An approach to the intercalibration of benthic ecological status assessment in the North Atlantic ecoregion, according to the European Water Framework Directive

Angel Borja a,*, Alf B. Josefson b, Alison Miles c, Iñigo Muxika a, Frode Olsgard d, Graham Phillips c, J. Germán Rodríguez a, Brage Rygg d

a AZTI-Tecnalia, Marine Research Division, Herrera Kai Portualdea, 20110 Pasaia, Spain
b Department of Marine Ecology, National Environmental Research Institute (NERI), 4000 Roskilde, Denmark
c Environment Agency (EA), Kingfisher House, Goldhay Way, Orton Goldhay, Peterborough PE2 5ZR, UK
d Norwegian Institute for Water Research (NIVA), Brekkeveien, 19, P. O. Box 173, Kjelsas, N-0411 Oslo, Norway

Abstract

The European Water Framework Directive (WFD) establishes a framework for the protection and improvement of transitional and coastal waters; its final objective is to achieve at least ‘good water status’ for all waters, by 2015. The WFD requires Member States (MSs) to assess the Ecological Status (ES) of water bodies. This assessment will be based upon the status of the biological, hydromorphological and physico-chemical quality elements, by comparing data obtained from monitoring networks to reference (undisturbed) conditions, and then deriving an Ecological Quality Ratio (EQR). One of the biological quality elements to be considered is the benthic invertebrate component and some structural parameters (composition, diversity and disturbance-sensitive taxa) must be included in the ES assessment. Following these criteria, several approaches to benthic invertebrate assessment have been proposed by MSs. The WFD requires that these approaches are intercalibrated.

This contribution describes the comparison of the different methodologies proposed by United Kingdom, Spain, Denmark and Norway. Results show a high consistency between the approaches, both with regard to determining the EQR and boundary settings for the ES.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: Intercalibration; Water Framework Directive; Benthic invertebrates; North-East Atlantic; Soft-bottom communities

1. Introduction

The European Water Framework Directive (WFD; 2000/60/EC) establishes a framework for the protection and improvement of all European surface and ground waters (including transitional and coastal waters); its final objective is to achieve at least ‘good water status’ for all waters bodies, by 2015. The WFD requires Member States (MSs) to assess the Ecological Status (ES) of water bodies. Status will be assigned through the assessment of biological, hydromorphological and physico-chemical quality elements, by comparing data obtained from monitoring networks to reference (undisturbed) conditions, thereby deriving an Ecological Quality Ratio (EQR). This ratio shall be expressed as a numerical value between zero and one, with ‘high’ status represented by values close to one and ‘bad’ status by values close to zero. In coastal and transitional waters, one of the biological quality elements to be considered is the benthic invertebrate fauna, an important component of which is the soft-bottom benthos.

The WFD defines the aspects of the biological quality elements that must be included in the ES assessment of a water body (annex V, WFD). Any proposed WFD classification scheme must, therefore, include methodologies that...
address those parameters defined for assessing the benthic quality status: ‘the level of diversity and abundance of invertebrate taxa’ and the proportion of ‘disturbance–sensitive taxa’. Following these criteria, to date several methodologies have been proposed by MSs for the status assessment of the benthic component (Borja et al., 2000, 2003, 2004a,b; Prior et al., 2004; Rosenberg et al., 2004).

All of these methodologies have focused upon the proportion of disturbance–sensitive taxa, with the AZTI Marine Biotic Index (AMBI) (Borja et al., 2000) being one of the most widely used in European countries.

Prior to the implementation of WFD assessment, any proposed methodology must be intercalibrated between the MSs within an ecoregion. Each MS shall divide the EQR scale for their monitoring system into five ecological status classes (high, good, moderate, poor, and bad) by assigning a numerical value to each of the class boundaries. The value for the ‘high/good’ and the ‘good/moderate’ class boundaries should be established through the intercalibration exercise. This is to ensure that the established class boundaries are consistent with the normative definitions of the WFD and are comparable between MSs.

As part of the official intercalibration exercise, a range of sites in surface water bodies in each ecoregion in the European Union have been identified, in order to establish an intercalibration network of sites (WFD Intercalibration Register). For each MS, for each surface water body type selected, the intercalibration network should consist of at least two sites corresponding to the boundaries between ‘high’ and ‘good’ status, and between ‘good’ and ‘moderate’ status, with status defined through the normative definitions. The sites were selected by expert judgement based on joint inspections of all available data on the quality elements and any other supporting information.

Within an ecoregion the monitoring system of each MS should be applied to those intercalibration network sites which lie in the surface water body type to which the system will be applied. The results of this application should then be used to set the numerical values for the relevant class boundaries in each MS monitoring system. The intercalibration exercise started in 2005 and is due to complete in June 2006.

This contribution describes a comparison of the initial methodologies proposed by several MSs in the Atlantic ecoregions, to derive the EQR and establish the benthic invertebrate fauna ES. The approach taken may present a useful way forward for other MSs, both for the North Atlantic and other ecoregions, and help develop the way the ES is assessed.

2. Methods

2.1. Data matrix

Although there exists a European intercalibration register, it has been recognised by both the European Commission and MSs that additional data from non-intercalibration sites, may be required to progress the intercalibration exercise. As such the benthic intercalibration group collated benthic invertebrate data from a range of geographical locations within the ecoregion, including intercalibration sites, and incorporating samples from distinct pressure gradients (ranging from ‘bad’ to ‘high’ ES) for (pressures and impacts, within the WFD, see Borja et al., 2006).

For this initial exercise, benthic invertebrate abundance samples were collated from the coastal water common European water body types NEA 1 and 26. These types are characterised by an exposed poly- to eutrophic (according to the Venice scheme) subtidal habitat, with soft sediment (muds, sandy muds, muddy sands). The samples were standardised for sample type (0.1 m², obtained in some cases by pooling data, see Table 1), sieve mesh size (1 mm), and sediment type (muds, muddy sands and sandy muds). By standardising data for this initial comparison, the high level of natural variability found in biological communities from different habitats was minimised, allowing changes in the benthic invertebrate communities to be more clearly associated with anthropogenic pressure.

The resulting data set was comprised of 589 benthic invertebrate abundance samples from different locations along the European Atlantic coasts (Fig. 1): Belgium (132), Denmark (72), Germany (64), Republic of Ireland (RoI, 14), Norway (12), Spain (45), and United Kingdom (UK, 250). As well as the benthic invertebrate abundance data, MSs submitted supporting parameters, such as water depth (m) and sediment size (% <63 μm) (Table 1). The pressure gradient data used relate to a variety of man-induced pressures and impact sources, such as eutrophication and hypoxia (Denmark); hypoxia (Stavanger, Norway); urban and industrial discharges from a submarine outfall (Basque Country, Spain) and sewage sludge disposal at sea (Garroch Head, UK). Other data were selected from well-known undisturbed areas, such as Trondheimsfjord in Norway.

Most of the Belgian data were taken from the west coast of Belgium, within the federal science project ‘HABITAT’ (Van Hoey et al., 2004). The other samples were taken from the eastern and middle part of the coastline (for details, see Fig. 1 and Table 1).

The Danish data were collected within the National Environmental Monitoring Programme of Denmark (Fig. 1) in order to evaluate the effects of eutrophication on Danish coastal areas. The monitoring has previously been described in annual reports (e.g. Årtebjerg et al., 2004) and international publications (e.g. biomass–nutrient relationships in Josefson and Rasmussen (2000); programme description and nutrient loads in Conley et al. (2002); species diversity in Josefson and Hansen (2004); and oxygen deficiency in Conley et al. (in press)). Some descriptive data and parameters can be seen in Table 1.

The Norwegian data were taken from an undisturbed site at 50 m depth in Trondheimsfjord (RAH1), from a sandy site at Utnes in southern Norway (U10), and from
Table 1

Sample description of data submitted by Member States, from the North Atlantic ecoregion, for the intercalibration exercise

<table>
<thead>
<tr>
<th>Member State</th>
<th>Site name</th>
<th>Sample method</th>
<th>Sample size (m²)</th>
<th>Stations (number)</th>
<th>Period</th>
<th>Replicates per station (number)</th>
<th>Samples submitted (number)</th>
<th>Depth (m)</th>
<th>Sediment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belgium</td>
<td>(1) Stations P2</td>
<td>VV</td>
<td>0.1026</td>
<td>1</td>
<td>1995–2000</td>
<td>1</td>
<td>1</td>
<td>6.7</td>
<td>Sand (97%)</td>
</tr>
<tr>
<td>Belgium</td>
<td>(2) Stations HA</td>
<td>VV</td>
<td>0.1026</td>
<td>12–37</td>
<td>1999–2000</td>
<td>49</td>
<td>9.5</td>
<td>Sand (85%)–</td>
<td></td>
</tr>
<tr>
<td>Belgium</td>
<td>(3) Stations MAREI44</td>
<td>VV</td>
<td>0.1026</td>
<td>1</td>
<td>2000</td>
<td>1</td>
<td>13.8</td>
<td>Mud (15%)</td>
<td></td>
</tr>
<tr>
<td>Belgium</td>
<td>(4) Stations M&amp;OD</td>
<td>VV</td>
<td>0.125</td>
<td>1–5</td>
<td>1996</td>
<td>6</td>
<td>14.2</td>
<td>Sand (99%)–</td>
<td></td>
</tr>
<tr>
<td>Belgium</td>
<td>(5) Stations O&amp;P</td>
<td>VV</td>
<td>0.125</td>
<td>1–8</td>
<td>1994, 1997</td>
<td>14</td>
<td>3.3</td>
<td>Sand (99%)–</td>
<td></td>
</tr>
<tr>
<td>Belgium</td>
<td>(6) Subtidal Stations</td>
<td>VV</td>
<td>0.1026</td>
<td>58</td>
<td>2002</td>
<td>58</td>
<td>3.7</td>
<td>Sand (93%)–</td>
<td></td>
</tr>
<tr>
<td>Denmark</td>
<td>(7) BF29 Øresund</td>
<td>HC</td>
<td>0.0143</td>
<td>1</td>
<td>2000–2003</td>
<td>25</td>
<td>27.9</td>
<td>Mud (50%)</td>
<td></td>
</tr>
<tr>
<td>Denmark</td>
<td>(8) BF23 Århus Bay</td>
<td>HC</td>
<td>0.0123</td>
<td>1</td>
<td>2000–2003</td>
<td>49</td>
<td>15.2</td>
<td>Mud (50%)</td>
<td></td>
</tr>
<tr>
<td>Denmark</td>
<td>(9) BF15 Lillebælt North Basin</td>
<td>HC</td>
<td>0.0123</td>
<td>1</td>
<td>2000–2003</td>
<td>45</td>
<td>180</td>
<td>19.4</td>
<td>Mud (50%)–</td>
</tr>
<tr>
<td>Denmark</td>
<td>(10) BF11 Rødgård</td>
<td>HC</td>
<td>0.129</td>
<td>1</td>
<td>2000–2003</td>
<td>35, 42, 43, 33</td>
<td>157</td>
<td>14.6</td>
<td>Mud (50%)–</td>
</tr>
<tr>
<td>Germany</td>
<td>(11) NS2 Vortrappetf</td>
<td>VV</td>
<td>0.1</td>
<td>1</td>
<td>1987–2004</td>
<td>3.5</td>
<td>54</td>
<td>15</td>
<td>Sand (94%)–</td>
</tr>
<tr>
<td>Norway</td>
<td>(12) Stavanger (SSA)</td>
<td>VV</td>
<td>0.1</td>
<td>1</td>
<td>1995</td>
<td>4</td>
<td>93</td>
<td>50</td>
<td>Mud (88%)</td>
</tr>
<tr>
<td>Norway</td>
<td>(13) Trondheimøfjord (RAH)</td>
<td>VV</td>
<td>0.1</td>
<td>1</td>
<td>2001</td>
<td>4</td>
<td>4</td>
<td>38</td>
<td>Sand (88%)</td>
</tr>
<tr>
<td>Norway</td>
<td>(14) Utne (U10)</td>
<td>VV</td>
<td>0.1</td>
<td>1</td>
<td>2001</td>
<td>4</td>
<td>38</td>
<td>38</td>
<td>Sand (88%)</td>
</tr>
<tr>
<td>R. of Ireland</td>
<td>(15) Kenmare River</td>
<td>DG</td>
<td>0.1</td>
<td>3</td>
<td>2003</td>
<td>3 × 4</td>
<td>12</td>
<td>45.9</td>
<td>Muddy sand</td>
</tr>
<tr>
<td>R. of Ireland</td>
<td>(16) Greetsianni Bay</td>
<td>DG&amp;BC</td>
<td>0.1 and 0.015</td>
<td>3</td>
<td>2003</td>
<td>3 × 4</td>
<td>12</td>
<td>40.1</td>
<td>Muddy sand</td>
</tr>
<tr>
<td>R. of Ireland</td>
<td>(17) Clew Bay</td>
<td>BC</td>
<td>0.015</td>
<td>3</td>
<td>2003</td>
<td>3 × 4</td>
<td>12</td>
<td>25</td>
<td>Fine sand</td>
</tr>
<tr>
<td>Spain</td>
<td>(18) San Sebastian– Passia</td>
<td>BC</td>
<td>0.166</td>
<td>9</td>
<td>2000–2004</td>
<td>3 (combined)</td>
<td>45</td>
<td>35.6-</td>
<td>Sand (90%)–</td>
</tr>
<tr>
<td>UK – Scotland</td>
<td>(19) Kilbrannan Sound (KIL)</td>
<td>DG</td>
<td>0.1</td>
<td>1</td>
<td>2004</td>
<td>10</td>
<td>50</td>
<td>50</td>
<td>Soft muds</td>
</tr>
<tr>
<td>UK –Scotland</td>
<td>(20) Liverpool Bay (LIV)</td>
<td>DG</td>
<td>0.1</td>
<td>3</td>
<td>2004</td>
<td>3 × 5</td>
<td>15</td>
<td>5.7</td>
<td>Sand (70%)–</td>
</tr>
<tr>
<td>UK –England</td>
<td>(21) Harwich (HAR)</td>
<td>DG</td>
<td>0.1</td>
<td>3</td>
<td>2004</td>
<td>3 × 5</td>
<td>15</td>
<td>6.4</td>
<td>Mud (85.3%)</td>
</tr>
<tr>
<td>UK –Wales</td>
<td>(22) Milford Haven (MIL)</td>
<td>DG</td>
<td>0.1</td>
<td>3</td>
<td>2004</td>
<td>3 × 5</td>
<td>15</td>
<td>4.6</td>
<td>Mud (78.5%)</td>
</tr>
<tr>
<td>UK –England</td>
<td>(23) Torbay (TOR)</td>
<td>DG</td>
<td>0.1</td>
<td>3</td>
<td>2004</td>
<td>3 × 5</td>
<td>15</td>
<td>13.7</td>
<td>Muddy sand</td>
</tr>
<tr>
<td>UK –Wales</td>
<td>(24) St. Bride Bay</td>
<td>DG</td>
<td>0.1</td>
<td>3</td>
<td>2004</td>
<td>3 × 5</td>
<td>15</td>
<td>37</td>
<td>Sand (88%)–</td>
</tr>
</tbody>
</table>

All samples were sieved with 1 mm mesh. Key: VV – van Veen grab; HC – Haps core; DG – Day grab; BC – Box core grab.
a fjord basin with organic load and some hypoxia, near Stavanger (S5A) (Fig. 1, Table 1).

The Spanish data were taken from the San Sebastián submarine outfall area (northern Spain) (Fig. 1, Table 1), which is located 1.2 km from the coast in an approx. 47 m water depth, and discharges $45.7 \times 10^6 \text{ m}^3 \text{ yr}^{-1}$ of urban and industrial waste water (Borja and Muxika, in press). Benthic community data from 5 months prior to the submarine outfall operation (year 2000), and four annual surveys after the commencement of discharges (from 2001 to 2004) have been used. The description of programme sampling can be seen in Borja and Muxika (in press) and some descriptive parameters can be seen in Table 1.

The UK data were taken within several intercalibration sites (LIV, Liverpool Bay; HAR, Harwich; MIL, Milford Haven; TOR, Torbay; SBB, St. Brides Bay; and KIL, Kilbrannan Sound) and a distinct sewage sludge disposal impact gradient (Garroch Head, data courtesy of Fisheries Research Services, Scotland) (Fig. 1, Table 1).

For details on the remainder samples from Germany and RoI, see Fig. 1 and Table 1.

It should be noted that assessments of the locations are based on the selected samples that were submitted for the intercalibration exercise and do not necessarily reflect the status of the whole water body. Moreover, in order to progress the intercalibration exercise, guidance produced by the WFD implementation working groups (Van de Bund, 2004) was used.

2.2. Pre-treatment of data

Although all samples were taken from similar soft sediment habitats, when combining benthic invertebrate data from different MSs, Agencies or laboratories it is generally necessary to standardize data before any accurate assessment can be achieved. This requirement arises due to differences in what is recorded in samples and in the level of taxonomic identification carried out between groups, factors which can affect classification assessment. Truncation of the data for this exercise was carried out to remove (i) non-benthic invertebrate taxa e.g. fish and algae; (ii) inconsistencies in the level of identification between laboratories; this in particular can affect classification tools as it can give a false impression of richness in a sample; and (iii) removal of non-soft sediment taxa; required due to inconsistencies in recording epibiota between laboratories. As for point (ii), taxa within the following taxonomic groups were combined to the level specified to negate the problem of inconsistent levels of identification: Oligochaeta; Nemertea; Platyhelminthes; Echiura; Sipuncula; Phoronida; Priapulida; several species (types a, b, etc.) to genus. Any of the taxa only identified to phylum level was also removed (with the exception of fauna from the above-mentioned Phyla). Other truncation rules, related with AMBI calculation, can be seen in Borja and Muxika (2005).

2.3. Methodologies used in assessing EQR

Once a standardised data matrix was established this was circulated to MSs in the ‘North East Atlantic Geographical Intercalibration Group’ (NEAGIG) for assessment by MSs’ current analysis methods. At the time of this comparison, only four main methodologies have been proposed by MSs for WFD assessment and could therefore be used in assessing EQR and the ecological status. These are from Denmark, Norway, Spain and UK (in this partic-
ular case, UK and RoI used the same methodology). Three of the methodologies use the AMBI in assessing the proportion of sensitive and indicator species (Borja et al., 2000). In the calculation of this index, the free software (http://www.azti.es) together with the guidelines from the authors (Borja and Muxika, 2005) has been used.

The Danish method is a multimoetric approach which takes into account the proportion of sensitive/tolerant species, measured by the AMBI; a diversity component, Shannon–Wiener index (Shannon and Weaver, 1963); and a factor to compensate for low densities and species numbers. All variables have equal weight in this approach, and the multimetric ranges from 0 to 1. The equation is

\[
DKI = \left(\frac{(1 - AMBI)}{7} + \frac{H}{H_{\text{max}}}\right) \times \frac{1}{2} \times \left(1 - \left(\frac{N}{1}ight)\right)
\]

where \(H\) the Shannon–Wiener index with log base 2, \(H_{\text{max}}\) the reference value that \(H\) can reach in undisturbed conditions, \(N\) the number of individuals and \(S\) the number of species.

The Norwegian method uses the ISI index based on the relative presence of pollution-sensitive species in the sample (equivalent to the AMBI index) and F3, a multimetric index including ISI, ES100 (a diversity index) and \(S\) (number of species), normalised to depth, grain size, and sampled area (for details on these methods, see Rygg, 1985, 2002; Solheim et al., 2004). The F3 was used in this study, changing the F3 scale to an EQR scale.

The Spanish method (used in the Basque Country) includes the use of AMBI, species richness and Shannon’s diversity as structural parameters. The EQR is calculated in Solheim et al. (2004) and Muxika et al. (2006), including the distance of a location to two virtual ‘high’ and ‘bad’ quality status locations (for details, on the methodology, see Borja et al., 2004a; Bald et al., 2005; Muxika et al., 2006). In this particular case, the EQR can be >1 when monitoring stations have values over those with virtual ‘high’ quality (see Borja et al., 2004a).

The UK method uses a multimetric index, combining AMBI (disturbance/sensitive taxa), Simpson’s diversity, number of individuals (abundance), and number of taxa

\[
\text{Index} = \left(2 \times \left(1 - \frac{\text{AMBI}}{7}\right) + \left(1 - \frac{\text{Lambda}'}{S}\right)\right) \times \frac{1}{2} \left[\left(1 - \frac{1}{S}\right) + \left(1 - \frac{1}{N}\right)\right]
\]

where \(\text{Lambda}'\) is Simpson’s Index, \(S\) is the number of taxa and \(N\) is the number of individuals (modifications of the UK index (version 2 was used for this comparison) are being trialled for applicability to a wider range of habitat types).

2.4. Reference conditions

Although the WFD requires the comparison of data against reference conditions, only the Spanish method determined those conditions. They are based on the approach described in Bald et al. (2005). High reference values for each of the structural parameters and communities were selected from Borja et al. (2004c); conversely, bad reference conditions were selected from azoic sediments, as described in Muxika et al. (2006, this issue).

Norwegian reference condition values are proposed only for diversity (ES100 = 26; \(S = 30; \text{ISI} = 9.9\)) (Solheim et al., 2004). Reference conditions by other MSs have not yet been fully determined; as many MSs consider it doubtful as to whether a true reference can be found with respect to diversity and species composition. However, in the present exercise, the Danish method used \(H_{\text{max}}\) (5) as a kind of reference, being in the good end of the diversity scale in this typology, while the UK method used multimetric assessments from impact gradients and associated reference sites to estimate reference conditions.

2.5. Boundaries and ES derivation

For this exercise each MS has proposed their own EQR class boundaries in assessing ES. These have been established in different ways: (i) Denmark has divided the EQR range into five equal parts, assuming that the part with the highest values corresponds to High quality conditions (Table 2); (ii) Spain has used the recommendations from the ‘Reference Conditions Working Group’ (REFCOND, 2003), as mentioned in Borja et al. (2004a) and Muxika et al. (2006); (iii) Norway has used determinations calculated in Solheim et al. (2004); and (iv) the UK has derived their boundaries from an anthropogenic pressure gradient, matching infaunal communities with the normative definitions provided for each status class.

2.6. Intercalibration comparison

Initial comparisons were carried out as pairwise comparisons, by means of linear regressions. The extent of agreement between pairs of MSs was then quantified, based on a single national boundary, corresponding to UK. This reported the boundary match/mismatch (‘High/Good’ and

<table>
<thead>
<tr>
<th>Ecological status</th>
<th>Denmark</th>
<th>Norway</th>
<th>Spain</th>
<th>UK and RoI</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>&gt;0.80</td>
<td>&gt;0.83</td>
<td>&gt;0.83</td>
<td>&gt;0.80</td>
</tr>
<tr>
<td>Good</td>
<td>0.60–0.80</td>
<td>0.72–0.83</td>
<td>0.62–0.83</td>
<td>0.65–0.80</td>
</tr>
<tr>
<td>Moderate</td>
<td>0.40–0.60</td>
<td>0.60–0.72</td>
<td>0.41–0.62</td>
<td>0.43–0.65</td>
</tr>
<tr>
<td>Poor</td>
<td>0.20–0.40</td>
<td>0.48–0.60</td>
<td>0.20–0.41</td>
<td>0.20–0.43</td>
</tr>
<tr>
<td>Bad</td>
<td>&lt;0.20</td>
<td>&lt;0.47</td>
<td>&lt;0.20</td>
<td>&lt;0.20</td>
</tr>
</tbody>
</table>
‘Good/Moderate’) between countries and the calculation allowed investigation into the consequences of changing boundaries. Following this, multiple boundary (‘high’ to ‘bad’) comparisons were investigated i.e. four-by-four comparison tables for each pair of MSs, by means of an Excel table which provides the agreement after modifying each of the boundaries. This table is available for any investigator, for further intercalibrations, upon request to the authors.

To analyse the agreement between MSs, a Kappa analysis was undertaken (Cohen, 1960; Landis and Koch, 1977). The level of agreement between the methods was established, based upon the equivalence table from Monseur and Leemans (1992). As the importance of misclassification is not the same between close categories (e.g. between high and good, or poor and bad) as between further categories (e.g. between high and moderate, or high and bad), Fleiss–Cohen weights were applied to the analysis (Fleiss and Cohen, 1973).

3. Results

The highest correlation between methods was found between the Spanish and Danish, and British and Danish methods, explaining 84% and 83% of the variability, respectively (Fig. 2). The remainder of the combinations explains only between 63% and 69% of the variability, the Norwegian method always showing the lowest correlation (Fig. 2). There is a group of stations (see Denmark–UK and Denmark–Spain, in Fig. 2), which appear as outliers in the analysis. Most of these samples correspond to station LIV in the Liverpool Bay Channel (UK), together with one sampling station from Ringgård Basin (Denmark). This difference may be explained by samples originating from a different habitat complex to that of the bulk of the samples or, as in the case of the DKI assessment of LIV, the use of an older AMBI taxon list with a different Lagis koreni assignment, which dominates the Liverpool Bay samples. For instance the Ringgård Basin is situated in an area on the border to mesohaline conditions (~19 psu) while other sites are from fully poly- to euhaline environments.

Taking into account the different boundaries used by each of the MSs (Table 2) the final ES assessment gives a ‘very good’ agreement (from a Kappa analysis) between the methods used by Denmark, Spain and UK (Table 3); however, the agreement between Norway and the remainder of the methods is only ‘good’. In this particular case, the relative number of cases in which the Norwegian method classified a station as ‘High’ or ‘Good’ and the remainder of methods as ‘Moderate’, ‘Poor’ or ‘Bad’ (or vice-versa), ranges between 27% and 33% (Table 3). For the other three MSs, the percentage ranges between 14 and 21% (Table 3).

![Fig. 2. Regression between the EQR, calculated by each of the methods.](image-url)
If the outliers are removed from the dataset the explained variability increases in all cases (87% in Denmark–UK; 89% in Denmark–Spain; 70% in UK–Spain; 67% in Norway–Spain; 68% in Norway–UK; and 70% in Norway–Denmark).

As the UK has determined their EQR boundaries in relation to an anthropogenic pressure gradient (as required by WFD) these values have been used to derive the boundaries of the other MSs, by using the regression equations in Fig. 2. The new boundaries, together with the confidence limits, can be seen in Table 4. These new values provide a ‘very good’ agreement between all methods, including that from Norway which previously presented a lower agreement. However, the relative number of cases in which a method classified a station as ‘High’ or ‘Good’ and the remainder of methods as ‘Moderate’, ‘Poor’ or ‘Bad’ (or vice-versa), only improves in the case of Norway, ranging now between 21% and 23% (note that both agreement and percentage of mismatch are not included in a table, only in the text).

Hence, it was necessary to check for new boundaries, by modifying all MSs boundaries together, as explained in Methodology, in order to get a better agreement between the methods, as required by the WFD. This 4 by 4 exercise provided an ‘almost perfect’ agreement between Spanish, British and Danish methods and a ‘very good’ agreement in the remainder of the combinations (Table 5). The new boundaries (Table 5) considerably reduced the percentage of cases in which a method classified a station as ‘High’ or ‘Good’ and the remainder of methods as ‘Moderate’, ‘Poor’ or ‘Bad’ (or vice-versa), ranging from 13% to 20% (Table 5). In this case, the obtained percentages are very similar between the four methods, allowing a better comparison of results and final ES assessment.
The EQR box-plot distribution for some of the locations, together with these new boundaries, is shown in Fig. 3. The locations with clear pressure gradients (San Sebastián, in Spain (location 18); and Garroch Head, in UK (location 25)) show the highest variability in the box-plot values, with an ample range of status. Conversely, some other locations show low variability and a high level of agreement between the four methods i.e. most methods coincide in classifying Oresund Funnel (in Denmark), and Utne and Trondheim (in Norway) as ‘High’ status; HA (Belgium), Arhus Bay (Denmark), German stations, Kenmare River (RoI), and Kilbrannan Sound (UK) as ‘Good’ status; and Ringgård Basin (Denmark) as ‘Moderate’ status. On the other hand, most methods classify the samples from stations O&P and subtidal locations (Belgium), Lillebælt North (Denmark) and MIL (UK) in the limit between ‘Moderate’ and ‘Good’ status; and Stavanger (Norway) and LIV (UK) in the limit between ‘Poor’ and ‘Moderate’ status. Finally, two locations (M&OD, Belgium; and HAR, UK) show different classifications, when comparing different methods.

4. Discussion

In this contribution a combination of pressures, including submarine outfalls, sewage sludge disposal, hypoxia, eutrophication, etc., have been studied in relation to the benthic invertebrate fauna. Following the WFD guidance in the intercalibration process (Pollard and van de Bund, 2005), as a conceptual model, when these pressures increase (i) a gradual decrease of ecological quality, in terms of decreasing diversity, (ii) a decrease of the ratio sensitive taxa/tolerant taxa (in this particular case, measured by the AMBI and the ISI index), and, (iii) probably, an initial...

---

**Fig. 3.** EQR box-plot for some of the locations (for details in numbers and names, see Table 1), calculated for each of the four methods, together with the boundaries for each of the methods determined in Table 5. Note: asterisks show the cases in which the total number of EQR values, calculated with the Norwegian method, is less than in the remainder of cases, due to the absence of calculation. Key: DK – Denmark; UK – United Kingdom; SP – Spain; N – Norway.
increase in abundance and biomass (in relation to organic load and eutrophication) should be expected. The methods used in this study agreed with this conceptual model and included all the biological structural parameters required under the WFD for assessing ES. However, some of the differences observed in the EQR and ES results can be explained by differences in the methodologies applied.

As mentioned by Van de Bund (2004), reference conditions are the starting point of WFD classification. Hence, agreement on reference conditions for the common intercalibration type is a requirement for intercalibrating the classification methodologies. However, at the time of this contribution only the Spanish and Norwegian methods incorporate clear reference conditions. The high level of agreement between classifications with the different methods suggests that the multimetric approaches used in the other methods, currently considering the highest values as reference, may be a valid approach.

Several international organisations, such as ICES or OSPAR, have produced guidelines for intercomparison and quality assurance of methodologies in sampling, laboratory analysis or data handling (Heip et al., 1992; Anonymous, 2002; ICES, 2002, 2004). However, one of the problems in intercalibrating methodologies for benthic quality assessment, within the WFD, is that when comparing biological elements there is not a precedent in such a process, so a comparison of approaches with previous intercalibration studies in other geographical areas is not possible. Recently, there has been a shift in emphasis within ICES and OSPAR towards comparable holistic evaluations of the biological status of the marine environment in relation to man’s activities (ICES, 2004). Guidance has also been provided for the intercalibration process within the WFD (Van de Bund, 2004), however, most of the concepts can be considered as subjective. It is hoped that this contribution can help in understanding the intercalibration process and provide practical guidance for other MSs and/or typologies when comparing their own methodologies.

The data used in this exercise were obtained on the basis of different grab types (van Veen, box-corer, Day grab, etc.), water depths (ranging from 3.3 to 180 m), geographical areas (from northern Spain to Norway), periods (from 1979 to 2004), seasons and pressures, as well as different taxonomy specialist teams (see Table 1). The Danish data required the pooling of several subsamples to make up 0.1 m², which may have created a bias towards higher EQR values for methods with high weight on species numbers. This is because there is a greater risk that several samples will be taken in more than one “habitat” compared to a single sample, resulting in an overestimation of the number of species. However, despite these differences, the final results are very consistent between the different methods used in assessing EQR and ES, even in absence of harmonised methodologies between the participants, as recommended in ICES (2004). Hence, probably, the total harmonisation in terms of sampling methodologies and analysis could not be as important (or relevant) as the requirement for habitat-specific reference conditions within water body types, in the final ecological assessment. As shown in this contribution, the inclusion of data from a different habitat, or changes in some species assignment, within AMBI, such as that from LIV (UK), can make difficult to compare results in terms of ES assessment.

It should be highlighted that the aim of the intercalibration exercise is to make all methods used by the MSs intercomparable, in terms of a final agreement in the ES assessment. Hence, although the MSs can use similar approaches, including the same or similar structural parameters, the intercalibration must determine the degree of agreement between methodologies. In this particular case, the achieved agreement between the four methods ranges from 77% to 86%. This range, in terms of biological elements, can be considered as very high, especially when taking into account the previous considerations on differences in sampling and analysis.

As stated by Gibson et al. (2000) core metrics in ecological assessment are those that will discriminate between good and poor quality conditions. Discriminatory ability of biological metrics can be evaluated by comparing the distribution of each metric at a set of reference sites with the distribution of metrics from a set of ‘known’ stressed sites within each site class (Gibson et al., 2000). This approach has been used in this study, and the results obtained are in relation to the different degree of degradation from reference conditions in the condition of benthic quality element. The degradation path is clearest in those MSs which have provided a clear impact gradient, such as Spain, UK or Denmark; this helping in interpreting and illustrating the normative definitions of the WFD. Hence, this gradient has provided a good basis for the determination of the boundaries between the different classes of ecological quality, and a further intercomparison of these boundaries between the different MSs methodologies.

On the other hand, British and Danish multimetric approaches give asymptotic EQR values. Hence, the real possibility to reach 1, in the EQR value, is impossible (see Fig. 2), making more relevant the need to determine different boundaries in the ES assessment, in order to avoid different final statuses between the different methodologies and MSs. This is, in fact, the aim of the European exercise of intercalibration: to achieve the same final ecological status assessment, for the same samples, using different methodologies.

5. Conclusion

The use of different methodologies in sampling and analysing benthic data was not an obstacle in comparing results from several impact gradients and geographical areas, in order to determine the EQR and ES. The methodology proposed here in the WFD intercalibration process, for NEA types 1 and 26, for poly- to euhaline subtidal sedimentary bottoms, is in accordance with the WFD guidelines, and allows to achieve a high level of agreement.
(77–86%) between the different methods. Hence, as the intercalibration within the WFD is a ‘work in progress’ (in fact, some of the methodologies presented here have been improved later to the acceptance of this paper, increasing the level of agreement between methods), this approach can be used by other MSs and applied to other typologies, in comparing their own methodologies and boundaries, and providing a basis for a comparable assessment of the ES through Europe, and, probably, for the new European Marine Strategy Directive (Borja, 2006).

Acknowledgements

This study was supported by different contracts, such as: the Department of Environment and Land Action of the Basque Government and AZTI-Tecnalia; Danish EPA; EU-REBECCA project SSP1-CT-2003-502158; Environment Agency of England and Wales and Scotland and Northern Ireland Forum for Environmental Research (SNIFFER). We wish to thank also to those investigators and organisations providing data sets, such as: Sofie Deroos, from the University of Ghent, Belgium; Joachim Voss, from Germany; Francis O’Beirn, from the Marine Institute of the Republic of Ireland; Julia Haythornthwaite, from the Scottish Environment Protection Agency (SEPA); Ian Davies, from Fisheries Research Services, Scotland; and Diputación Foral de Gipuzkoa, from Spain. Finally, the comments from two reviewers (Dave Jovett (Environment Agency, UK) and Anna Occhipinti-Ambrogi (University of Pavia, Italy)) have considerably improved the initial draft.

References


