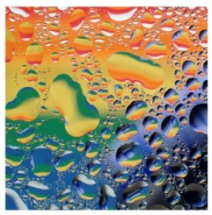
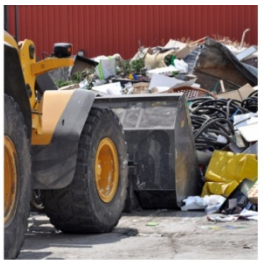



	Client:	Environment Agency	Report Ref: 14713-0/RH
Report Date:	November 2011	Contract No:	14713-0
Author:	Andrew Davey, Victoria Bewes		

Statistical review of the one-out, all out rule for classifying water bodies

1. Introduction

Whenever multiple quality elements (QEs) are monitored for the EU Water Framework Directive (WFD), the ecological status of a waterbody is determined by the worst-performing QE (the so-called 'one-out-all-out' (1OAO) rule). The 1OAO rule means that the risk of accidentally down-grading a water body increases with the number of QEs measured, which introduces a pessimistic bias to the classification results. The Agency wishes to assess the potential severity of this bias when reporting at a national scale and to understand the options for correcting classification results.

This short technical note explains the nature of the problem and discusses the pros and cons of possible solutions.

2. Background

2.1 The classification process

The WFD requires the Agency to provide a broad-ranging assessment of the environmental health of surface waters (rivers, lakes, estuaries and coastal waters) by monitoring various environmental parameters (termed Quality Elements, QEs). These parameters are divided into biological QEs (e.g. fish, invertebrates, plants, algae), physico-chemical QEs (e.g. temperature, nutrients, salinity, oxygenation), and hydromorphological quality elements (e.g. hydrological regime, morphological condition). The results from all these QEs are combined to yield an overall assessment of ecological status.

The assessment is carried out for individual management units called water bodies; water bodies are sections of a river or coastline, or a whole lake or estuary. In most cases[1], the number of quality elements monitored in each water body is determined by a prior assessment of the local environmental problems and which QEs are at risk of failing their environmental objectives. This means that the number and combination of QEs monitored varies from water body to water body.

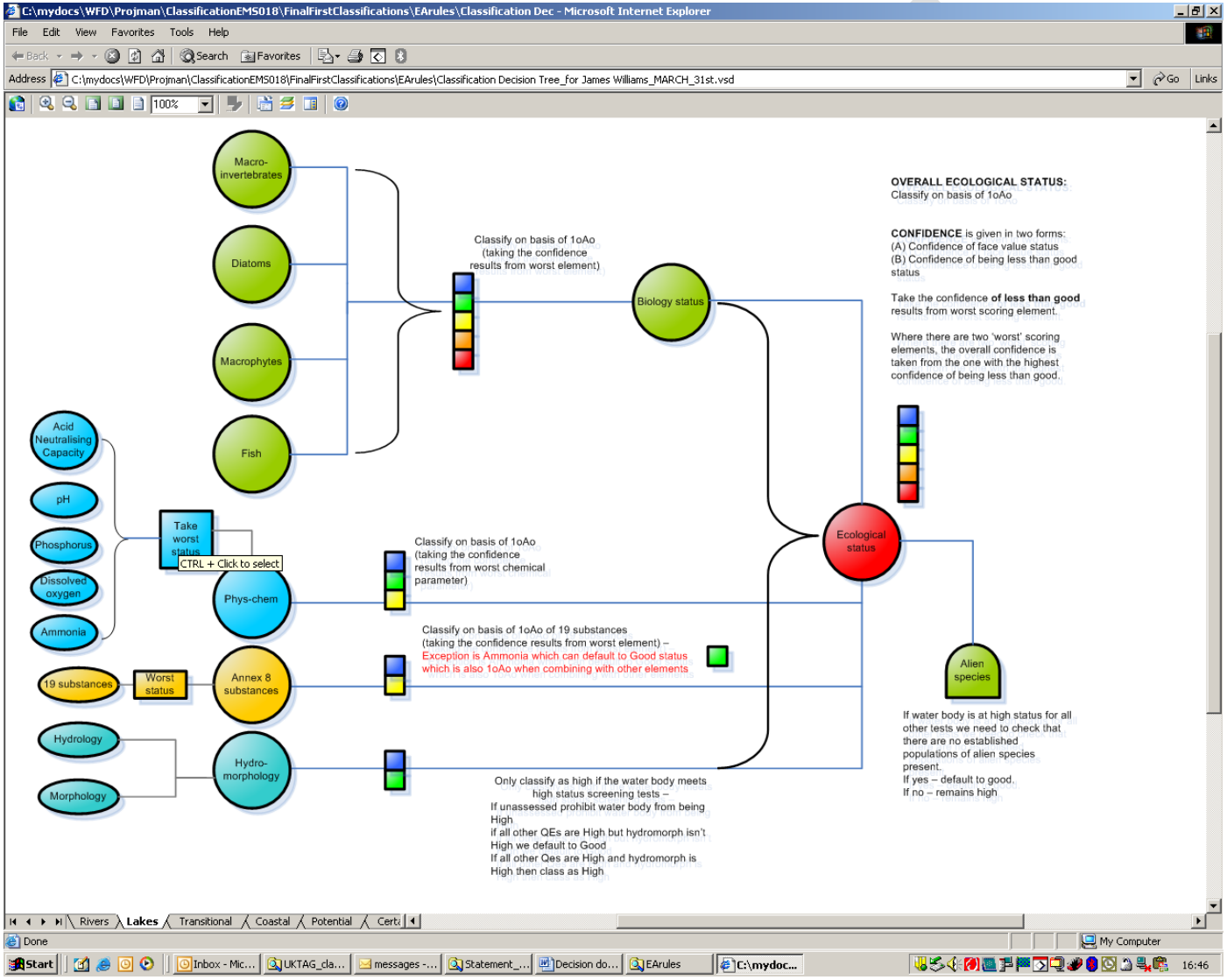
Each QE is used to classify the water body into one of five classes: High, Good, Moderate, Poor or Bad. The exact details of how the results from individual QEs are combined to yield a classification result differ slightly among different types of water body but in general the ecological status of the water body is determined by the worst-performing QE (the so-called 'one-out-all-out' (1OAO) principle).[2] Figure 1 illustrates this process for river water bodies.

Except where improvements would be disproportionately expensive or technically infeasible, the WFD requires all water bodies to achieve at least Good Ecological Status (GES) by 2027 at the latest. The proportion of water bodies at GESis a key indicator of the health of the aquatic environment and is reported annually by the Agency.

[1] A small proportion of water bodies are part of the surveillance monitoring network and are monitored for all quality elements. The remainder are subject to pressure-based operational monitoring where only those QEs thought to be as risk of failing are monitored.

[2] Annex V, 1.4.2(i) of the WFD requires: "For surface water categories, the ecological status classification for the body of water shall be represented by the lower of the values for the biological and physico-chemical monitoring results for the relevant quality elements".

Figure 1 Flow diagram showing how quality elements combine to classify the ecological status of river water bodies



2.2 Mis-classification and its consequences

In an ideal world, the status of each QE would be known without error and each water body would be correctly assigned as meeting or not meeting GES.

In reality, measurements of biological, physico-chemical and hydromorphological parameters are subject to various sources of error and the observed data may give a falsely optimistic or pessimistic reflection of the true environmental conditions. For example, the true (but unknown) mean phosphorus concentration in a river over a three year reporting period might be 85 µg/l, but the 36 monthly samples collected by the Agency might estimate the mean concentration at 94 µg/l because some samples just happened to be taken on days with unusually high phosphorus concentrations. The disparity between the observed and true values is known as sampling error.

In many cases, sampling error will not affect the reliability of the classification result – i.e. the QE will be placed in the same status class as it would have been had the Agency had a complete and prefect record of phosphorus concentrations minute by minute throughout the three year period. But of course there is a risk that sampling error will result in mis-classification – i.e. based on the observed data, the QE may be placed in a higher or lower class than it should be.

Mis-classification of individual QEs can cause a water body to be falsely classified as Good or better (a false positive), or as Moderate or worse (a false negative). These two types of classification error are illustrated in Table 1.

Table 1 Types of classification error for individual QEs

		Judgement based on monitoring data	
		Good or better	**Moderate or worse**
Reality	**Good or better**	Correct result	False negative
	Moderate or worse	False positive	Correct result

3. How likely are false negatives and false positives?

3.1 False negatives

As we've seen above, mis-classification of individual quality elements can lead to false positives and false negatives when determining whether a water body is at GES.

A false negative can only occur if the *true* ecological status of the water body is Good or better (i.e. all monitored QEs are of at least Good status; see Table 1). Now suppose, as an example, that a water body is assessed for two QEs, which are both truly at Good status. Also suppose that each QE has a 90% chance of being classed as Good, and a 10% chance of being falsely classed as Moderate. The probability that at least one of the QEs is classed as Moderate, and therefore that the water body as a whole is falsely classed as Moderate (i.e. a false negative), is 19% (Table 2).

Table 2 Ecological status of a water body based on observed status of two quality elements truly at Good status

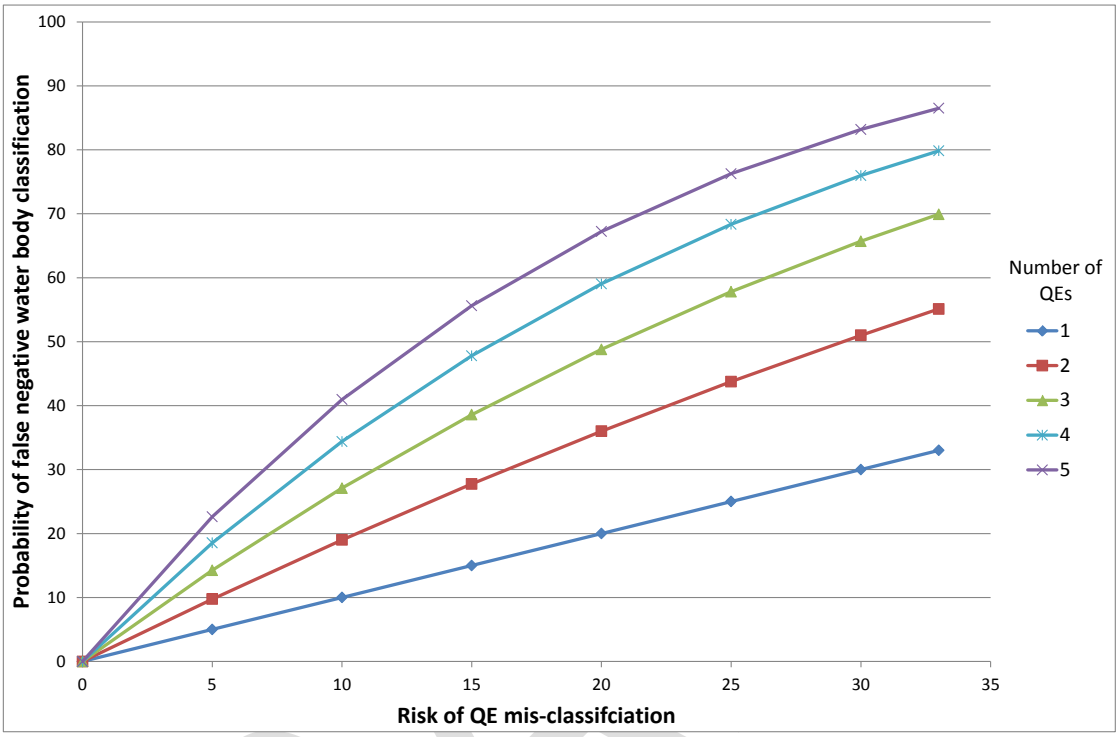
Observed status of first QE	Observed status of second QE	
	Good (90%)	**Moderate (10%)**
Good (90%)	Good (81%)	Moderate (9%)
Moderate (10%)	Moderate (9%)	Moderate (1%)

The probability of a false negative (i.e. falsely classifying a water body as Moderate or worse) increases with:

1. the number of QEs monitored - because the more QEs that are monitored, the greater the chance that at least one will be falsely classed as Moderate, or worse; and,
2. the risk of each individual QE being mis-classified.

Figure 2 shows how the probability of a false negative increases as the number of QEs increases from 1 to 5 and as the risk of each QE being falsely classified as Moderate or worse increases form 0% to 33%.

Figure 2 Probability of under-estimating the ecological status of a water body as a function of the number of quality elements monitored and the confidence of each quality element classification results



3.2 False positives

Now consider the opposite side of the coin. A false positive can only occur if the *true* ecological status of the water body is Moderate or worse (i.e. at least one monitored QE is of Moderate, Poor or Bad status; see Table 1). Now suppose, as an example, that a water body is assessed for two QEs, which are both truly at Moderate status. Also suppose that each QE has a 90% chance of being classed as Moderate, and a 10% chance of being falsely classed as Good. The probability that both QEs are classed as Good, and therefore that the water body as a whole is falsely classed as Good (i.e. a false positive), is only 1% (Table 2). Contrast this with the 19% chance of a false negative in the previous example.

Table 3 Ecological status of a water body based on observed status of two quality elements truly at Moderate status

Observed status of first QE	Observed status of second QE	
	Good (10%)	**Moderate (90%)**
Good (10%)	Good (1%)	Moderate (9%)
Moderate (90%)	Moderate (9%)	Moderate (81%)

Thus, the 1OAO rule false negatives much more likely to occur than false positives, all else being equal.

3.3 Techniques for controlling false negatives

As we have seen, the more QEs that are monitored, the greater the risk that a water body will be falsely classified as Moderate or worse (a false negative). Similarly, the more bites that we have at the cherry, the greater the risk that the water body will fail the GES test with high confidence, and the greater the risk of implementing expensive and unnecessary mitigation measures. In statistical terms this is known as the 'multiple comparisons' problem.

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 'Bonferroni' method (Ellis 2007).[3] 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 Bonferroni 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 = 5 quality elements 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 water body as a whole has failed (Ellis 2007).

As an example, consider the three QEs in Table 4, and suppose that a water body fails the GES test if there is at least C=95% confidence that the water body is worse than Good.

[3] Ellis, J.C. (2007) Combining Multiple Quality Elements and Defining Spatial Rules for WFD Classification. WRc report to SNIFFER/EA. Although termed the Bonferroni procedure by Ellis 92007), it is actually the Dunn-Sidak modification of the standard Bonferroni procedure.

Table 4 Example of confidence that class is Good or better and face value class for three QEs in a waterbody

QE	Confidence of Good or better	Face value class
1	60 %	Good
2	12 %	Moderate
3	4 %	Bad

Under the 1OAO rule, the ecological status of the waterbody would be classed as worse than Good with 96% confidence, to match the the worst performing QE, and would therefore fail the GES test.

Using the Bonferroni procedure, however, each QE is tested at a confidence level of:

$$C_{BF} = 100(C/100)^{(1/3)} = 98.3\%.$$

The water body would still be classed as Bad status, but none of the QEs fail at the C_{BF} level so we cannot say with 95% confidence that the waterbody is worse than Good status.

The Bonferroni procedure can also be applied to face value classification results. In this situation, the water body fails the GES test if there is at least C=50% confidence that the water body is worse than Good. Table 5 shows the Bonferroni adjusted confidence levels for between 1 and 10 QEs. Applying the Bonferroni procedure, a water body monitored for 8 QEs would be judged to fail the GES test only if the worst performing QE had at least 91.7% confidence of being worse than Good.

Table 5 Bonferroni adjusted confidence levels for between 1 and 10 quality elements

Number of QEs	Bonferroni adjusted confidence level
1	50.0%
2	70.7%
3	79.4%
4	84.1%
5	87.1%
6	89.1%
7	90.6%

Number of QEs	Bonferroni adjusted confidence level
8	91.7%
9	92.6%
10	93.3%

Although the Bonferroni procedure does guard against the risk of false negatives, it has some drawbacks:

1. It is very conservative, particularly with large numbers of QEs – i.e. the power to detect genuinely failing water bodies is diminished.

2. The Bonferroni test can also be manipulated by entering large numbers of good quality QEs, which reduce the calculated confidence level so that the test becomes more difficult to fail.

3. The Bonferroni test makes the assumption that all QEs are independent, whereas in reality QEs may respond to similar pressures and therefore be correlated to some extent.

4. How many false negatives and false positives are there at a national level?

The fact that a false negative can arise more easily than a false positive when applying the 1OAO rule suggests that the proportion of water bodies achieving GES could be under-estimated at a national scale – i.e. that national WFD classification results might be subject to a pessimistic bias. The 2010 WFD classification results were examined to try to ascertain whether this might be the case.

The ecological status of each water body is classed as 'Good or better' or 'Moderate or worse' based on the available monitoring data. There is no way of knowing whether an observed result is a false negative or a false positive because we do not know the true status of the water body (if we did then there would be no need to undertake monitoring!). It is possible however, to calculate the probability of each classification result being wrong.

4.1 False negatives

For water bodies classed as Moderate or worse, the risk of a false negative is estimated by calculating from the data the probability that the water body could actually be Good or High.

Taking the example in Table 6, two QEs are classed Good and one is classed Moderate. The worst performing QE has 20% confidence of being Good or better, BUT the risk of a false negative is not 20% because the two QEs each have 20% confidence of being Moderate or worse. For the water body as a whole to be at GES, all three QEs would have to be Good or better; the probability of this is 0.8 x 0.8 x 0.2 = 0.132. the probability of a false negative is therefore 13%

Table 6 Illustration of a false negative

QE	Face value class	Confidence of Good or better	Confidence of Moderate or worse
QE1	Good	0.8	0.2
QE2	Good	0.8	0.2
QE3	Moderate	0.2	0.8

This logic was applied to the 2010 WFD classification results. 4417 water bodies were classified as Moderate, Poor or Bad. [4] Exact confidence figures were not available, so the following approximations were used:

- QEs at High status = 95% confident of Good or better;
- QEs at Good status = 75% confident of Good or better;
- QEs at worse than Good status, but uncertain = 33% confident of Good or better;
- QEs at worse than Good status, and quite certain = 25% confident of Good or better;
- QEs at worse than Good status, and very certain = 5% confident of Good or better;
- Water body at GES based on expert judgment or mitigation measures assessment = 75% confident of Good or better; and,

[4] 74 out of a total of 5818 were not assessed.

- Water body not at GES based on expert judgment or mitigation measures assessment = 25% confident of Good or better.

Based on these assumptions, 262 water bodies were estimated to be falsely classed as Moderate or Worse. The vast majority of these (210) were those classified using expert judgment or mitigation measures assessment.

4.2 False positives

For water bodies classed as Good or better, the risk of a false positive is estimated by calculating from the data the probability that the water body could actually be Moderate, Poor or Bad.

Taking the example in Table 7, all three QEs are classed Good. The worst performing QE has 30% confidence of being Moderate or worse, BUT the risk of a false positive is not 30% because the other two QEs might also be Moderate or worse. For the water body as a whole to be at GES, all three QEs would have to be Good or better; the probability of this is 0.9 x 0.8 x 0.7 = 0.51. The probability of a false positive is therefore 49%.

Table 7 Illustration of a false positive

QE	Face value class	Confidence of Good or better	Confidence of Moderate or worse
QE1	Good	0.9	0.1
QE2	Good	0.8	0.2
QE3	Good	0.7	0.3

Applying this logic to the 1330 water bodies classified as Good or High in 2010, and making the same assumptions about the level of confidence, 504 water bodies were estimated to be falsely classed as Good or better. Only 47 of these were classified using expert judgment or mitigation measures assessment. The majority (419) were water bodies with six or more QEs where each individual QE had a low confidence of being moderate or worse, but the combined confidence was quite high (as in Table 7).

5. Discussion and Conclusions

This analysis of the 1OAO rule has revealed an apparent contradiction between theory and practice. In theory, false negatives are expected to be more prevalent than false positives because a water body can be falsely classified as Moderate or worse by the mis-classification of a just a single QE. This has prompted concern WFD results might under-estimate the proportion of water bodies at Good ecological status.

In practice, however, false positives appear to be more prevalent that false negatives, suggesting that national WFD results may in fact over-estimate the proportion of water bodies at Good ecological status. Why is this?

In section 3, we took the true water body status as the starting point, and calculated the probabilities of a correct result, a false positive or a false negative. Three of the four possible observed outcomes give a false positive (Table 8), while only one of four gives a false positive (Table 9). It would follow from this that a false positive tends to be less likely than a false negative.

Table 8 True status known to be Good

QE	True status	Observed status			
QE1	Good	Good	Good	Moderate	Moderate
QE2	Good	Good	Moderate	Good	Moderate
Water body	Good	Correct result	False negative	False negative	False negative

Table 9 True status known to be Moderate

QE	True status	Observed status			
QE1	Moderate	Good	Good	Moderate	Moderate
QE2	Moderate	Good	Moderate	Good	Moderate
Water body	Moderate	False positive	Correct result	Correct result	Correct result

In fact this analysis is unrealistic, because the true status is unknown, and in section 4 we start with the observed status. If it good, a false positive arises in 3 cases, with either or both QEs being misclassified (Table 10). A false negative arises in only 1 case out of 4 (Table 11). It follows that a false positive is not unlikely.

Table 10 Observed status is Good

QE	Observed status	True status			
QE1	Good	Good	Good	Moderate	Moderate
QE2	Good	Good	Moderate	Good	Moderate
Water body	Good	Correct result	False positive	False positive	False positive

Table 11 Observed status is Moderate

QE	Observed status	True status			
QE1	Moderate	Good	Good	Moderate	Moderate
QE2	Moderate	Good	Moderate	Good	Moderate
Water body	Moderate	False negative	Correct result	Correct result	Correct result

In summary, potential pessimistic bias in the national WFD classification results caused by the 1OAO rule is cancelled out by an even greater optimistic bias that arises from not taking into account the possible mis-classification of QEs reported as being of Good and High status. Application of the Bonferroni procedure is therefore not recommended as this would give an unnecessarily generous level of protection against false negatives and severely reduce the Agency's ability to detect water bodies that are genuinely failing to meet Good ecological status.