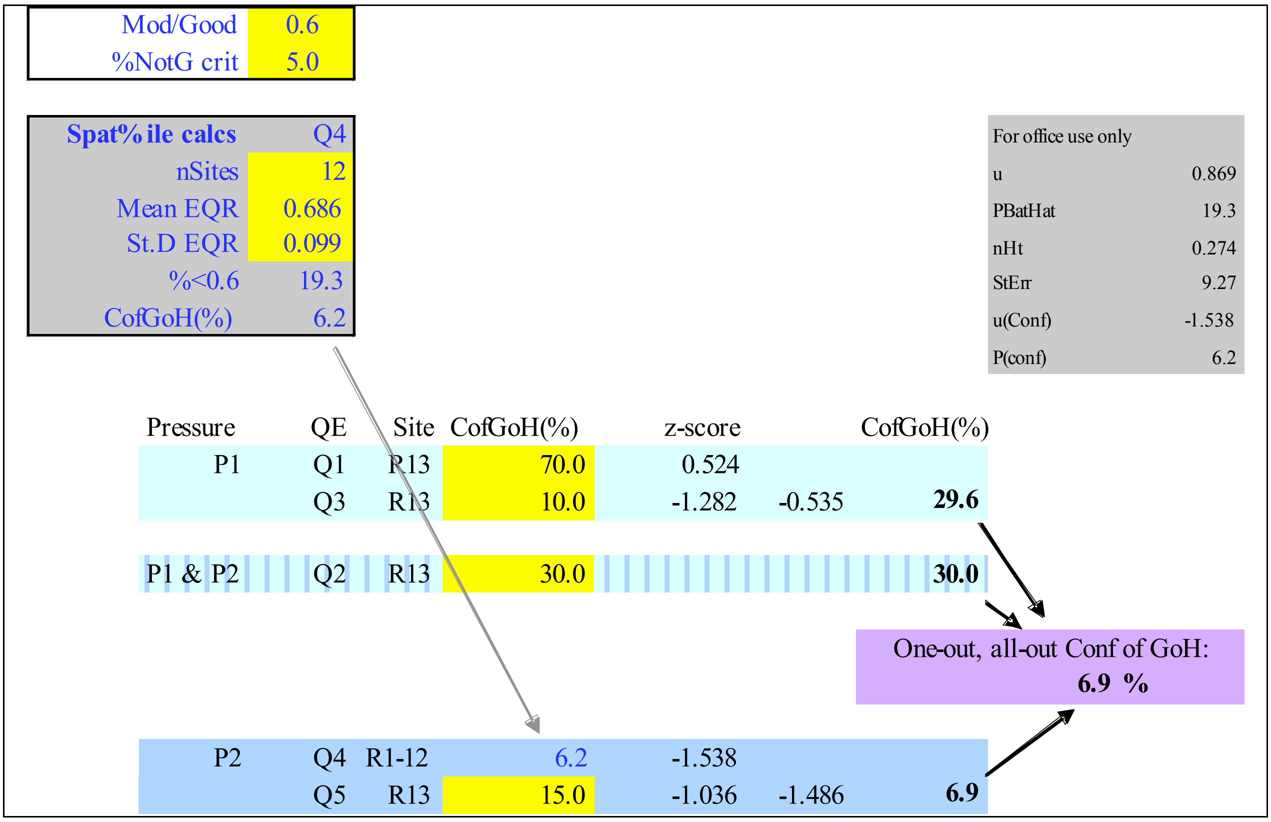
Pressure	QE	CofGoH(%)	z-score		CofGoH(%)
P2	Q2	69.6	0.513		
	Q3	99.9	3.081	2.541	**99.4**

One-out, all-out Conf of GoH:
70.0 %

P1	Q1	70.0			**70.0**

- The first grey panel shows the weighted average calculation for the Q2 results at sites S1 and S2. Overall quality is estimated by the weighted average EQR of 0.62, and there is 69.6% confidence that this is greater than the Mod/Good boundary of 0.6.
- Similar weighted average calculations are carried out in the second grey panel for the Q3 results at sites S1 and S2. For this QE we have 99.9% confidence that the WB is Good or High.
- These two CofGoH values both relate to the same Pressure (P2), and so these are transferred to the next panel and combined by the z-scores method to give an overall CofGoH of 99.4%.
- There is only one site and QE for the other Pressure (P1), and so the CofGoH is obtained directly from the input data as 70%.
- Finally, the poorer of the CofGoH values for the two Pressures is selected under the 1oAo rule to produce the overall CofGoH figure for the WB, viz **70%.**
- Thus we are 70% confident that the WB is Good or High - or, equivalently, 30% confident that it is worse than Good.
- As we cannot declare with 95% confidence that the WB is failing the Mod/Good boundary, **we may classify the WB as Good or High**.

A3.2 Example 13



Mod/Good	0.6
%NotG crit	5.0

Spat% ile calcs	Q4
nSites	12
Mean EQR	0.686
St.D EQR	0.099
%<0.6	19.3
CofGoH(%)	6.2

For office use only	
u	0.869
PBatHat	19.3
nHt	0.274
StErr	9.27
u(Conf)	-1.538
P(conf)	6.2

Pressure	QE	Site	CofGoH(%)	z-score		CofGoH(%)
P1	Q1	R13	70.0	0.524		
	Q3	R13	10.0	-1.282	-0.535	**29.6**
P1 & P2	Q2	R13	30.0			**30.0**
P2	Q4	R1-12	6.2	-1.538		
	Q5	R13	15.0	-1.036	-1.486	**6.9**

One-out, all-out Conf of GoH:
6.9 %

- The CofGoH values for the three QEs reflecting Pressure P1 are combined by the z-scores method, and this produces an overall CofGoH of 23%.
- Next, the detailed Spat%ile calculations for Q4 are carried out in the grey panel on the left. The resulting CofGoH value for Q4 (17.9%) feeds through into the panel for Pressure P2.
- The CofGoH values for the three QEs reflecting Pressure P2 are combined by the z-scores method to produce an overall CofGoH of 5.8%.
- Finally, the poorer of the CofGoH values for the two Pressures is selected under the 1o,Ao rule to produce the overall CofGoH figure for the WB, viz **5.8%.** Equivalently, we are 94.2% confident that the WB is worse than Good.
- Thus, as we are not quite able to declare with 95% confidence that the WB is failing the Mod/Good boundary, **we may classify the WB as Good or High**.

References & Bibliography

Environment Agency (2006). Uncertainty estimation for monitoring results by the WFD biological classification tools. Science Report.

Environment Agency (2007). Uncertainty estimation for monitoring results by the WFD biological classification tools - Workshop to obtain common view and application of deriving uncertainty estimates with classification results. WFD Report, in press.

UK TAG (2005). Classification Schemes in River Basin Planning: An Overview.

Glossary of terms

Ecological status is an expression of the quality of the structure and functioning of aquatic ecosystems associated with surface waters, classified in accordance with Annex V of the Water Framework Directive (2000/60/EC)

Good ecological status is the status of a body of surface water, so classified in accordance with Annex V of the Water Framework Directive (2000/60/EC).

A Water Body is a Body of surface water. means a discrete and significant element of surface water such as a lake, a reservoir, a stream, river or canal, part of a stream, river or canal, a transitional water or a stretch of coastal water (2000/60/EC).

Surface water status is the general expression of the status of a body of surface water, determined by the poorer of its ecological status and its chemical status (2000/60/EC).

List of abbreviations

1oAo	One-out, all-out - describing a method for selecting the worst-case outcome from a number of different QEs
Ave	Average
B-of-D	Benefit of Doubt (one of the possible stances towards the burden of proof in assessing compliance)
Chla	Chlorophyll a
CofC	Confidence of Class
CofGoH	Confidence of Good or High
Conf(NotGood)	Confidence of Moderate or Worse
CPET	Chironomid Pupal Exuvial Technique
EA	Environment Agency
EFI	European Fish Index
EQR	Ecological Quality Ratio
NotGood	Term describing a site that is not Good or High - that is, it is Moderate or worse.
QE	Quality Element
SEPA	Scottish Environment Protection Agency
Spat%ile	Spatial Percentile
TraC	Transitional and Coastal
UK Tag	United Kingdom Technical Advisory Group
WB	Water Body
WFD	Water Framework Directive (2000/60/EC)
Wtd Ave	Weighted Average
z-scores	A method of pooling information from a collection of statistical significance tests on measures all purporting to reflect the same underlying effect.

We welcome views from our users, stakeholders and the public, including comments about the content and presentation of this report. If you are happy with our service, please tell us about it. It helps us to identify good practice and rewards our staff. If you are unhappy with our service, please let us know how we can improve it.

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

Environment first: This publication is printed on paper made from 100 per cent previously used waste. By-products from making the pulp and paper are used for composting and fertiliser, for making cement and for generating energy.